Database Systems Journal BOARD

Contact
CaleaDorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro
E-mail: editordbjournal@gmail.com

CONTENTS

A Reinforcement Learning Approach for Smart Farming

Gabriela ENE
The Bucharest University of Economic Studies, Romania
gabriela.ene02@gmail.com

At a basic level, the aim of machine learning is to develop solutions for real-life engineering problems and to enhance the performance of different computers tasks in order to obtain an algorithm that is highly independent of human intervention. The main lying ingredient for all of these, is, of course, data.
Data is only valuable if it is transformed into knowledge, or, experience and the machine learning algorithm is only useful if it can make a prediction with high accuracy outside the examples in the training set. The field of machine learning intersects multiple domains such as data science, artificial intelligence, statistics, and computer science, but has appliances in any possible field that relies on decision making based on evidence, including healthcare, finance, manufacturing, education, marketing and recently, more and more in agriculture and farm-related management systems. As the Internet of Things and Cloud-Based solutions are introducing artificial intelligence in farming, the phenomenon of Big Data is going to impact the whole food-supply network. Machines that are connected with each other through a network or that are equipped with deep learning software or just with measurement systems are making the farming processes extremely data-driven. Fast decision-making capabilities might become a game-changing business model in this field.
Keywords: *machine learning, reinforcement learning, artificial intelligence, smart farming, Thompson Sampling, Q-Learning*

1 Introduction

Precision agriculture — a suite of information technologies used as management tools in agricultural production— has already advanced and will continue to change farm management, from the way farmers consider their commodity mix, scout fields, and purchase inputs, to how they apply conservation techniques, and even how they price their crops and evaluate the long-run size of their operations [1]. Mainly, the main focus of researchers and one big improvement for the farmers is the analytical causality between seeds and fertilizers or be-tween irrigations and crop quality. Traditional methods to determine relationships between such inputs and outputs relied on experiments or estimating data by mixing observed data sets with behavioral models, such as two-stage least square technique.
For a vast majority of farmers, the small plot experiments are mainly focused on few inputs and restricted to a determined time/season/location and cannot be often generalized, so the results may not be relevant and such implementation might be costly. The intersection of machine learning and agriculture might offer the starting point of a broader solution de-signed to optimize crop management. The aim of this article is to cover the implementation and the impact of reinforcement learning algorithms in smart farming, starting from the problem that many farmers face when they choose between using past production methods that bring income and exploring the value of new practices that can increase income. This problem fits under exploration vs. exploitation paradigm, and the focus of this paper is to conceptualize it as a multi-armed bandit problem. Also, on the same note, considering the increased costs of transportation, a conceptual implementation of a self-driven truck in an established environment is presented.

2 Content details

2.1 Types of machine learning algorithms

Machine learning algorithms can be classified into three categories:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

First-class needs a labeled data set in order to acquire the optimal knowledge. One good example would be the classification of a never-before-seen item based on a trained model that includes many items recorded in the data set with a corresponding label. Saying that z is the feature vector, as, in the data instance, the equivalent label of z would be f(z), known as the ground truth of z. The feature vector can be a multi-dimensional vector of different features that are relevant to the item, and the value of f(z) is one of a couple of classes of which the item belongs, so the model, is basically a classifier. If f(z) can have multiple values and the outcome values are ordered, then the model is a regressor.
Considering a prediction model of z as p(z), the success of the model under the influence of a parameter, $p(z|\theta)$, depends on the distance between these two vectors: f(z) and $p(z|\theta)$. This distance is known as the cost. The main goal of supervised learning is to minimize the cost, so to determine the parameters of the model that among all the data points of z, result in minimum cost.

Unsupervised learning assumes modelling data without knowing the associated labels.
Dimensionality reduction and clustering are very powerful tools that are broadly used to gain knowledge from data alone. The first one implies removing redundant features, in order to lower the dimensional space of the feature vectors, and clustering manages the process of distributing the data in specific classes without considering the pre-defined labels.
Unlike training labeled datasets provided by an external „teacher", and different from the approach of finding patterns in unlabeled datasets, reinforcement learning challenges the trade-off between exploration and exploitation.

2.2 Reinforcement learning

The very basic definition of reinforcement learning is acquiring knowledge through interaction with an environment. An agent acts in a specified environment and adapts its behaviour based on the rewards that it receives. The roots of the trial-and-error process are in behaviourist psychology [2], the agent main goal being to learn a strategy *[policy]* that would maximise the cumulative reward.
Reinforcement learning theory is already contributing to our understanding of natural reward, motivation, and decision-making systems, and it can contribute to the improvment of human abilities to learn, to remain motivated, and to make decisions [3].
The agent in the reinforcement learning algorithm, at a predefined time step, t, detects a state, s_t. The interaction with the environment assumes taking an action a_t, that will trigger the transition of both the agent and the environment to a new state s_{t+1}, defined by the previous one and the taken action. The state consists of sufficient statistics in order to offer the agent all the needed data in order to proceed in the best direction.
The rewards given by the environment determine the optimal sequence of actions, formally called „*policy*". The change of the state consists also in providing feedback to the agent, as a scalar reward r_{t+1}. Knowing the state, the policy will return a single action or a set of actions to perform.
One efficient technique to describe the

environment in an RL problem would be the *Markov Decision Process* approach, which provides an efficient model that can perform probabilistic inference over time. [4]

Markov Decision Process elements are as follows:

- The set of states - *S*
- The set of actions - *A*

Each s(i) state has its corresponding action or set of actions A(s(i)).

- The transition probabilities model P $\{S_{t+1} = s \mid St = s, A_t = a\}$

The probability of going from state s_t to state s_{t+1} depends only on the action and on the state.

- Reward function - *R(s)*
- Discount factor: $\gamma \in [0, 1)$

Once the agent takes an action a_t, selected from a set of actions that correspond to the state s_t, the agent gets the expected value of the reward, *R(s,a)* and, given the transition probability, the state of the process moves to the next one, so the model builds a path of transited states. Policy π is a mapping from states to a probability distributions over actions $\pi(s,a)$. So, it describes the way of acting. The function depends on the action and the state and returns the probability of taking the action in the specific state.

$$\sum_{a} \pi(s, a) = 1$$

The scope of the RL is to get the maximum reward from all states, with the optimal policy:

$$\pi^* = \frac{\operatorname{argmax}_\pi \mathrm{E}[R|\pi]}{\pi}$$

Learning the optimal policy implies using one of the two types of value functions available in machine learning: an action-value function - *Q(s,a)* – or a value function *V(s).*

Following a policy in state *s*, the expected return would be given by the formula:

$$V^{\pi}(s) = \mathrm{E}_{\pi}[R_t|s_t = s]$$

Even though the state is the same, the value function varies depending on the policy. The action-value function returns the value added by taking an action in a specified state when approaching some policy.

$$Q^{\pi}(s, a) = \mathrm{E}_{\pi}[R_t|s_t = s, a_t = a]$$

We can rewrite the value function in this manner:

$$V^{\pi}(s) = \mathrm{E}_{\pi}[r_{t+1} + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots|s_t = s] = \mathrm{E}_{\pi}[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s]$$

Taking into account the transition probability, and the expected reward that the agent receives by taking the action *a* and moving to state $s_{t+1,}$ we obtain the Bellman equation for the action value function:

$$V^{\pi}(s) = \sum_a \pi(s, a) \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^{\pi}(s')]$$

And also, we can do this for the action-value function:

$$Q^{\pi}(s) = \sum_{s'} P^a_{ss'} \ [R^a_{ss'} + \gamma \sum_{a'} \pi(s', a') Q^{\pi}(s', a')$$

This equation is important because allows expressing values of one state as values of another state, so if we know the value of a specific state we can determine the value of another.

3. The Thompson Sampling Algorithm

Thompson sampling is an algorithm for online decision problems where actions are taken sequentially in a manner that must balance between exploiting what is known to maximize immediate performance and investing to accumulate new information that may improve future performance [5].

Although it was first proposed in 1933, it is only in the past years that interest into its potential developed, and currently, it has been successfully applied in a broad variety of domains, especially in website management, A/B testing, portfolio management, or recommendation systems. The concept of the n-armed bandit problem is as follows: among a set on n actions, the agent is asked to make a choice.
Every choice will be rewarded with a numerical value selected from a stationary probability distribution. The objective of the agent is to maximize the value received after each action over a number of fixed iterations, or time steps. The greedy approach of this assumes selecting the action that will return the highest reward, a phase that corresponds to the exploitation part. Improving the estimate of a reward, by choosing one of the nongreedy actions is the exploration phase.
The multi-armed bandit problem is often presented as a slot machine with *n* arms. By pulling one arm at a time step, a reward is given and over M number of rounds, the player's scope is to obtain the maximum sum of the rewards.
Given the fact that the rewards are random, each one of the *n* arms defines for k ∈ {1, 2, … , n} a stochastic process $\{X_{i,m}\}$ in the form of a distributed sequence of random variables, with an unknown mean μ_i. One specific type of the bandit problem is the *Bernoulli Bandit*, which models the probability of an event occurrence, that follows a binomial distribution, with N = 1, which is basically, the Bernoulli distribution.
This model can be adapted to the problem of the farmer that needs to decide which plots to select as experimental plots for different seeding rates.
Supposing that the farmer manages many fields, the purpose of our model is to decide where to place experimental plots in order to obtain improved yield response.
Each field has different soil characteristics, such as nutrients, acidity/alkalinity, organic matter or type.
The decision implies using all this information for better placement of the experimental plots in the field. The feature vector, *v*, of a field consists of a predefined number of similar characteristics for each field and an area in the fields is defined by $\mathbf{M} = \sum_i |\mathbf{v(i)}|$, the total of the feature values. Let's assume that *v(1)* describes parts of the field by nutrients content, with values varying in different ranges: less than 4%, between 4.5% and 5%, 5% and 7% and over 7%, so the field would be divided into four areas.

Following the same approach, *v(2)* can classify the field into five areas depending on the pH value, so we would have 9 parts of the field that can overlap. Coming back to our model, each of these parts nine parts is an "arm' of the multi-armed bandit problem. By selecting an area and placing a plot there, the farmer observes and figures if the plot improved the total reward, in this case, the yield response.

The model we follow to track the yield response to the seeding rate is as follows, as propesed in [6] :

$$\text{Yield} = Ymax \times (1 - e^{-\beta \times SR}) \text{ [6]}$$

Ymax is the estimated asymptotic yield maximum, and β determines the responsiveness of yield as seeding rate increases.

Therefore, a smaller β indicates that a higher seeding rate is needed to reach maximum yield for that seed treatment. [6]

The nonlinear least squares (NLS) was used to estimate the parameters *Ymax* and β separately, an estimation that can be achieved by the algorithm [6].

Each area from the fields is assigned, at each step, a probability that selecting a plot that belongs to it will improve the estimation of the parameters.
The probability function is based on the previous steps, which resulted in the better or worse estimation of the parameters.

A reward of *1* is added if selecting the field area improved the accuracy of the prediction, and *0* otherwise. After that, the area with the greatest probability of improving the estimation is selected. At each iteration of sample selection, a new sample will be added to the training dataset.
The expected rewards are modeled using a probability that follows Bernoulli distribution with parameter $\pi_i \in [0, 1]$. We maintain an estimate of the likelihood of each πi given the number of successes αi and failures βi observed for the field area. Successes (r = 1) and failures (r = 0) are defined based on the reward of the current iteration. It can be shown that this likelihood follows the conjugate distribution of a Bernoulli law, a Beta distribution *Beta(αi , βi)* [7] :

$$P(\pi i \,|\alpha i, \beta i) = \frac{\Gamma(\alpha i + \beta i)}{\Gamma(\alpha i)\Gamma(\beta i)} (1 - \pi i)^{\beta i - 1} \times \pi i^{\alpha i - 1}$$

Thompson Sampling for Sample Selection [7]

1: Ymaxi = 1, βi = 1,S={}, M={areas}, A={1,2,…M},N=field areas ,∀i ∈ {1, . . . , M}
2: for t = 1, , ...N do
3: for i 1, . . . , N do
4: Daw ˆπi from Beta(Ymaxi, βi)
5: Reveal sample ht = {xt, yt, mt} from field areas Cj
where j := arg maxi ˆπi.
6: Add sample ht to S and remove from all field areas.
7: Obtain new model parameters Ymaxi, βi
8: Compute reward rt based on new prediction:

$$\text{Yield} = \text{Ymax} \times (1 - e^{-\beta \times SR})$$

9: if rt == 1 then Ymaxj = Ymaxj + 1
10: else βj = βj + 1

4 Q-Learning Algorithm

The underlying philosophy of this algorithm is based on the following method: the agent takes an action at a particular state and the feedback consists of a reward or a penalty. The agent can evaluate the feedback by estimating the value of the state to which it was taken. So, the learning is the process of going through different stages with the scope of maximizing the future return, R. The return from a specific time step, r_t, can be defined also by using the discount factor, γ, where $0 < \gamma < 1$, defined before as an element in the Markov Decision Process. The important thing to consider is that if the value of this factor is smaller, the agent would be inclined to choose only the immediate reward and not take into consideration the up-coming rewards.
If $\gamma = 1$, then all rewards are equally considered. The algorithm makes use of the action-value function and estimates the optimal function, with no regards of the policy that it follows. But, the policy is used also in this approach in order to map the pairs of states and actions that were updated.

The applications of this algorithm in farming are many, but we will consider one simulation, a delivery truck that is self-driven.
The truck's job is to get the crop from a determined place and to deliver it to another. Basically, the reinforcement algorithm that we will model will follow pre-defined steps as: environment observation, deciding upon the action to take based on the strategy of maximizing the obtained reward, acting, receiving the penalty or the reward, accumulating experience and improving the strategy and, finally, iterating until the optimal function is found.
A high positive reward is going to be obtained for a successful arrival at the location, and a penalty will be given if the truck arrives in the wrong place. The discount factor will be used when not getting to the destination after every step, meaning that late-arriving is better than making wrong moves.
The state-space consists of all situations that the truck may encounter and consists of useful data needed in the decision-making process.
Assuming that the field is the training area of the truck, we don't have to consider many obstacles that might be encountered, but only the area of the field, which we can divide in small plots, viewed as a matrix M.
For the purpose of the example, we would consider 36 possible plots, some of them may contain the silo and some can contain the harvested crop.
The actions will be defined as crop-load, crop-delivery, west, north, east and south. In the code, we would assign a penalty for every stop at the wrong silo location. The algorithm will only make use of the state space and the action space, and we will assign, in the defined order, a value from 0 to 5 to each action. For each state, the optimal action is the one that adds the most to the total reward.
When the environment is defined, a reward table or a matrix [number of states as rows, number of actions as columns] is created, named the Q-table. The table is used by the agent to acquire knowledge from it and hold the values of action-value functions, initially populated with zeros, but, during the training, with values that will optimize the agent strategy for the maximum total reward.
The first step of the algorithm is the creation of Q-table with 0 values. Secondly, the algorithm will iterate trough each state and select any of the actions that are available for the chosen state. As a result of the action taken, the agent „goes" to the next state and sees which action has the greatest Q-value.
The Q-table will be updated afterwards with values obtained from the Boltmann equation (2) and the next state will become the current state. This will be repeated until the goal is reached. After training with a large data set, it is proven that the agent has effectively learned the best move in a predefined matrix. Over time, the hyper parameters as the learning rate and the exploration level should decrease, as the gained knowledge increases, and the discount factor would increase as well because receiving the desired reward very fast is preferable.
Q-Learning learns the optimal policy even when actions are selected according to a more exploratory or even random policy [8].

5 Implementation

5.1 Multi Armed Bandit Algorithm Evaluation

For the Multi-Armed Bandit algorithm, we make use of the *pandas* library in Python. *Pandas* is a very powerful tool that enables a lot of tools for data processing with very high optimization.
Because of the lack of data that is needed for the conceptual problem described, the method of implementing the algorithm relies on a randomly generated a set of data.
For the sake of simplicity, we assume already observed data for the plots, and the yield already computed based on the model

described above[6]. We encode the obtained yield over fixed values as a good one, and we annotate it with „1", and whatever is below the value, with „0".
Using pandas, a DataFrame object is created with the randomly obtained values for 200 observations.

```
data = {}
data['A1'] = [random.randint(0,1) for x in range(200)]
data['A2'] = [random.randint(0,1) for x in range(200)]
data['A3'] = [random.randint(0,1) for x in range(200)]
data['A4'] = [random.randint(0,1) for x in range(200)]
data['A5'] = [random.randint(0,1) for x in range(200)]
data['A6'] = [random.randint(0,1) for x in range(200)]
data['A7'] = [random.randint(0,1) for x in range(200)]
data['A8'] = [random.randint(0,1) for x in range(200)]
data['A9'] = [random.randint(0,1) for x in range(200)]
data['A10'] = [random.randint(0,1) for x in range(200)]
data = pd.DataFrame(data)
print("\n\nPlot Data = ", data)
```

A list is initialized for the rewards associated with each plot, and one for all the penalties that belong to the plots.
For each observation, we iterate through each machine and based on the highest random beta distribution, the plot selected is updated with the plot/machine used. Once the plot is selected, the data corresponding to it is verified and we updated the list of rewards/penalties accordingly.

```
for m in range(0,total_observations):
    plot = 0
    beta_max = 0
    for i in range(0, machines):
        beta_d = random.betavariate(rewards[i]+1, penalties[i]+1)
        if beta_d > beta_max:
            beta_max = beta_d
            plot = i
    machine_used.append(plot)
    reward = data.values[m,plot]
    if reward == 1 :
         rewards[plot] = rewards[plot] +1
    else:
         penalties[plot] = penalties[plot] +1
    total_reward = total_reward + reward

print("\n\nRewards by machine = ", rewards)
print("\n\nTotal rewards = ", total_reward)
print("\n\n Machine used at each round: \n", machine_used)

plt.bar(['A1','A2', 'A3', 'A4', 'A5', 'A6','A7', 'A8', 'A9', 'A10'], rewards)
plt.title('MABP')
plt.xlabel('Bandits')
plt.ylabel('Rewards By Machine')
plt.show()
```

The output is :

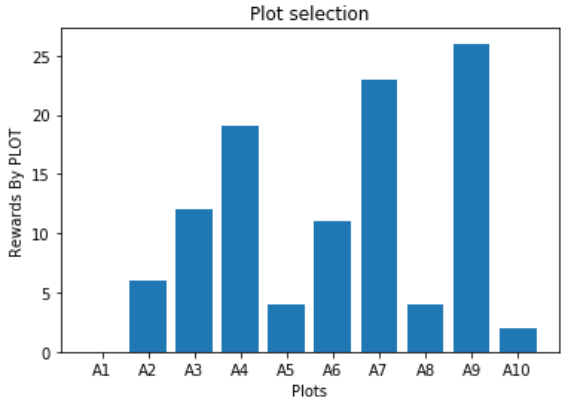


Fig. 1. Plot selection

In order to make sure that the algorithm selected the most optimal plot over time, we make a histogram of the plot that we use over time.

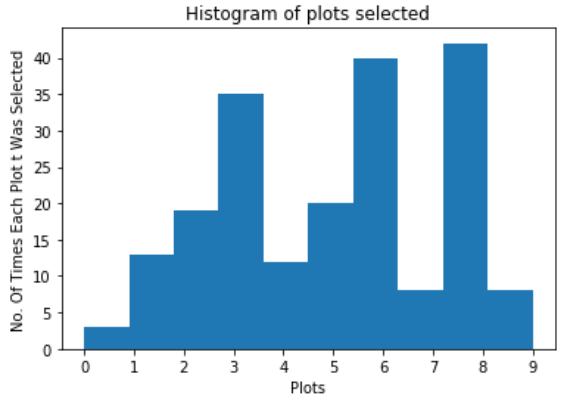


Fig. 2. Histogram of selected plots

By comparing the two graphs, we can see that the optimal plot was selected every time.

5.2 Q-Learning Algorithm evaluation

According to the GitHub repository, openAI Gym is a toolkit for developing and comparing reinforcement learning algorithms. For the implementation, the available collection of environments is useful for testing the agent, because the library also provides the required information as in states, scores or actions. In

openAI Gym, the environment replies with rewards, namely, scores.
We make use of *Env*, the core gym interface, and the predefined methods available: reset, step and render. Fortunately, openAI Gym provides a built-in environment, named „Taxi-v2", but the environment uses only a matrix of 25 possible agent locations and four possible destinations/locations.

In order to extend the locations, a new environment is created in openAI Gym. For achieving this, the environment is registered by calling gym's *register()* function, and by running the command *pip install –e* that takes as argument the location of the setup file where we defined the new environment. The custom made environment will be available with the call of:

```
env = gym.make('truck-v0')
env.render()
```

The focus is to break down the agent's learning experience into episodes.
Each episode starts by setting the first state of the agent randomly selected from the distribution. The agent iterates through episodes with the scope of maximizing the expectation of total reward/episode.

For our TruckEnv, we initilize „the field", like this matrix:

```
FIELD = [
    ['-','-','-','-','-','-','-','-','-','-','-','-','-','-','-'],
    ['-','1','0','1','0','0','1','0','0','0','0','0','0','-'],
    ['-','0','0','0','0','0','0','0','0','0','0','0','0','-'],
    ['-','0','0','0','0','0','0','0','0','0','0','0','0','-'],
    ['-','0','0','0','0','0','0','0','0','0','0','0','0','-'],
    ['-','1','0','0','0','0','1','0','0','0','0','0','0','-'],
    ['-','-','-','-','-','-','-','-','-','-','-','-','-','-','-']
```

Fig. 3. Field matrix

The state space is represented by :"truck_row","truck_col", „crop_location", „silo_destination".

For our example, we use a matrix with 5 possible crop locations and silo destinations. So, the total number of states is 6 * 6 [the possible positions of the truck in the matrix] * 5[the possible crop locations] * 5 [the silo destinations] = 900 possible states;
When the environment interface is built, the initial matrix of states and actions is created.
In order to view the structure, with the call of *env.P()*, we can see for each of the five possible actions the probability, the next state in which the agent will be if the action at the indicated index is taken, the amount of the reward[based on the type of the action and the position], and a boolean value, namely, „done", which indicates if the episode is successful.
In the environement, the P structure is initialize like this:

P = {state: {action: []
for action in range(number_of_actions)}
for state in range(number_of_states)}

We leave the other functions of the environment as they are defined in openAI gym default environment, because, basically, the truck will use the same context to train.

For training, the following hyperparameters were used:

- ***alpha***, the learning rate as 0.1
- ***gamma***, the discount factor and we set like this the importance of the future reward as 0.5
- ***epsilon***, the quantification of the exploration phase

Explaining it in a simple manner, it would be the decision of whether to check random actions or to make use of the already computed values in the Q-table. Epsilon is set in the code as 0.1

The number of episodes is set to 100000 for the training. We iterate through all the

episodes, and we reset the environment at each iteration to get a clean step.
For each step, we check the epsilon value against a random value from 0 and 1 and decide if the action is a random one or we just exploit the already known actions that have the greatest value in the Q-table.

Afterwards, the action is taken and the next step becomes the current one, the reward and the „done" boolean value are actualized along with the Qtable value for the specific state and action based on the formula described above.

```
for episode in range(total_episodes):
    state = env.reset()
    step = 0
    done = False
    for step in range(steps):
        if random.uniform(0, 1) < epsilon:
            action = env.action_space.sample() # Explore action space
        else:
            action = np.argmax(qtable[state,:]) # Exploit learned values
        # Take the action and observe the outcome state and reward
        new_state, reward, done, info = env.step(action)
        # Q(old):= Q(old) + lr [R + gamma * max Q(new) - Q(old)]
        qtable[state, action] = qtable[state, action] + alpha * (reward + gamma *
                                np.max(qtable[new_state, :]) - qtable[state, action])
        # Let new_state be state
        state = new_state
        # If done : finish episode
        if done == True:
            break
    # Reduce epsilon (because we need less and less exploration)
    epsilon = min_epsilon + (max_epsilon - min_epsilon)*np.exp(-decay_rate*episode)

if episode % 100 == 0:
    clear_output(wait=True)
```

For evaluating the performance of the agent, we make use of three indicators: the average number of steps taken to reach the destination, the average number of rewards and penalties per move, and so, after the training, we iterate through the range of episodes until the done indicator is true. By selecting the action only with the use of the Q-table, we can calculate these parameters and observe the performance.

```
env.reset()
rewards = []

total_rewards= 0
total_penalties = 0
total_test_episodes = 50
total_steps = 0
for i in range(total_test_episodes):
    state = env.reset()
    rewards, penalties = 0, 0
    done = False
    while not done:
        action = np.argmax(qtable[state])
        state, reward, done, info = env.step(action)
        if reward == -10:
            penalties += 1
        else:
            rewards += 1
        steps += 1

    total_penalties += penalties
    total_steps += steps
    total_rewards += rewards

print(f"Results after {total_test_episodes} episodes:")
print(f"Average timesteps per episode: {total_steps / total_test_episodes}")
print(f"Average penalties per episode: {total_penalties / total_test_episodes}")
print(f"Average rewards per episode: {rewards / total_test_episodes}")
```

The output is as follows:

```
Average timesteps per episode: 405.18
Average penalties per episode: 0.0
Average rewards per episode: 0.34
```

Fig. 4. Performance of the agent

6 Conclusions

The evolution of Big Data and Machine Learning will change the methods of farm management and is actually changing the research methods. Optimization is definitely improved by the prevalence of data and rapid estimation of causality relationship of inputs is overpassing the traditional approaches. Also, the adoption of data-driven technologies will play a big role in conserving resources and expanding the returns. Analysis of data from software that manages irrigation reduces water consumption and impacts environmental management. Predictions based on the historical data are being replaced with a comprehensive analysis of the crops, based

on real-time input. Machines can also classify and detect plant disease reducing costs this way and improving the quality of the crops. Progress in machine learning has been driven by low-cost computation opportunities as well by the availability of online resources, data and the development of learning algorithms.

References

[1] D. J. Russo, B. Van Roy, Benjamin, A. Kazerouni, A. Osband, "A Tutorial on Thompson Sampling. Foundations and Trends", *Machine Learning*, 11. 10.1561/2200000070, 2017

[2] S. Sukhbaatar, A. Szlam, R. Fergus. "Learning Multiagent Communication with Backpropagation", *NIPS*, 2016.

[3] P. Sterling, S. Laughlin, "*Principles of Neural Design*", MIT Press, Cambridge, MA, 2015

[4] CC Bennett, K Hauser, "*Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach*", Artificial intelligence in medicine, Elsevier 2013

[5] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen (2018), "*A Tutorial on Thompson Sampling*", Foundations and Trends® in Machine Learning: Vol. 11: No. 1, pp 1-96

[6] A.P. Gaspar, P.D. Mitchell, S.P. Conley, "Economic risk and profitability of soybean fungicide and insecticide seed treatments at reduced seeding rates". *Crop Sci*, 55, 924-933, 2015

[7] B.Gutiérrez, L. Peter, T. Klein, C. Wachinger, "A Multi-Armed Bandit to Smartly Select a Training Set from Big Medical Data", *Computer Vision and Pattern Recognition*, arXiv:1705.08111, 2017

[8] R. S. Sutton, A.G. Barto, "*Reinforcement Learning: An Introduction*", MIT Press, 1998

Gabriela ENE has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2015. Currently she is a PhD Student at the same university. Main fields of interest are big data technologies, artificial intelligence, web development, algorithms and data structures.

Monitoring and Controlling Electricity Consumption Application for Smart Buildings

Maria Irène CĂLINOIU, Simona-Vasilica OPREA
Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies, Romania
calinoiumaria@yahoo.com, simona.oprea@csie.ase.ro

In the context of global concern about climate change and energy security, saving and optimizing electricity usage has become a priority of the European Union's energy policy. In the member states, households consume about 27% of the total electricity used [1]. *Thus, controlling electricity consumption remains an important objective for this sector as well. Developing a reliable application for electricity monitoring and optimization is a way to reduce individual consumer consumption, as well as a solution to avoid overloading the distribution network.*

Keywords: *electric energy consumption feedback, energy literacy, consumer profile, shifting optimization algorithm, mobile application design*

1 Introduction

Electric energy is a basic need, whose importance is seldom acknowledged. In the absence of an efficient, real-time monitoring system, consumers cannot perceive energy as an essential aspect, whose usage depends strictly on their habits and way of life.

Energy literacy is defined as "the understanding of energy concepts necessary to make informed decisions on energy use at both individual and social levels". [2]. This education can be formal, obtained during the schooling process, or acquired through everyday experiences.

In order to make the concept of energy more transparent, information concerning energy consumption should be made available to all consumers, with the purpose of avoiding waste and properly managing usage where the demand is bigger than the offer. Thus, consumers will be made aware of their behavior's direct impact on total consumption, usage costs, and the environment. Consequently, they will be more motivated to adapt their lifestyle in order to alter consumption habits.

Feedback, known as the process of obtaining information regarding behavior that can be used as a basis for improvement, can help increase energy literacy, therefore impacting consumer behavior [3].

Currently available technology offers the means for analyzing usage patterns. Among those devices, Smart Meters (SM), Home Energy Displays or In-Home Displays (IHD) are the most common. SM are advanced metering devices capable of recording consumption in detail and communicating it through a network to the provider in order to monitor it and simplify the billing process. Installing SM presents multiple advantages, among which: offering feedback in real time, eliminating the need to be read periodically by an employee, discouraging fraud, remotely controlling the system, and quickly identifying malfunctions [4]. IHD are modern gadgets that show more detailed information regarding consumption data in real time, by measuring data recorded by a standard meter, or by receiving data from SM.

The literature concerning feedback on energy consumption shows that, despite the initial attraction towards these devices, consumers eventually lose interest if the data provided is not considered relevant, intuitive, readable, and understandable. Moreover, the studies stress the importance of accessibility, and show that, in order to be effective, the information presented needs to be displayed somewhere the user finds convenient and within reach.

Web and mobile software products complement the physical devices designed to monitor consumption, offering flexibility and interactivity, and filtering out information deemed unimportant depending on user preferences and on relatable motivational factors.

Differential tariffs, also known as *Time-of-use tariffs*, or **ToU**, work by establishing varying costs per kWh based on the time interval during which usage takes place. They represent an alternative to fixed tariffs and could offer advantages to both consumers and energy providers.

ToUs allow optimizing consumption distribution by discouraging consumption at peak hours, mainly by raising the energy costs for certain busier time intervals. To encourage the desired behavior changes, energy companies offer low tariffs outside of peak hours, e.g. at nighttime. Implementing these tariffs is possible only through smart grids, where consumers use SM that record their usage dynamically. Through such technological advances, the former, more centralized approach is transformed into a dynamic process, adapted to client demand, known as **Demand-Side Management** (DSM) [5].

2 Objectives

The mobile application described in this article, called **Watt Manager**, is destined to put the aforementioned concepts in practice. It provides users with quick access to current and historical consumption data, allowing them to track and to optimize their usage. Consumption is monitored through goal setting, optimized by scheduling appliances in order to minimize the consumption peak, and visualized through interactive and concise diagrams and reports. Moreover, it aims to raise awareness regarding household usage habits, and to increase energy literacy.

What sets Watt Manager apart from other consumption monitoring applications is the increased focus on the consumer, and the flexibility regarding user needs. It is customized from all points of view, from units of measurement to active features, giving the user full control over the information provided. Thus, it is suitable for all household members, regardless of their age or level of involvement.

A survey completed at signup generates a consumer profile, the basis on which the app is run. This method is more effective in registering the client's preferences than a bland settings page, because it captures the user's interest, and it helps establish their motivational factors and determine their level of adaptability. Therefore, the application is customized according to the client's billing method and needs, displaying only the fitting information.

The application promotes education in the energy domain through a **Trivia** game. The game's goal is to develop the clients' knowledge and to raise awareness regarding the impact of energy consumption on the environment. Moreover, a feature that schedules appliances, based on the restrictions imposed by the client, offers better control of energy usage.

However, drastic lifestyle changes are not the app's objective. Instead, it aims to obtain gradual shifts in behavior, varying from one household to another and based on habits and consumer flexibility.

Through implementing a **shifting optimization algorithm**, the electricity consumption curve is flattened, generating a more homogenous grid load throughout the day [6]. The scheduling algorithm must take both personal and natural restrictions into account. Such restrictions could refer to certain dependencies between appliances, or to particularities of the programmable devices considered, such as whether they can be scheduled with interruptions or not.

Watt Manager eliminates the uncertainty concerning household energy consumption by displaying usage in real time. Based on the detailed information provided by the application, the clients can visualize their total usage, or grouped by category, by time interval, or by individual meter, helping them become more aware of the way their habits impact their consumption.

In addition, the app allows users to monitor consumption through goals. It also signals tariff details in the case of ToU tariffs. These elements offer the necessary information to identify the opportunities for achieving behavioral changes, while also motivating users to reduce consumption.

The app keeps a record of the household's bills, offering the possibility to accessing all the data concerning energy consumption on a single platform. The graphs and the reports generated allow a detailed analysis, depending on the type of graph, the level of aggregation, the measurement unit, and the time interval selected by the consumer.

The Trivia game offers users a chance to find out more about the energy field and, as a result, improve their consumption habits.

3 Designing the Application

Watt Manager is designed in order to facilitate the implementation of the features described above.

The advantages of a relational database, such as the well-defined relationships between its entities and the possibility to eliminate redundancy, make it the chosen format for the app's database. Its logical schema is presented below, in Figure 1.

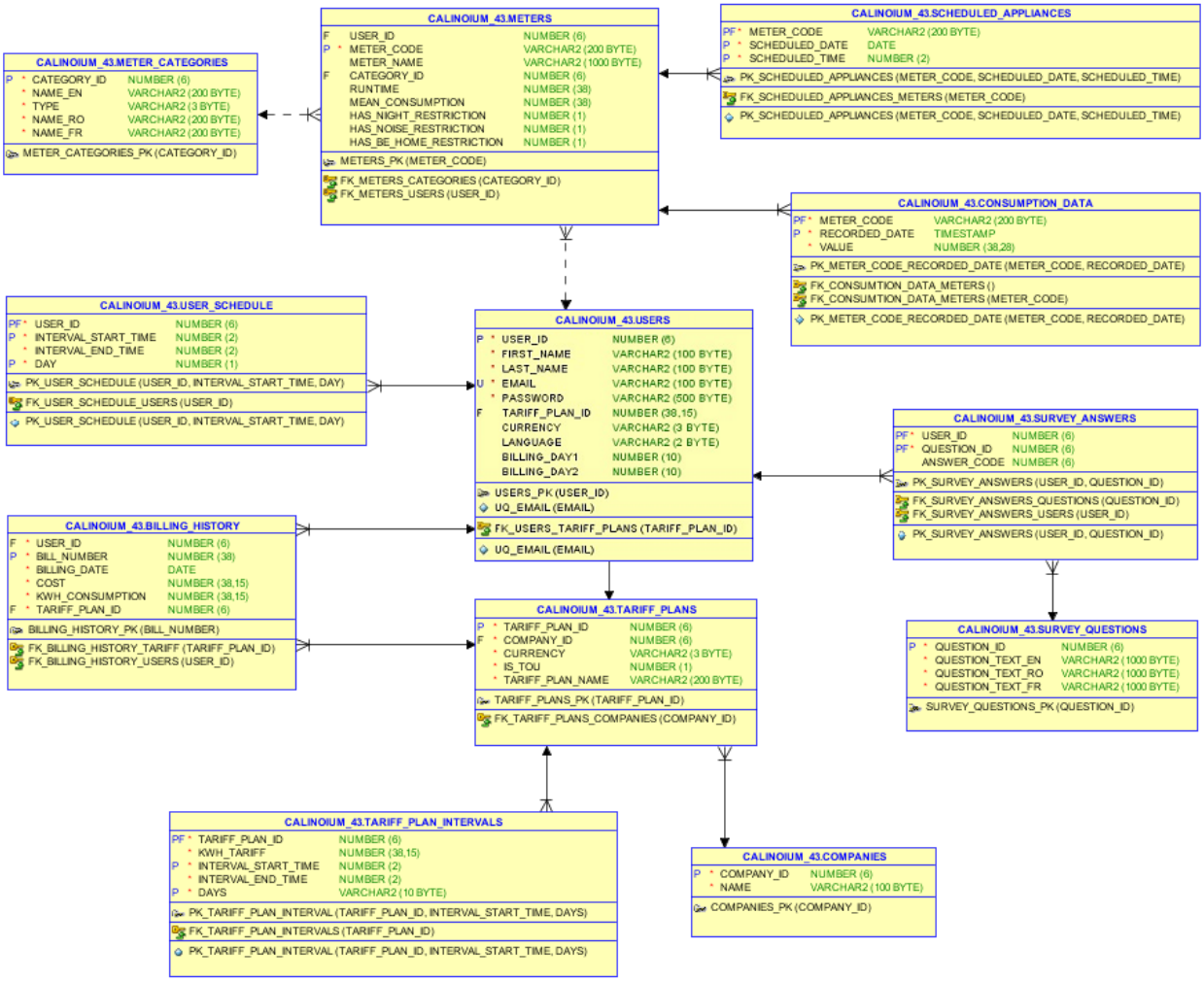


Fig. 1. Logical Schema of the Database

Reaching as many consumers as possible is one of the app's main objectives. Therefore, an Android mobile app provides an excellent platform, being available to most of the population that uses a cellular phone. Furthermore, a mobile app is the most accessible, requiring the least amount of effort and technical skill to use.

Mobile user interfaces in Android are represented by activities and dialogs. Following the description of the mobile application's functionalities and the data

flow, Watt Manager's screens are shown schematically below, in Figure 2. Links between interfaces show paths that can be accessed by the users within each screen.

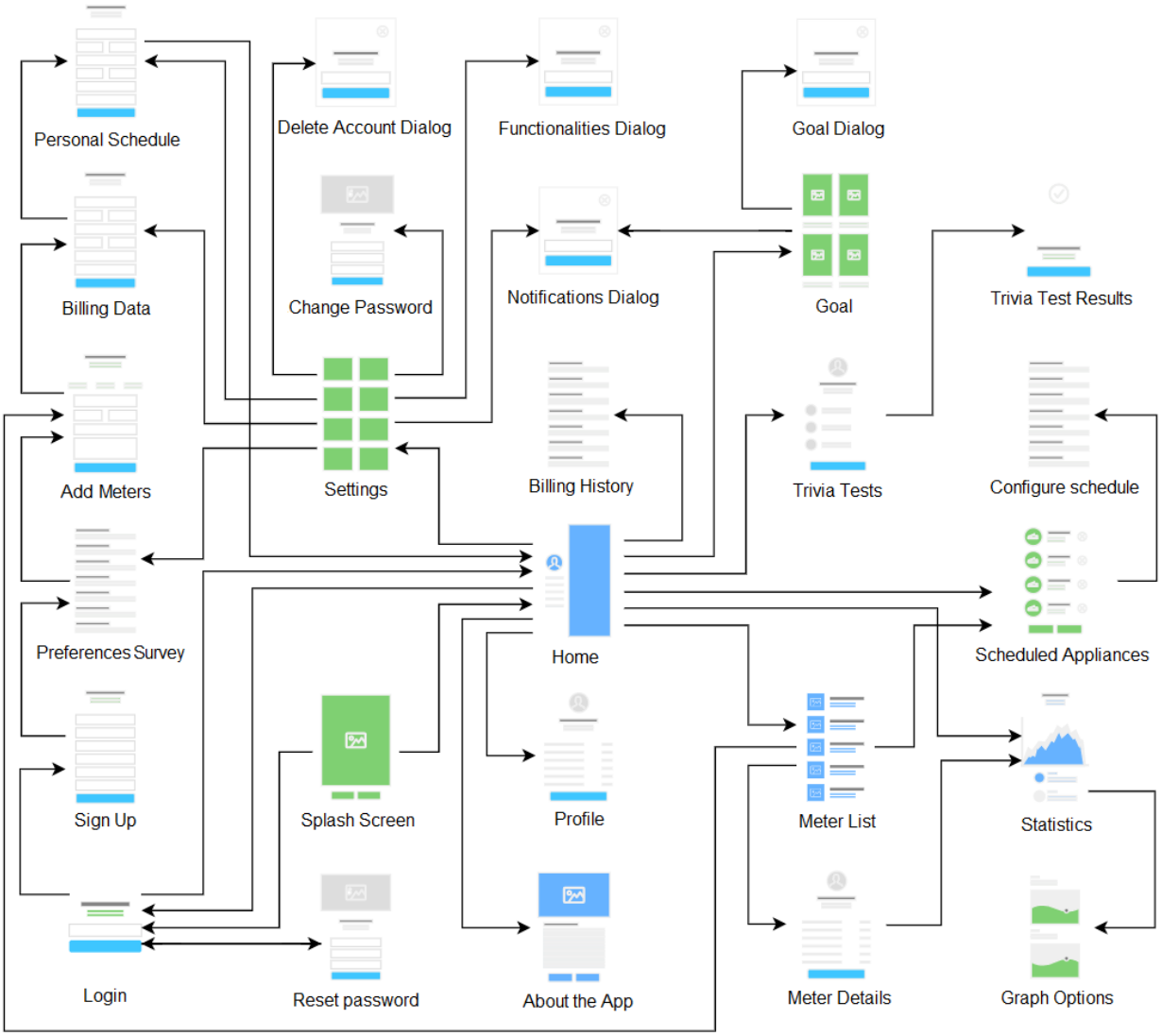


Fig. 2. User Interfaces Diagram

4 Software Technologies

Developing the application involves creating three software components: an Oracle relational database, a Flask REST server, and an Android mobile app.

JSON (JavaScript Object Notation) is a text format that uses key-value pairs. This format is used to transfer messages between the server and the app.

Extensible Markup Language, or **XML**, encodes documents using tags marked with symbols "<>", which describe the transmitted data. They are used in the mobile app as compiled resources to define the visual elements that make up the user interface, character strings, menus, colors, styles, and sizes.

Representational State Transfer (**REST**) is a client-server software architecture that acts as an intermediate level between the application and the database. A **REST-API** (Application Program Interface) separates the user interface from the data storage.

Isolating these components has several advantages. First, the application becomes scalable. Also, connecting to the database in a controlled way makes the process more secure. At the same time, it ensures portability, and the possibility of extension without constraints regarding the connection

to the database. Moreover, it allows the database or mobile application to be changed, or to extend the system by adding an iOS application, a Web platform or a module for power suppliers, without requiring major changes to the other components. Finally, since the architecture implements a stateless protocol (**HTTP**), it is not necessary to load the server with data prior to each request because the information is sent in such a way that it can be understood without taking into account previous messages.

The mobile application places requests to the server through a uniform resource identifier (**URI**), the web address the server is running. The **GET** method is used to retrieve data, the **POST** and **PUT** methods are used to store the transmitted information into the database, and the **DELETE** method erases information from the database.

Python is an intuitive, high-level, dynamic, general-purpose programming language. The software principles on which Python was developed are synthesized in a collection of 19 expressions, "The Zen of Python", by one of the developers of language, Tim Peters. Among them, are: “Beautiful is better than ugly”, “Explicit is better than implicit”, “Simple is better than complex”, “Flat is better than nested”, “Sparse is better than dense”, “Readability counts”, “In the face of ambiguity, refuse the temptation to guess” [7].

The Python standard library is very large compared to other programming languages. Python Package Index, the official repertory, offers over 130,000 libraries. They facilitate both software development and rapid data processing in the exact sciences. Since most libraries are open-source, development in Python offers the programmer flexibility.

There are a number of Python development environments, such as *PyCharm, Jupyter Notebook, or Amazon Web Service Cloud9*. **Amazon Web Service Cloud9** is an integrated development environment based on Amazon Cloud. The advantage offered by Cloud9 is providing the URI required for HTTP calls to the REST server. Normally, non-cloud-based development environments either allow the creation of a local server or require the configuration of an external server. Since the application architecture is client-server, it is essential that the remote server exists to allow the interaction between the REST-API and the mobile devices running the application. This platform provides the programmer with an external IP that makes this possible.

Flask is a Python micro-framework designed for server development, following the REST paradigm for system interaction. The Watt Manager server was built with Flask. The messages transferred between the app and the server are parsed by the server with the **JSON** library. To retrieve data from the Oracle database, the connection is established with the help of the **cx_Oracle** library, and the queries use its *Cursor* object. User passwords were secured by encoding them in the database as *hash* values using the *Security* module in the **Werkzeug** library. The libraries used for data processing include **NumPy** and **Pandas**. NumPy is the fundamental package for mathematical calculations, with many predefined mathematical functions. Pandas provides data structures (such as *Series* and *DataFrame*) and functions for manipulating data.

Java is an object-oriented programming language, created to facilitate the development of distributed applications, which can be executed on heterogeneous networks [8]. Java is also used to develop mobile applications for devices that use the Android operating system.

Android Studio is the official Android app development environment, created by Google. Watt Manager was developed in Android Studio for a minimum level API 21, the equivalent of Android 5.0 (Lollipop), and a target level API 28 (Android 9). Thus, approximately 89.3% of existing Android devices can run the application [9]. The app only requires an Internet permission, being based on the client-server architecture which requires online communication.

The **Gson** library is implemented by Google in order to allow the conversion of Java objects or of user-defined serializable objects to JSON. Also, Gson can transform JSON character strings into objects. This

quickly generates the POST and PUT requests, eliminating the need for manual definition. **OkHttp** is an HTTP client which facilitates the creation and execution of the requirements in Android towards the remote server. **MPAndroidChart** is an Android library used to build the charts in the Statistics screen, offering the ability to draw and personalize complex graphs with ease.

Relational databases are built on relational set theory. They use attribute integrity restrictions to control the data's accuracy, and associations between entities to ensure consistency. The tables are normalized, giving them a structured character, with minimal and controlled redundancy. Moreover, they support relational algebra operations, such as: union, intersection, difference, Cartesian product. In addition, they also support specific SQL operators. The advantage of using an SQL database is the mathematical, intuitive character of its design, construction, and query.

Structured Query Language, or **SQL**, is a programming language used for interrogating and manipulating relational databases. Procedural Language/Structured Query Language, or **PL/SQL**, is an Oracle extension of the SQL language. It contains modular code blocks stored directly in the Oracle database, enabling data control, manipulation and validation. SQL is used when creating, modifying, and querying the app's database, while PL/SQL is used to create triggers and sequences that generate indices automatically.

Oracle SQL Developer is a development environment used to build and to manage Oracle relational databases, implementing SQL and PL/SQL. It allows concurrent access and ensures data persistence using a remote database server.

5 Methodology

The shifting optimization algorithm represents the most complex processing sequence in the app's server. It receives the scheduling date and a list of devices to be scheduled, together with the initial hours proposed by the user as input values.

The database stores the restrictions imposed for the algorithm. The first type of restriction is given by each device's category. If the dryer and the washing machine are programmed on the same day, for example, the dryer should operate immediately after the washer, being scheduled accordingly. A device can have three optional restrictions: it can't be planned at night, it can't be planned during quiet hours, and it can't be planned in the absence of the user (they must be home during the appliance's runtime). At signup, users who benefit from ToU rates are invited to provide their approximate weekly schedule. This is the point of reference for the third restriction described above.

The algorithm's goal is to flatten the consumption curve by minimizing the peak usage, considering the constraints outlined above. Starting with the hours proposed by the user for each appliance, the algorithm follows an iterative process to gradually reduce maximum consumption, until it can no longer be improved.

The algorithm is adapted from the implementations presented in [5] and [6], with added user-imposed constraints for each device. Also, the way dependent and interruptible devices are handled is different. To manage dependent devices, they are grouped together, forming a single device with a summed runtime. By contrast, interruptible devices are divided into several uninterruptible devices, one for each hour of the initial device's runtime.

The algorithm is described with the help of the logical schema presented in Figure 3.

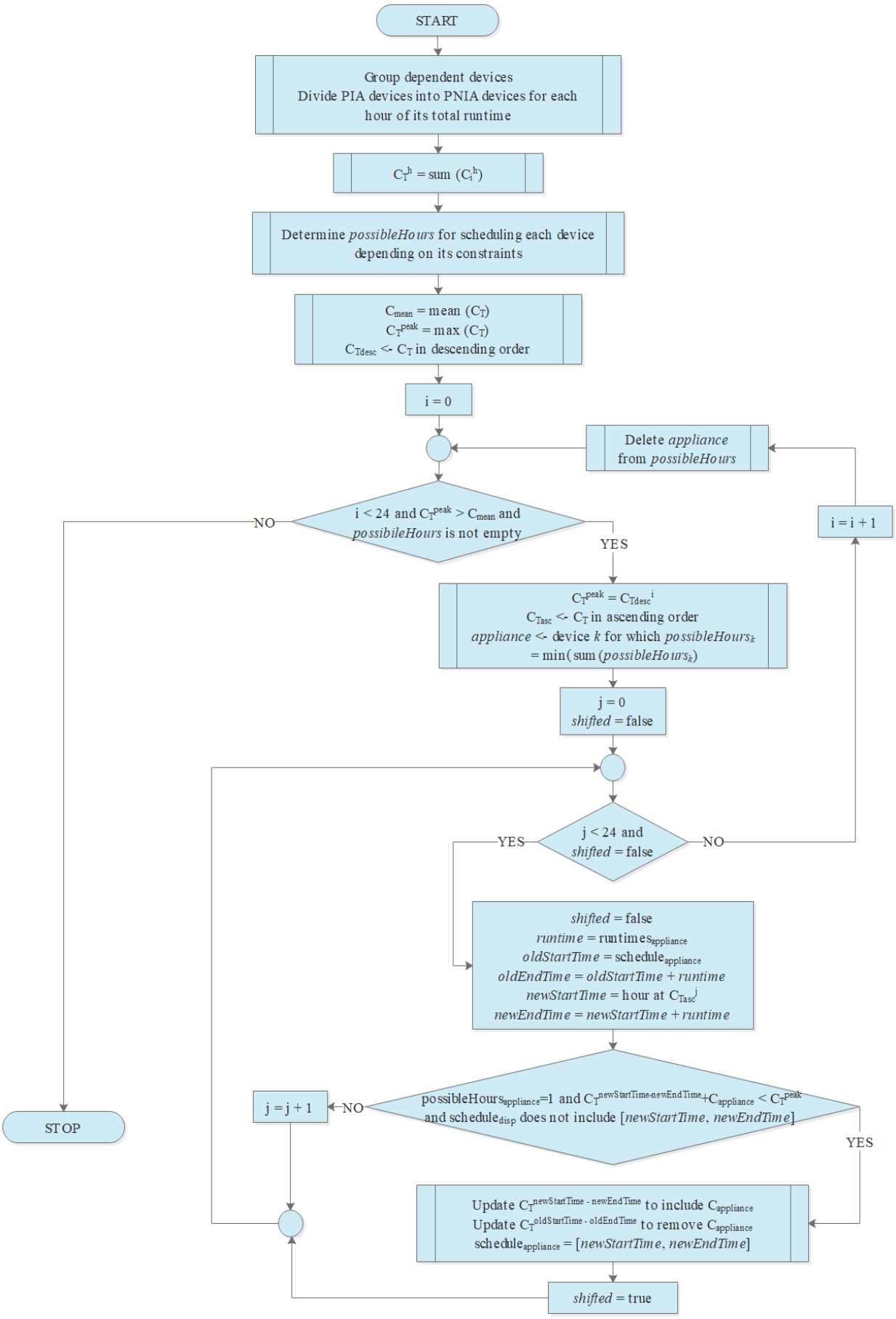


Fig. 3. Logical Schema of the Shifting Optimization Algorithm

C_T	Total consumption array for each hour

possibleHours	Boolean matrix with indices representing scheduled devices and hours representing columns, indicating possible scheduling times for device i at hour j, depending on its constraints
C_{mean}	The mean of the C_T total consumption array
C_T^{peak}	Peak consumption, calculated as the maximum of C_T
C_{Tdesc}	Array C_T in descending order
C_{Tasc}	Array C_T in ascending order
appliance	The appliance with the least number of possible scheduling hours, calculated by obtaining the index of the row with the minimum sum in the *possibleHours* array
shifted	Boolean variable signaling whether the current device has been shifted during the current iteration
runtimes	Array that stores the approximate runtime for each device scheduled
oldStartTime	Scheduled time for the device before it is shifted
oldEndTime	Time until the device runs before it is shifted, obtained by adding its runtime to *oldStartTime*
newStartTime	Scheduled time for the device after it is shifted
newEndTime	Time until the device runs after it is shifted, obtained by adding its runtime to *newStartTime*

6 Application Interface Depiction

Using the Watt Manager mobile application begins at signup, with the registration of a customer who owns smart meters that record consumption data. First, they introduce their personal data. Then, they complete the preference survey, add their meters by code, complete the billing form, and add their personal schedule.

Adding a meter requires a user-provided code. If the code exists in the database, it is validated and associated with the user, along with the name and category given. If the selected category represents programmable devices, the user also enters its runtime and its schedule restrictions in order to ensure the correct operation of the planning algorithm in the future.

Then, after adding their meters, users continue to the billing data form. If the selected tariff plan is a ToU plan, they introduce their personal weekly schedule. After completing the process, the client accesses the Home screen of the application. This screen also differs depending on the tariff type.

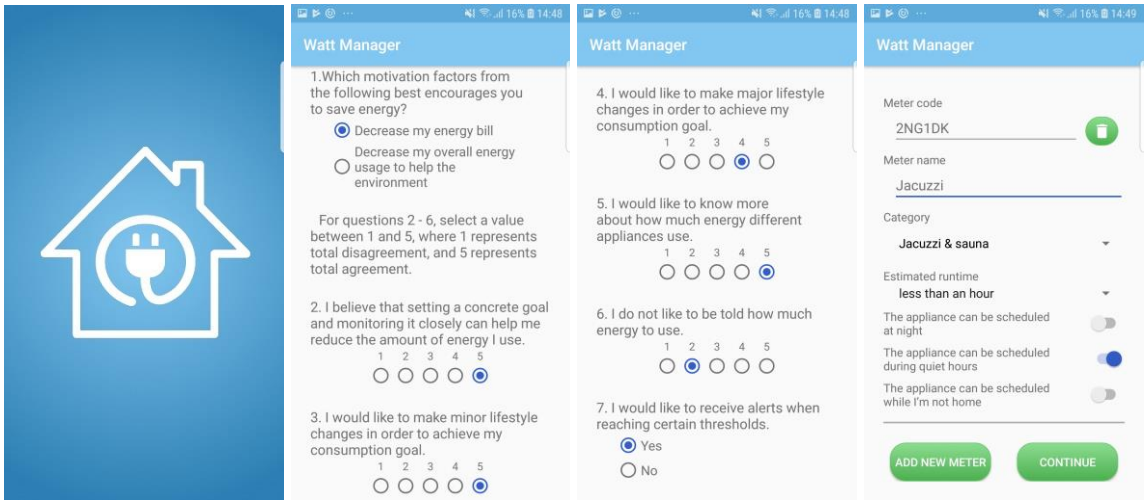


Fig. 4. Loading Screen, Survey, and Adding Meters

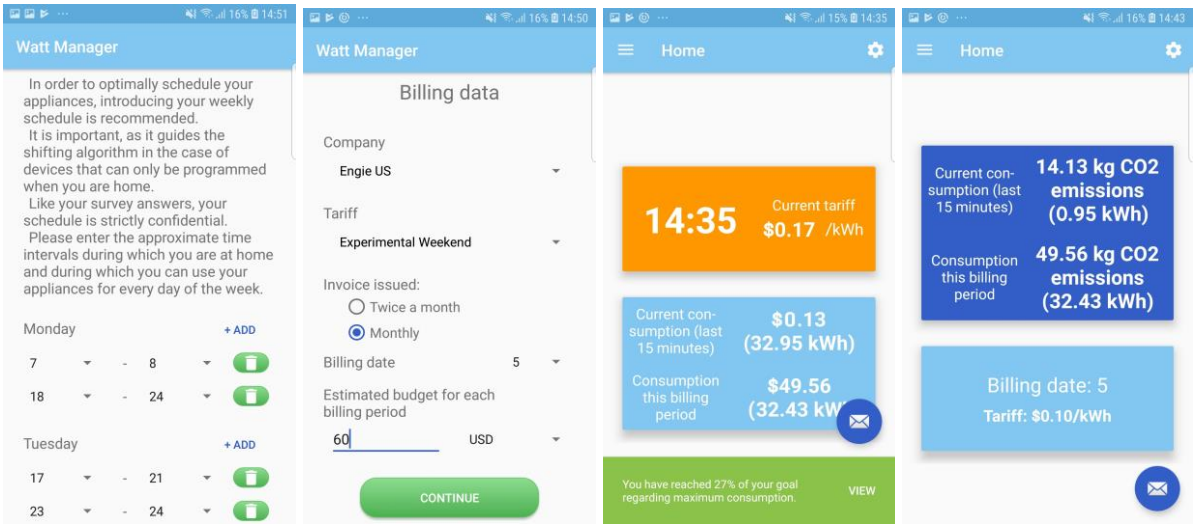


Fig. 5. Personal Schedule, Billing Data, and Home Screens

From the app's menu, users can access all of the app's functionalities.

The main screen contains two boxes and a Floating Action Button. The button shows the user's goal progress, and pressing it opens the goal screen. If the user benefits from a ToU tariff plan, the first box displays the current hourly rate. It is colored green, yellow, orange, or red, depending on the current cost of electricity relative to the minimum, maximum, and average tariff in the customer's plan, visually representing the current tariff. The second box shows current and total consumption for the time unit selected. If the user does not have a ToU tariff, the first box indicates total usage, and the second box shows billing details and the constant rate per kWh.

Selecting the box showing the consumption sends the user to the Statistics screen. By default, it displays separate consumption data for each meter so that the customer can better understand their consumption habits. If the user wants to generate another graph, they provide the measurement unit, the time unit, and the type of chart desired. Depending on this information, the user chooses the type of consumption (total, aggregated by category, aggregated by tariff, or disaggregated) and the app generates a chart.

Selecting the box that displays the tariff leads to the user's profile, including a more detailed view of the tariff plan.

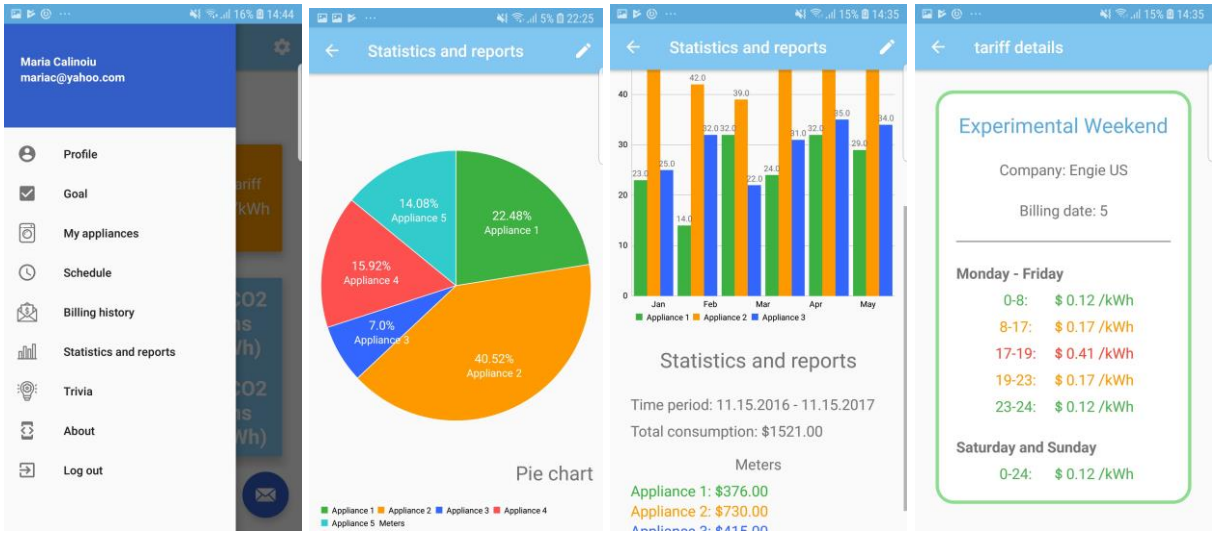


Figure 6 – App Menu, Statistics, and Tariff Details

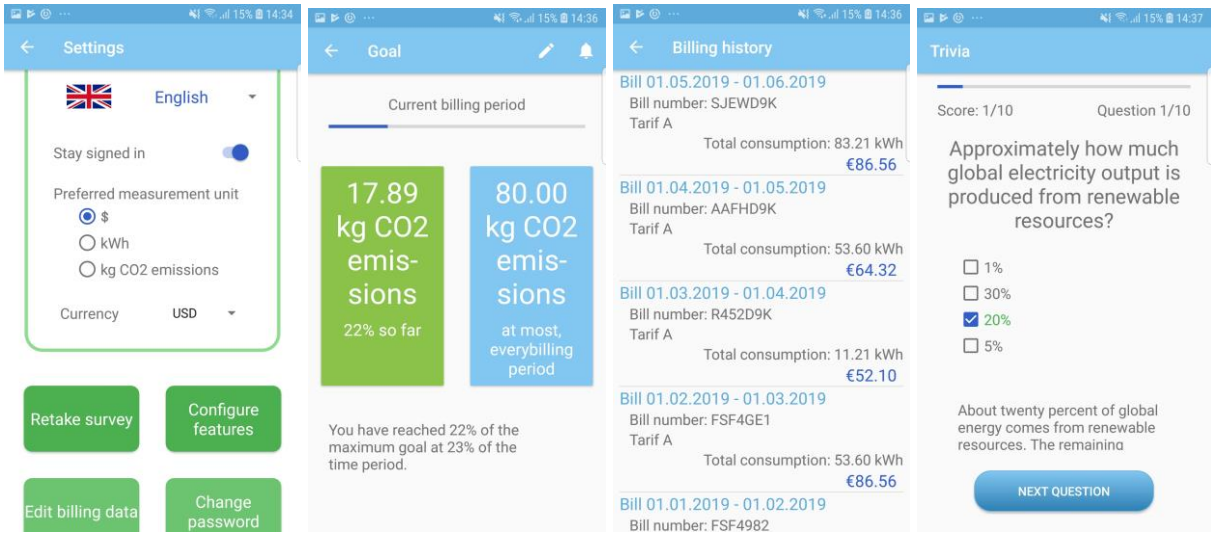


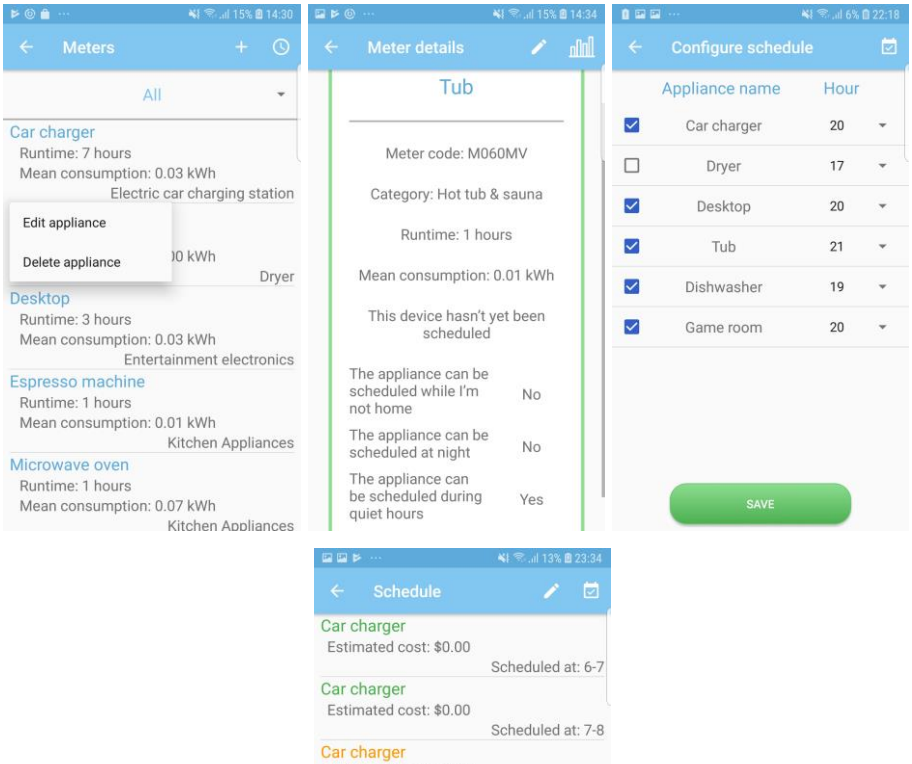
Fig. 7. Settings, Goals, Billing History, and Trivia

App settings allow the user to edit the data recorded at signup, except for their personal data. The user can also change their password and delete their account, enable or disable features, and configure notifications.

The Goals screen shows the current progress during the period for which the target is set. Like the Home Page box, the progress box is colored from green (0% of the maximum usage reached), to red, (exceeding the maximum usage set). To set a goal, the user accesses the dialog from the header menu. There, they can also enable or disable goal notifications.

Trivia tests contain ten questions regarding electric energy, aimed to develop consumer knowledge and to raise awareness regarding the impact of daily usage on the environment. After selecting and confirming an answer, the correct response is indicated, along with an explanation.

The appliance list displays the user's meters. Clicking a device shows its details, and long-clicking it displays the context menu that allows editing or deleting it. The scheduling feature can be accessed from the Appliances header menu, leading to a screen where the user can select which devices to program, the estimated scheduling time, and the date. After running the algorithm on the server, the app returns its result in the scheduling screen. There, the names of the planned devices are colored according to the tariff value at their programmed time.



7 Conclusions

The objective of the paper was to analyze the best way to implement an application for monitoring and controlling electricity consumption, regarding the functionalities included and the technical aspects involved. The positive impact on the environment produced by decreasing usage and adapting consumer behavior to avoid overloading the distribution network provided the motivation for choosing this subject.

The proposed IT solution facilitates consumption surveillance by providing the most relevant feedback to each user. The profile established on the basis of a questionnaire completed at signup captures the user's motivational factors and preferences. This determines the content of the application, in order to determine users to reduce their consumption or to optimize it. The application follows a three-tier architecture: an SQL relational database, an intermediate REST server developed in Python Flask, and an Android mobile app developed in Java.

The existing application can be improved by building the necessary hardware components for measuring consumer data. In addition, considering multiple sensors can provide additional details on consumer behavior. Thus, recommendations become even more personalized. Also, adding a Watt Manager version compatible with iOS devices or other operating systems would be an important step to cover a larger proportion of consumers.

The system can be expanded by developing a Web platform, especially for energy providers, where they can visualize the users' consumption profiles, as well as changes in behavior based on pricing, season, or other stimuli tested. Suppliers can also obtain estimates based on consumer data to manage their business effectively.

European policies in the energy field, technological developments, and the need to control electric energy usage highlight the importance of such an application.

8 Acknowledgments

This paper presents the scientific results of the project "Intelligent system for trading on wholesale electricity market" (SMARTRADE), co-financed by the European Regional Development Fund (ERDF), through the Competitiveness Operational Programme (COP) 2014-2020, priority axis 1 – Research, technological development and innovation (RD&I) to support economic competitiveness and business development, Action 1.1.4 - Attracting high-level personnel from abroad in order to enhance the RD capacity, contract ID P_37_418, no. 62/05.09.2016, beneficiary: The Bucharest University of Economic Studies.

References

[1] European Comission, "Eurostat," European Comission, 17 08 2018. [Online]. Available: https://ec.europa.eu/eurostat/web/energy/data/main-tables. [Accessed 17 06 2019].

[2] R. S. Brewer, Fostering sustained energy behavior change and increasing energy literacy in a student housing energy challenge, Manoa: ProQuest LLC, 2013.

[3] B. Karlin and J. F. Zinger, "The Effects of Feedback on Energy Conservation: A Meta-Analysis," *Psychological Bulletin*, vol. 141, no. 6, pp. 1205-1227, 2015.

[4] S. Darby, "Smart metering: what potential for householder engagement?," *Building Research & Information*, vol. 38, no. 5, pp. 442-457, 2010.

[5] S. V. Oprea, A. Bâra, A. I. Uță, A. Pîrjan and G. Căruțașu, "Analyses of Distributed Generation and Storage Effect on the Electricity Consumption Curve in the Smart Grid Context," *Sustainability*, vol. 10, pp. 2264-2289, 2018.

[6] S. V. Oprea, A. Bâra and G. Ifrim, "Flattening the electricity consumption peak and reducing the electricity payment for residential consumers in the context of smart grid by means of shifting optimization algorithm," *Computers & Industrial Engineering*, vol. 122, pp. 125-139, 2018.

[7] T. Peters, "PEP 20 -- The Zen of Python," 19 August 2004. [Online]. Available: https://www.python.org/dev/peps/pep-0020/.

[8] Sun Microsystems, Inc., "JavaSoft Ships Java 1.0," Sun Microsystems, Inc., 23 01 1996. [Online]. Available: https://tech-insider.org/java/research/1996/0123.html. [Accessed 18 05 2019].

[9] Google, Inc., "Distribution dashboard," Google, Inc., 07 05 2019. [Online]. Available: https://developer.android.com/about/dashboards. [Accessed 18 05 2019].

Maria Irène CĂLINOIU (b. August 9, 1997) is a graduate of the Faculty of Economic Cybernetics, Statistics, and Informatics at the Bucharest University of Economic Studies, under the tuition of PhD. Lecturer Simona Vasilica Oprea. She graduated “Tudor Vianu” National High School of Computer Science in Bucharest in 2016. Her scientific fields of interest include Programming, Databases, Data Science, Natural Language Processing, and Machine Learning.

Simona-Vasilica OPREA (b. July 14, 1978) received the MSc degree through the Infrastructure Management Program from Yokohama National University, Japan, in 2007, the first Ph.D. degree in Power System Engineering from the Bucharest Polytechnic University in 2009, and the second Ph.D. degree in Economic Informatics from the Bucharest University of Economic Studies in 2017. She is involved in several national and H2020 European research projects as member or project manager. She is currently project director for a H2020 project entitled Multi-layer aggregator solutions to facilitate optimum demand response and grid flexibility (acronym SMART-MLA).

Trading Fragmentation Methodology to Reduce the Capital Exposure with Algorithmic Trading

Cristian PĂUNA
The Bucharest University of Economic Studies, Romania
cristian.pauna@ie.ase.ro

This paper presents a practical methodology to reduce the capital exposure by early exits from the financial markets using algorithmic trading. The method called trading fragmentation uses several automated trading software applied on more unrelated markets and a particular risk management strategy to obtain a higher profit level with a lower risk. An advanced capital management procedure is used to integrate all into an unitary risk management system applied into a single trading account. It was found that the method presented here is the proper way to avoid large loss trades and to reduce the time when the capital is blocked into negative positions for the recovery process. In this way the efficiency of the capital usage is improved and the profit is made faster with lower risk level. The method was tested with real capital for more than five years and positive results were obtained. Comparative trading numbers will be also included in this paper in order to reveal the efficiency and the advantages obtained with the trading fragmentation methodology.
Keywords: *algorithmic trading, capital exposure, risk management, trading fragmentation*

1 Introduction

In the context of the electronic trading release and implementation all over the world in financial markets, algorithmic trading has become a major research interest theme nowadays. "With rapid advances in technology, enterprises today frequently search for new ways to establish value positions" [1]. Integrating the automated trading software in the business intelligence systems of the financial investment companies is a trend and a necessity today.

Modern methodologies to automate the trading decisions and the risk management offer more and more advanced solutions and advantages in order to make profit. "Business intelligence is the result of the natural evolution in time of decision support systems and expert systems, systems that aimed at replacing humans in the decision making process or, at least, at offering solutions to the issues they are concerned of" [2].

"The real-time data analysis for prediction and risk management in the electronic trading systems place the automated trading systems to be the main engine in the business intelligence system of a financial or investment company" [3]. The design and implementation of any algorithmic trading strategy into an automated software starts from the principles of the manual trading activity.

There are two major question marks in the trading activity: when to entry on the market and how to exit from the opened trades in order to obtain the desired profit with a minimal risk level.

For the first question there are a higher number of researches and studies testing and developing trading strategies with positive edge to locate the right moment to buy equities on financial market. Using the computers to process the price evolution in time, with the right data mining process applied for the time price series, the entry in the markets are built as trading signals and can be automatically executed by the trading software. This will not be the subject of this particular paper.

How to exit from the market in order to reduce the capital exposure is a subject less treated in the academic literature today. This article will present a methodology in order to exit from the markets. The new approach called "trading fragmentation" will permit to

decrease the capital exposure and to obtain the desired profit faster and with a lower effective capital risk. The paper will present the basics for this methodology that can be applied in any financial markets for any entry used strategy. One strong point is that this exit method can be completed automated in order to be integrated into an automated trading system for any business intelligence system applied for any investor or company type.

2 Classical exit methods

In this chapter it will be presented on short several classical known exit strategies to close a trade and to exit from the market. Once a trade was opened, the usual exit method is to close that trade when the profit is equal with a specified value a priori established before to open that trade. This is the most common exit method used in algorithmic trading and especially in high-frequency trading systems. The fixed target level (FTL) method sets the take profit level from the beginning and wait the market to evolve until the price reaches that price target. The method is very simple and good to test and optimize any trading strategy. For a specified profit target level, all functional parameters of the trading strategy will be optimized in order to maximize the profit level and to minimize the capital exposure. It was found that the third optimization criterion is "the Longest Time Trade period" [4] (LTT). "This factor makes the difference between trading and investment" [4] and is the main indicator to establish how long the capital is blocked into the recovery process of a trading strategy.

Other trading strategies use different exit methods become classical because they are used since years. As example it will be mentioned here the "Fixed Time Exit Strategy" (FTE) which "tells us to exit when a certain amount of time has passed" [5], "First Up Close Exit Strategy" (FUC) meaning to exit after the market "has its first up close versus the previous day" [5], "New High Exit Strategy" (NHE) is about to exit the position on the market "after it closes above a new high" [5], "Close above the Moving Average" (AMA) method that indicate simply to close the trade dynamically after the market "has closed above its simple moving average" [5] and "2-Period RSI Exit" (2PR) which indicate to exit the trade when the 2-Period RSI has a value higher than a specified limit value (65% or 70% according to the source [5]).

Other known exit method, applied especially for those trades that were opened using a trend indicator, is to close that trade when the trend indicator used tells us that the trend is no longer exists or the power of the trend is decreasing. Methodologies to measure the power of the trend are presented in [6] and [7] and can generate good exit signals.

With all of these exit methodologies any trading strategy can be optimized in order to have the proper functional parameters set to obtain the desired profit. It was found that, using several automated trading software in a single trading account, none of these exit methods assure the optimal solution. Each exit methods mentioned above can produce a solution for any trading strategy but all of them generate two types of trades. With a proper optimization the most of the trades can touch the exit criteria in a short period of time. Because there is no perfect trading method, each exit strategy will generate also a small number of trades that will last a longer period of time. In these cases, after the trade was opened, the market reversed and evolved against the direction of the open trade. Even the trade will be closed on profit, this process can last sometimes hundreds or thousand of trading hours. This is a case that must be avoided. In all this time the capital is blocked in that particular trade and cannot be used to make new profits. These cases reduce drastically the efficiency of the capital usage. The exit strategy presented later in this paper will solve the problem and will gives a method for capital efficiency optimization.

3 Diversification in financial trading

To trade on only one financial market is not a proper solution. To optimize the capital usage several markets must be traded in the same time. This is called diversification. It is recommended "diversifying into a minimum of three unrelated markets. At any point of time, one market might be in a major uptrend and one might be in a major downtrend, with a third trendless. The odds of catching major moves increase with the number of markets traded. One caveat: they should be unrelated markets" [8].

In the actual conditions of the high price volatility, "the price variations per time unit have become extremely fast and in order to capture and optimally use the price differences it is necessary a fast calculation for the buy and sell decisions and at the same time the transmission of these orders as fast as possible to the execution" [9]. To ensure the low-latency condition, each market will be traded by particular trading software using a particular set of trading strategies with particular optimization sets for the functional parameters.

In this point is obviously that we deal with "software on demand". "In the present time, when companies' businesses are growing more and more, the software developers may adapt to a change from the industry direction and must continuously analyze and optimize current solutions. By developing new strategies to automate services, the architects and developers contribute to more flexible and efficient solutions that provide support for business integration and agility" [10].

With all of these the design direction for the trading system is clear: we need to trade several unrelated markets to ensure the diversification with one special software especially designed and optimized for each of these markets to ensure the low-latency condition.

When it is about the exit decisions, the process to design, develop, test and optimize the trading software includes one of the classical exit strategies presented in chapter 2 of this paper. Usually the FTL exit methodology is a good and simple choice for this step. With this, the trade profit will be calculated using the natural formula:

$$P_t = V(p_t - p_e) \qquad (1)$$

where V is the traded volume, p_t is the target price and p_e is the entry price. Once a trade is opened, the V and p_e are known. For each P_t profit level wanted, the target price p_t can be calculated from the formula (1). Starting from these considerations, the exit decision can be automated using the exit signal for each i trade using the Boolean variables:

$$Exit_i = (p_i \geq p_t) \quad or \quad Exit_i = (P_i > P_t) \qquad (2)$$

where p_i is the current price level and P_i is the current profit for each trade. The trading software will continuously compare the current price level with the target level and will close the trades when the exit condition (2) is met. This is the classical FTL exit condition transposed in the Boolean variable in order to automate the exit decision. In the formula (2) it was included a second form for the exit decision variable to prepare the next considerations.

Once the diversification and the low-latency conditions were met in order to have a reliable and optimal trading system, using several automated trading software will not assure the capital is well used. Even each trading strategy is optimized longer trades will be always present in the trading reports. The practice shows us a drastically reduction of the trading efficiency for those trading strategies optimized to make very shorter trades. A case study was presented in [9]. By filtering the data mining processes to obtain only very short trades, "an algorithm has produced a profit of one hundred times smaller than its counterpart that performs ten times longer transactions" [9]. A new method was searched in order to prevent longer trades for the same profit level. This will be presented in the next chapter.

4 The trading fragmentation method

The trading fragmentation exit method starts from the principle "run your profits, cut your losses" [11], which is the most important principle in the trading psychology in the current paper author's opinion.

Before to define the method let's see how the trading activity is evolving when we use several trading systems in order to trade on more unrelated markets. All software will open one or more trades. The most of these trades will ran in the direction of the trends and will be close on profit in a short period of time after they were opened. Some of trades, fewer, will be on negative amount. These are the cases when the market changed the direction and goes against the trading methodology. Almost all trading strategies generate losing trades. Some of these trades will be closed on profit after the market will reverse again and will recover all that losses. Other part of these trades will be close on loss, after a significant number of trading hours, when the loss became higher enough to touch the stop loss used by each trading software. All of this process will block a part of the trading capital and will reduce the trading efficiency.

The trading fragmentation principles consist to close all opened trades once a specified small profit level was achieved into the trading account. Using a several number of trading strategies on more unrelated markets will generate a high number of small profitable trades and a low number of losing trades. Instead to wait to recover all that negative trades, once a specified profit level was achieved into the trading account, all the negative trades will be closed. In this way, the capital blocked in those losing trades will be released and will be involved by other trading strategies into new trades in order to make profit.

The profit level when all trades will be closed is called profit fragment and usual has a small value between 0.1% and 2% of the traded capital. The exit condition with the trading fragmentation method can be also automated using the formula:

$$Exit_i = \left(\sum_{j=1}^{N} P_j + \sum_{k=1}^{M} P_k > \xi \right) \qquad (3)$$

Pj is the profit value realized for all closed trades using the formula (2), M is the total number of the closed trades, Pj is the current profit of the open trades, N is the total number of opened trades and ξ is the profit fragment.

With other words we do not start the trading software and let them to open trades continuously and follow exits only by formula (2). Time to time, setting a profit fragment ξ, we will close all opened trades to avoid large period for loss accumulation. This will methodology will reduce the longest time trade for the entire trading system and will improve the capital efficiency usage involving the blocked capital into new trades.

A very important observation is that the trading fragmentation methodology works only if the traded markets are unrelated. This is because the loss trades closed at the moment when the ξ profit level is achieved are practically recovered by other positive trades made by other trading software in the same trading account. To be sure if two markets are unrelated it can be calculated the Galton-Paerson correlation coefficient [12] between the two time price series. If this coefficient is closed to zero, we will have unrelated markets.

The trading fragmentation methodology can be applied for correlated markets also, if the trading strategies involved in each trading system are different and use totally different principles. What important is, is to have a good coverage of positive trades to cancel the losses of the wrong trades.

There is not a general prediction for the valued of the profit fragment ξ. This is a functional parameter that must be optimized for each trading system.

5 Comparative results

In order to reveal the efficiency of the trading fragmentation methodology we present the next trading results obtained with the automated trading system theServer [13]. The system traded 12 financial markets using 20 trading software components which includes 60 different trading strategies optimized for each market. Each strategy has its own FTL level and all trades are individually closed using the (2) exit formulas. The risk management strategy used the "Global Stop Loss" methodology presented in [14] with a maximal capital exposure level of 1% for each trading software and with a 10% maximal capital exposure for the whole trading system.

The results presented below were obtained in a real time trading test organized in three different accounts. In one account the trading fragmentation methodology was not enabled (Table 1). In the second account the trading fragmentation methodology was applied with a profit fragment ξ=0.5% of the trading capital (Table 2). In the third account the trading fragmentation exit methodology was applied with a profit fragment ξ=1% of the trading capital (Table 3).

All accounts were hosted by the same brokerage company, they all had the same leverages, commissions and slippage taxes and all accounts had the same trading capital amount on the beginning. All the functional parameters of the trading system were the same and the latency of the price time series was the same. The purpose of this test was only to reveal the influence of the trading fragmentation methodology. For this reason, the only one parameter that was different between the three trading accounts used was the profit fragment value. The test was intentionally closed at a limited date, before touching the profit target in one of the three accounts in order to see the differences. Here are the trading results obtained:

Table. 1. Trading results obtained without trading fragmentation exit methodology

Start trading	01.01.2018
Stop trading	30.06.2018
Initial capital	50,000 euro
Profit target	50%
Profit fragment	not applied
Number of trades	634 trades
Total profit	16,423 euro
Longest trade	1,422 hours
Max. drawdown	4,842 euro
Abs. drawdown	2,104 euro
Maximal RRR	1:3.39
Absolute RRR	1:7.81

Table. 2. Trading results obtained with trading fragmentation for ξ=0.5%

Start trading	01.01.2018
Stop trading	30.06.2018
Initial capital	50,000 euro
Profit target	50%
Profit fragment	ξ=0.5%
Number of trades	852 trades
Total profit	18,861 euro
Longest trade	147 hours
Max. drawdown	2,214 euro
Abs. drawdown	811 euro
Maximal RRR	1:8.51
Absolute RRR	1:23.25

Table. 3. Trading results obtained with trading fragmentation for ξ=1.0%

Start trading	01.01.2018
Stop trading	30.06.2018
Initial capital	50,000 euro
Profit target	50%
Profit fragment	ξ=1.0%
Number of trades	784 trades
Total profit	19,283 euro
Longest trade	2097 hours
Max. drawdown	2,633 euro
Abs. drawdown	811 euro
Maximal RRR	1:7.32
Absolute RRR	1:23.77

The RRR is the risk to reward ratio. In the tables above were highlighted with red the negative part of the initial trading system (without fragmentation) and with blue the positive advantages obtained using the trading fragmentation methodology.

6 Conclusions

Looking at the Table 1, the negative factor that should be corrected is the longest time trade. Even the trading strategies included in each automated trading software were optimized in order to maximize the profit and to reduce the capital exposure level, the longest trade periods is still important. 1,422 hours means about 60 days and this time is due to the usage of six automated investment software included in theServer. In all this time a capital stake is blocked into the negative trade that it is waiting to be recovered by the trading software. In all this period that capital will produce nothing, and even that trade will be closed on profit, the efficiency of that capital will be low.

Looking to the results from the Table 2 we can see that using the trading fragmentation model, the maximal trade period was reduced by almost ten times. In this case, with ξ=0.5%, the longest time trade was 147 hours instead 1,422 hours in the first case. Using the new exit methodology those negative trades from the first case were covered by some positive trades made by other trading software. In this way the capital was blocked for a shorter period of time and a 14,84% supplementary profit stake was obtained. In addition, the maximal and the absolute drawdown were improved and by consequences the RRRs obtained numbers are significantly better.

Thinking to the trading fragmentation methodology we can suppose that the best case is to use a smallest ξ, but this is not the true case. Looking to the Table 3. we can see that using a ξ=1.0% a better profit level was obtained. The explanation is that with a larger ξ, some of the losing trades were recovered in time and produced a supplementary part of the profit. For the case with ξ=1.0% the profit is with 2,23% higher than the case with ξ=0.5% and with 17,40% higher than the case without fragmentation.

A first conclusion is that the trading fragmentation exit methodology is a significant improvement factor for an advanced automated trading system.

An important notice is about the profit fragment level. Even in the numbers presented above we have seen that the case with ξ=1.0% is the better case, this conclusion is good only for the trading system used as example. The profit fragment level must be optimized for each trading system. This factor depends on the trading strategies used, on the profit level used for each individual exit methodology by each trading software and it depends also on the capital exposure level allocated for each trading strategy. Further researches indicate that the ξ level depends also on the spreads, commissions and slippage taxes asked by the brokerage company.

Another factor hard to be predicted without tests is that the ξ level depend on the markets used to trade, on the correlation coefficient between those markets together and on the independence between the trading strategies used. With all of these considerations the ξ level must be optimized for each trading software and for each financial market used by the system.

Another important factor is that the ξ level cannot be optimized using simulation or back tests. We have the methodologies to simulate the functionality of any trading system but to simulate 10 or 20 trading software together is a very hard that can ask a huge computing power and time. To avoid these complications trading tests like those presented in the chapter 5 can be made using different valued for the ξ level.

Together with the “Global Stop Loss” methodology presented in [4], the trading fragmentation method is the best method I have found in my research to reduce the capital exposure and to increase the trading efficiency using algorithmic trading in any advanced automated trading system that trade in several unrelated markets.

References

[1] A. Bâra, I. Botha, V. Diaconița, I. Lungu, A. Velicanu and M. Velicanu, A model for Business Intelligence Systems' Development, Informatica Economică Journal, vol. 13, no. 4/2009. ISSN: 1453-1305

[2] A. Bologa and R. Bologa, *Business Intelligence using Software Agents*, Database Systems Journal vol. II, no. 4/2011. ISSN: 2069-3230

[3] C. Păuna, Automated Trading Software - Design and Integration in Business Intelligence Systems, Database Systems Journal vol. IX, 2018. ISSN: 2069-3230

[4] C. Păuna, *Capital and Risk Management for Automated Trading Systems*, Proceedings of the International Conference on Informatics in Economy, May 2018., Iași, Romania. Available at: https://pauna.biz/ideas

[5] L. Connors and C. Alvarez, *Short Term Trading Strategies That Work. A Quantitative Guide to Trading Stocks and ETFs*, TradingMarkets Publishing Group, 2009, ISBN: 978-0-9819239-0-1. pp. 112.

[6] C. Păuna, *Trend Detection with Trigonometric Interpolation for Algorithmic Trading*, Scientific Annals of Economics and Business, ISSN: 2501-1960

[7] C. Păuna, *A Price Prediction Model for Algorithmic Trading*, Romanian Journal of Information Science and Technology, ISSN: 1453-8245

[8] G. Kleinman, *The new commodity trading guide. Breakthrough Strategies for Capturing Market Profits*, FT Press, 2009, ISBN:0-13-714529-2. pp. 126.

[9] C. Păuna, The psychology of the automated decision-making algorithms usage in the financial information systems, Revista de Psihologie, ISSN: 0034-8759

[10] R. Zota and L. Ciovica, *Designing software solutions using business processes*, Proceedings of the 7th International Conference on Globalization and Higher Education in Economics and Business Administration, GEBA 2013. Published by Elsevier B.V. ISSN: 2212-5671. doi:10.1016/S2212-5671(15)00125-2

[11] S. Ward, *High Performance Trading. 35 Practical Strategies and Techniques to Enhance Your Trading Psychology and Performance*, Harimann Hours, 2009, ISBN: 978-1-905641-61-1. pp. 125.

[12] I. Purcaru, *Informație și corelație*, Editura Științifică și Enciclopedică, 1988, pp. 91.

[13] C. Păuna, theServer automated trading system online presentation, 2015. Available at: https://pauna.biz/theserver

[14] C. Păuna, *Capital and Risk Management for Automated Trading Systems*, Proceedings of the 17th International Conference on Informatics in Economy, 2018, pp 183-188. Available at: https://pauna.biz/ideas

Cristian PĂUNA graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Economic Studies Academy in 1999 and he is also a graduate of the Aircraft Faculty from the Bucharest Polytechnic University in 1995. He got the title of Master of Science in Special Aerospace Engineering in 1996. In the last decades he had a sustained activity in the software development industry, especially applied in the financial trading domain. Based on several original mathematical algorithms, he is the author of more automated trading software for financial markets. At present he is the Principal Software Developer of Algo Trading Service Ltd. and he is involved as PhD student in the Economic Informatics Doctoral School from the Economic Studies Academy.

Factors that contribute programming skill and CGPA as a CS graduate: Mining Educational Data

Md Aref BILLAH, Sheikh Arif AHMED, Shahidul Islam KHAN
Department of Computer Science and Engineering (CSE)
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh
mdarefbillah818@gmail.com, sheikharif1993@gmail.com, nayeemkh@gmail.com

Computer Science (CS) has become one of the most popular under graduate program in last few years. According to UGC roughly 116 universities out of 136 are offering computer science program which indicates a massive number of students are choosing this program as their undergraduate program. But statistically significant number of students are failing to become skilled and effective CS graduate because many students are taking CS without accessing their chance in this program. Success in academic and professional life require to choose right under graduate program. Considering CGPA and Programming Skill as two of the most significant factors to determine student's success in CS, we have predicted these two by taking students personal interest, academic results, analytical skill and problem solving skill into account. We also extracted most significant features of a prospective CS student by using gain ratio.

***Keywords**: Computer Science Student, Predicting Performance, Machine Learning Techniques, Data Mining, Programming skill, CGPA*

1 Introduction

Computer Science has now become a buzzword in the global community. Being one of the developing countries Bangladesh Government has already taken the challenge of outshining in the ICT department, so as the students. But when the question of skill, show casing talent and achievements in national and international level comes it seems significantly important percent of students are failing to do so. Without proper analysis, substantial amount of students are taking this program and eventually the performance of larger part of students of this Program is performing poorly which is hurting their

Academic and professional life. Though massive number of students are rushing into this program in Bangladesh, many IT industries are still hiring IT professionals from India because of limited number of skilled graduates [1].

So, to predicting the performance of prospective CS graduates before they start is what they need. If they can know the factors on what their performance as a CS graduate depends, they can decide whether they are going to take CS or not. There is a possibility that they can change themselves as the demands to be a good CS graduate.

In recent past, Students' final result predictor and Classification model for determining students' future was built by taking different types of feature sets like as previous academic result, family income, family expenses, medium of teaching, marital status, parent's occupation, parent's qualification, family size, attendance, assignment, lab work: [2],[3],[4],[5],[6],[7]. Different types of classification model were built by implementing Decision tree, Support Vector Machine and Naive Bayes Classifier algorithm on students' academic results: [2], [3].

Purpose of this work is to build a model to predict students' final result or we can say CGPA of their CS program and also the programming skill which is very good indicator of a good CS graduate before they start.

To predict students' performance in any program it is important to take the previous

academic results into account [8]. Studies show that CS students' need mathematical skill for increasing programming and other skills like problem solving and analytical skills: [9], [10]. So besides academic results, to make our proposed model more accurate in this paper we used their previous academic results, their experience with ICT course at HSC level, online class experience, Internet browsing reasons, Participation in Mathematics or Science Olympiad, Interest in Competitive programming and students problem solving skills etc. To train our model we have used current CS student's data and when our model is ready we predicted both CGPA and programming skill. Where CGPA tells about student's theoretical knowledge and programming skill tells about their practical knowledge.

2 Literature Review

Many researcher worked for predicting the performance of students.

Romero et al. (2013) conducted a study on students participating in on-line discussion forum and predicted students' final performance in Spain. They collected forum interaction data such as number of messages post/read, ask and reply relationship between students. Afterwards compared them in between classification and classification via clustering approach. [11]

Alharbi et al. (2016) did a case study whether they can highlight performance problems early on and propose remedial actions using data mining techniques. They collected students data during admission and after completing their academic first year and eventually predicted good honors outcomes with reasonable accuracy by using classifying model with highlighting students that are predicted to low achievers with high probability module results [12].

Baradwaj & Pal (2012) suggested a classification model for Predicting Performance improvement on the Educational databases which contains invisible information for improvement of students' performance. They collected 300(74 females, 226 males) students record from Dr. R. M. L. Awadh University, Fazibad India and used Bayes classification approach. They further conducted a research by taking students class test, assignments, attendance, lab work and seminars into consideration and analyzed students' performance in the semester final examination. They used decision tree for predicting students' performance: [2], [3].

Daud et al. (2017) conducted a study where using Educational Data mining approach they considered student's family expenditure, personal information and predicted whether he will be able to complete his degree or not. They used WEKA tool to classify 100 students record with 23 features each from different universities of Pakistan [4]

Goga et al. (2015) proposed a tool by using .Net framework which takes students various information as input and predicts students' grade. They first collected the student's enrollment records from Babcock University, Nigeria and then built models using classification trees and a multi-layer perception learning algorithms operating on WEKA. In the domain of this study random tree adopted as the best algorithm and served as a building block of this generic system [13]

Arsad & Buniyamin (2013) conducted a study to predict student's final result at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia. They showed Artificial Neural Network model served as a vital to predict students result. They found that Students' first and thirds semesters fundamental course results reflected their final result. Therefore, fundamental subjects should be fully understood because without clearing fundamental knowledge it is very difficult to face advance courses [14].

Huang & Fang (2012) proposed a mathematical model for making early prediction of students' final score and Engineering dynamic courses using the results of first semester and validated using next three semester data. They collected data from 1900 engineering students of four

semesters. Four different mathematical modeling approaches: radial basis function neural networks, multi-layer perception neural networks, multivariate linear regression, multi-layer perception neural networks, and support vector machines were applied and experimental results showed anyone can be used to predict the desired result [15].

3 Techniques

We have used few ***data miming techniques*** and existing algorithms and tested their efficiency through various techniques. Here is a short description of these tools and techniques

The term Data mining is a misnomer, as it refers to extracting knowledge from large amount of data [16]. The data mining techniques is featured to create model which will help to find new data using unknown data [17]. Data mining can be basically of two types, Predictive and Descriptive [18].

Predictive Data Mining Model

This method studies previous historical data and predict and forecast what is going to happen to future data set e.g.: classification, regression, time series analysis etc. [19].

C 4.5: It's a decision tree based algorithm for classification both numeric and nominal classes. It was written by J. R. Quinlan [20].

Support Vector Machine: Support vector machine is the supervised algorithm for using as a classifier or regression algorithm for pattern, nested problems or mining of texts [21]. It uses a hyper plane to partition two different classes where support vectors are those which instances are used as the margin [22].

Descriptive Data Mining Model

Descriptive data model is a predictive model consisting of clustering, summarization, association rule etc. It finds pattern in large data and further works in intelligence system decision-making.

Data mining is of two types according to its class i.e. supervised and unsupervised learning algorithms.

Unsupervised algorithms: No information about class or class label. E.g.: Clustering and association rules etc.

Supervised algorithms: Class data is known. E.g.: classification, regression etc.

4 Methodology

Researchers from different regions found that study on forecasting student's final result or CGPA for under graduation program is very important. They considered students' academic result, parents data, hours spent in study, activity in online discussion forum, marital status & student's class attendance for forecasting their final performance and building classification model which can help them take decision which track they should follow for the under graduation program. Different from the literature, we considered only Computer Science program as our purpose of study for its immeasurable popularity and the kind of challenges it encounters. Especially, for someone who is not ready to take those challenges is sure to suffocate in CS. For CS programming skill measurement is also very important. So we build a model to predict CGPA and programming skill using different regression algorithm to build a prediction model based on 23 features including their academic results, experience with ICT course, personal interests, personal experiences of current students' with similar academic achievements and problem solving skills.

Our proposed method has few stages-

1. Data collection
2. Pre-processing
   - Data cleaning
   - Transformation
   - Integration
   - Standardization
   - Feature selection
3. Data mining, model generation and Performance measurement of algorithms
4. Finally we will get a model to use.

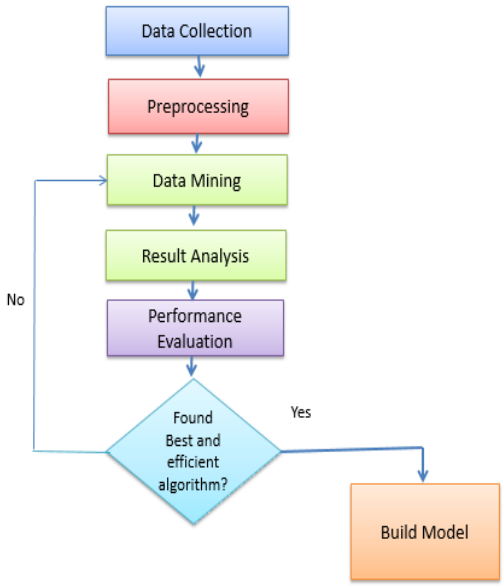


Fig. 1. Overview of Model. It shows the whole process of building our model step by step. Starts with data collection and ends with building a model.

5 Implementation and Result

Primary data was collected by doing a survey on current students of Computer Science. We collected CS student's data having 23 features in these six following categories. i.e.: Personal information, personal Interest, academic results, their experience with ICT course in HSC level, level of their problem solving skill and their online course experience. We assigned numeric values to these questions responses for data analysis and research purpose. We have removed inconsistency from our data set and transformed them into the proper format to use them for analysis. **Table 1** Shows the options with correspondent values and short term of the survey questionnaire for students and here is the questionnaire-

Section A- Personal Information

A1- What is your name?

A2- Gender?

A3 -What is the Location of your University?

Section B- Academic result

B1- What is your SSC GPA out of 5.00?

B2- What is your HSC GPA?

Section C- Experiences with ICT Course

C1- How did you find ICT course (Your experience with ICT)?

C2- How was your academic result in ICT?

Section D- Personal Interest

D1-Why do you browse internet mostly?

D2- Participated in number of math or science Olympiad?

D3- Online courses you've followed related to your study?

D4- Rate your interest in Competitive Programming?

Section E- Personal Experience

E1- "Tendency of using online resources help a lot in Computer Science":

E2- "Having Patience help in Computer Science":

E3- "Knowing PC configuration helps in reading Computer Science":

E4- 'Computer gamer have better chance in CS':

E5- "Capability of Self-study makes significant difference in Computer Science":

E6- "HSC ICT Course result reflects once potential in Computer Science":

E7- Knowledge (about the CS program) you had before starting your Program?

Section F- Problem solving skills and others

F1- Rate your programming skill:

F2- Your skill in Basic mathematics (SSC level):

F3- Your skill in Higher Level Mathematics (HSC level):

F4- Rate your patience out of 10:

F5- Rate your capability of self-study out of 10:

Table 1. Options with correspondence values and short terms

Options	Short Term	Values
Strongly Agree/Very Good /5:00P/Many /Very Interesting /9 to 10 ratings/Social media and Entertainment	SA/VG/C3.8AA/5P/MT/VI/9-10R/FLP/SMAE	5

Agree/Good /4.50P and above/More than 5/Interesting/7 to 8 ratings/Study and Social Media	A/G/C3.4AA/4.5PAA/MT5/I/7-8R	4
Neutral/Satisfactory/ /4.00P and above/More than Once/5-6 ratings/, Reading Blogs only	N/S/C3.0AA/4.0PAA/MTO/5-6R	3
Disagree/Less than satisfactory /3.50P and above/Once/3-4 ratings/Online Courses only	D/LTS/C2.5 AA/3.5PAA/O/3-4R	2
Strongly disagree/Poor/ Less than 3.50P/Never/1-2 ratings/Online Course and Reading Blogs	SD/P/CLT2.5/LT3.5P/Ne/1-2R	1

Table 2. Students related variables that illustrates the questions we asked the students and probable answers.

Variables	Description	Possible Values
Gender	Students Gender	{Male, Female}
UniLoc	University Location	{Dhaka, Chittagong, Other City, Outside City}
GSSC	Students grade in SSC	{5.00 >4.50 and <5.00, >4.00 and <4.50, >3.5 and <4.00, >3.5}
GHSC	Students grade in HSC	{5.00 >4.50 and <5.00, >4.00 and <4.50, >3.5 and <4.00, >3.5}
ICTResult	Students ICT result in HSC	{Very Good, Good, Satisfactory, Less than satisfactory, poor}
ProRatings	Students programming skill ratings	{ 9 to 10,7 to 8, 5 to 6,3 to 4,1 to 2}
Bmath	Students skill in basic mathematics	{Very Good, Good, Satisfactory, Less than satisfactory, poor}
HMath	Students skill in Higher level mathematic	{Very Good, Good, Satisfactory, Less than satisfactory, poor}
PaRatings	Students Patience's ratings	{ 9 to 10,7 to 8, 5 to 6,3 to 4,1 to 2}
ICTExp	Students experience with ICT	{Very Interesting, Interesting, Neutral, Difficult, Very Difficult }
IBReason	Most important reason behind browsing internet	{Online courses and Reading Blogs, Online Courses only, Reading Blogs only, Study and Social Media, Social media and Entertainment}
MSOlym	Student participation in number of math or science	{many , less than 5, more than once, once, never}

	Olympiad	
NOC	Number of online courses followed	{many, less than 5, more than once, once, never}
Cpro	Students competitive programming Interest ratings	{ 9 to 10,7to 8, 5 to 6,3to 4,1 to 2}
SSR	Students capability of self-study ratings	{ 9 to 10,7to 8, 5 to 6, 3to 4,1 to 2}
ORH	Tendency of using online resources helps	{Strongly agree, agree, neutral, disagree, strongly disagree}
PH	Having Patience helps	{Strongly agree, agree, neutral, disagree, strongly disagree}
SSH	Self-study helps	{Strongly agree, agree, neutral, disagree, strongly disagree}
CGC	Computer gamers chance in Computer Science	{Strongly agree, agree, neutral, disagree, strongly disagree}
ICTRR	ICT result reflects success in CSE	{Strongly agree, agree, neutral, disagree, strongly disagree}
PC	Knowing about PC configuration helps in reading Computer Science?	{Strongly agree, agree, neutral, disagree, strongly disagree}

We have collected data from current CS student of various university from various city of Bangladesh. **Table 3** shows the frequency, percent valid percent and cumulative percent of dataset in respect to Gender

Table 3. Frequency Table (Gender of Students). 329 of 501 students are male and 172 is female. That is 65.7% students are male and 34.3% students are female.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	329	65.7	65.7	65.7
	Female	172	34.3	34.3	100.0
	Total	501	100.0	100.0	

Table 4. Programming skill * CGPA cross tabulation. We can see most of the good programmers (with rating 4 and 5) have CGPA 3.5-4.00 and out of 90 students with programming skill 5, 72 students have CGPA 3.7-4.0.

Programming skill * CGPA Cross tabulation							
		CGPA					Total
		2.49-bellow	2.5-2.99	3.0-3.49	3.5-3.69	3.7-4.00	
Programming skill	1	5	9	10	0	0	24
	2	0	36	63	15	3	117
	3	0	15	62	21	12	110
	4	0	0	42	**60**	**58**	160
	5	1	2	1	**14**	**72**	90
Total		6	62	178	110	145	501

5.1 Preprocessing

Often collected data is not understandable, inconsistent, lacking in important criteria or can contain various errors. Preprocessing makes data understandable by various process and solve those issues. Hence we have removed inconsistency from our data set and transformed them into the proper format to use them for analysis. Also selected important features to reduce the complexity of our process. For this we have used a filter method which is feature selection by gain ratio. Table-3 shows the comparison selected features-

Table 5. List of selected features. Here we can see for CGPA programming knowledge is an important factor. For programming skill self-study, patience and skill of higher level mathematics is most important.

Selected Features (For CGPA)	Gain Ratio	Selected Features (For CGPA)	Gain Ratio
ProRatings	0.474	SSR	0.279
PaRatings	0.465	PaRatings	0.268
GHSC	0.457	HMATH	0.255
SSR	0.452	Cpro	0.246
Cpro	0.448	MSOlym	0.224
GSSC	0.439	NOC	0.212
HMATH	0.439	CGPA	0.211
MSOlym	0.410	BMath	0.183
NOC	0.407	ICTResult	0.182
IBReason	0.404	IBReason	0.172
ICTResult	0.389	GHSC	0.168

5.2 Data Mining and Model Generation Using Various Algorithms

For predicting the CGPA and programming skill we have used several algorithms i.e. SVR, C4.5 etc. While predicting two standard evaluation metrics (MAE, RMSE) are used as cost function to evaluate the o performance of our prediction.

Mean Absolute Error (MAE): measures the difference between two continuous values. It uses absolute values and gives intuition (the "average error").

Root Mean Square Error (RMSE): refers to the standard deviation of the prediction errors. By using RMSE we can tell how concentrated the data is around the line of best fit.

Now we will see the results of regression algorithms for predicting the CGPA and programming skill in respect to these cost functions.

Table 6. Error evaluation for Regression (CGPA prediction)

Evaluation Criteria	**Multiple Linear Regression**	**C4.5**	**SVR (Linear kernel)**	**SVR (Poly kernel)**
MAE	0.1180	0.1555	0.1123	0.1409
RMSE	0.1712	0.2517	0.1604	0.2117

Table 7. Error Evaluation for Regression (Programming Skill Prediction)

Evaluation Criteria	**Multiple Linear Regression**	**C4.5**	**SVR (Linear kernel)**	**SVR (Poly kernel)**
MAE	0.5226	0.4242	0.5036	0.4596
RMSE	0.6899	0.7881	0.6880	0.6584

Taking all these results to account we can say MLR is best for predicting CGPA and SVR with linear kernel is best for predicting programming skill.

Suppose 'X' and 'Y' trying to figure out their eligibility in CS. **Table 8 shows** their data by important features from both programming skill and CGPA.

Table 8. Example of CGPA and Programming Skill prediction

Features	**Values**		
Name (Real Name not used)	X	Y	Z
Rating of programming skill (out of 10)	?	?	?
Number of participated math or science Olympiad	Never	More than once	Less than 5
Capability of self-study (out of 10)	3 to 4	5 to 6	9 to 10
Interest in Competitive Programming	1 to 2	3 to 4	5 to 6
Number of online courses followed	Never	Once	Many
CGPA	?	?	?
Rating of patience (out of 10)	3 to 4	3 to 4	9 to 10
Skill in Higher Level Mathematics	Poor	Satisfactory	Good
Reason behind internet browsing	Social Media Only	Social Media only	Online Study and Social Media
Skill in Basic mathematics	Satisfactory	Good	Very Good
Result in ICT	Satisfactory	Good	Good

Table 9. Predicted CGPA and programming skill of example data

Student	**CGPA**	**Programming Skill**
X	2.8	1 to 2
Y	3.3	3 to 4
Z	3.8	9 to 10

6 Conclusions

We can see CGPA and programming skill are connected with each other for most of the case. While selecting important features we have found few rules like students with good CGPA have good programming skill. For good programming skill students need patient, self-study, good skill in higher level mathematics etc. For predicting the CGPA multiple linear regression is the efficient one because it has the lowest error rate for both RMSE and MAE. And for predicting programming skill Support vector regression is more efficient than other algorithms. C4.5 gave less error but it was found over fitted for our dataset.

Contribution

The main contributions of this work are:

- We have discovered eleven most influential features to get success in CSE
- We have predicted student's final result (CGPA) and programming skill.
- We have discovered factors behind a good CGPA and Programming Skill

Future work

Other factors like course number of a specific course can be taken to predict the final result for the student of other discipline. Technology is changing day by day. Also the educational system. A lot more improvement can be done to this study. Study can be done taking data of other students from other important discipline like Bachelor of Medicine, pharmacy, Electronic engineering etc.

References

[1] Daily Industry Bangladesh becomes 4th largest remittance source for India. Retrieved from http://www.dailyindustry.news/bangladesh-becomes-4th-largest-remittance-source-india, 2 July 2018

[2] B.K. Baradwaj, P. Saurabh "Mining educational data to analyze students' performance." *arXiv preprint arXiv:1201.3417*, 2012

[3] T. Beaubouef "Why computer science students need math.", *ACM SIGCSE Bulletin*, 34, no. 4, 2002, pp. 57-59.

[4] A. Daud, A. Naif Radi, R. Ayaz Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi. "Predicting student performance using advanced learning analytics.", *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 415-421. International World Wide Web Conferences Steering Committee, 2017.

[5] M.Ramaswami, R. Bhaskaran. "A CHAID based performance prediction model in educational data mining." *arXiv preprint arXiv:1002.1144*, 2010

[6] M. Tair, M. Abu, A. M. El-Halees. "Mining educational data to improve students' performance: a case study.", *International Journal of Information*, 2, no. 2 , 2012, pp.140-146.

[7] E. Osmanbegović, M. Suljić. "Data mining approach for predicting student performance." *Economic Review*, 10, no. 1, 2012, pp. 3-12.

[8] Q.A. Al-Radaideh, E. M. Al-Shawakfa, M. I. Al-Najjar. "Mining student data using decision trees.", *International Arab Conference on Information Technology (ACIT'2006)*, Yarmouk University, Jordan. 2006.

[9] T. Beaubouef, R. Lucas, J. Howatt. "The UNLOCK system: enhancing problem solving skills in CS-1 students." *ACM SIGCSE Bulletin*, 33, no. 2, 2001, pp. 43-46.

[10] S. Kumar, Venkata Krishna, S. Padmapriya. "An efficient recommender system for predicting study track to students using data mining techniques." *International Journal of Advanced Research in Computer and Communication Engineering*, 3, no. 9, 2014, pp.7996-7999.

[11] C. Romero, M.I. López, J-M. Luna, S. Ventura. "Predicting students' final performance from participation in

on-line discussion forums." *Computers & Education*, 68, 2013, pp.458-472.

[12] Z. Alharbi, J. Cornford, L. Dolder, B. De La Iglesia. "Using data mining techniques to predict students at risk of poor performance.", *SAI Computing Conference (SAI)* , 2016.

[13] M. Goga, S. Kuyoro, N. Goga. "A recommender for improving the student academic performance." *Procedia-Social and Behavioral Sciences* 180, 2015, pp.1481-1488.

[14] P.M. Arsad, N. Buniyamin. "A neural network students' performance prediction model (NNSPPM).", *Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference*, pp. 1-5. IEEE, 2013.

[15] S. Huang, N. Fang. "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques.", *Frontiers in Education Conference* (FIE), 2012, pp. 1-2. IEEE, 2012.

[16] Y. Zhao, "*Data mining techniques*.", 2015.

[17] S. M.Ali, M. R. Tuteja, "*Data Mining Techniques"*, 2014.

[18] D.J. Hand, "Principles of data mining." *Drug safety*, 30, no. 7, 2007, pp. 621-622.

[19] J. Han, J. Pei, M. Kamber, "*Data mining: concepts and techniques*". Elsevier, 2011.

[20] J. R. Quinlan, "Induction of decision trees." *Machine learning*, 1, no. 1, 1986, pp.81-106.

[21] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, "Support vector machines." *IEEE Intelligent Systems and their applications*, 13, no. 4, 1998, pp.18-28.

[22] S. R. Gunn, "Support vector machines for classification and regression." *ISIS technical report*, 14, no. 1, 1998, pp.5-16.

Shahidul Islam Khan obtained his B.Sc. and M.Sc. Engineering Degree in Computer Science and Engineering (CSE) from Ahsanullah University of Science and Technology (AUST) and Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2003 and 2011. He is now a Ph.D. Candidate in the Department of CSE, BUET, which is the highest ranked technical university of Bangladesh. His current fields of research are Data Science, Database Systems, Machine Learning, Data Security & Privacy, and Health Informatics. He has more than thirty published papers in refereed journals and conferences. He is also an Associate Professor in the Dept. of CSE, International Islamic University Chittagong (IIUC), Bangladesh.

Sheikh Arif Ahmed obtained his B. Sc. Engineering Degree in Computer Science and Engineering (CSE) from International Islamic University Chittagong, Bangladesh in 2018. He also worked as an undergraduate Teaching Assistant (TA) in International Islamic University Chittagong. His current research fields are Data Science, Machine Learning, Data mining, Security and Privacy etc. He has two papers in International Conferences.

Big Data: Technologies and Software Products

Adriana BĂNUŢĂ, Cătălina DRAGOMIR, Mădălina GHEȚU
Bucharest University of Economic Studies, Bucharest
adrianabnta@gmail.com, catalina93.d@gmail.com, madalina.ghetu@gmail.com

The main tendency in technology leans towards huge amounts of data to be stored, analyzed and processed, in order to obtain valuable information on various topics. Regardless the domain of interest, storing data is always a must and it must be done in an efficient, secure and accessible way. Then, it can be used for statistics, studies or as training sets in the field of machine learning. The aim of this paper is to give a brief overview of the concept known as big data, as well as to present and compare the main technologies and software products used to store and manipulate this type of data.
Keywords: *Big Data, Hadoop, NoSQL, Traditional Database*

1 Introduction

Big data is a broad term, generally referring to data sets that cannot be processed in a more "traditional" manner (i.e. using relational databases), due to the voluminous data they have to store and process.

Storing capacities have evolved at a fast pace, as the technological context imposed that, needing more and more memory to save and process data. Going into the digital era, storage of data became a vital need of every organization and more advanced ways had to be found, for keeping up with both companies and people's needs. These needs include storing data of millions of users, or storing historical data to be analyzed for statistics and predictions.

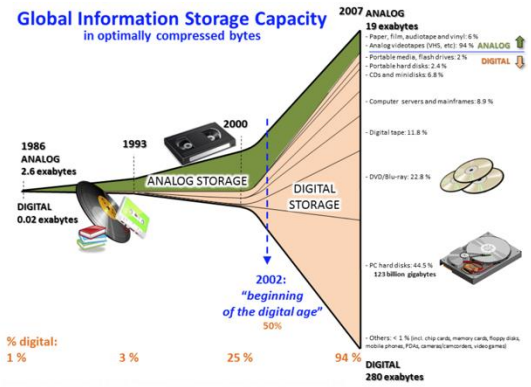


Fig. 1. Evolution of data

As shown in Fig. 1., digital storage can reach the order of Exabytes. Because of these "pretentious" specifications, the idea of big data became more and more popular and utilized.

This type of data relies on specific concepts, highlighted in Fig. 2., that come along with various challenges.

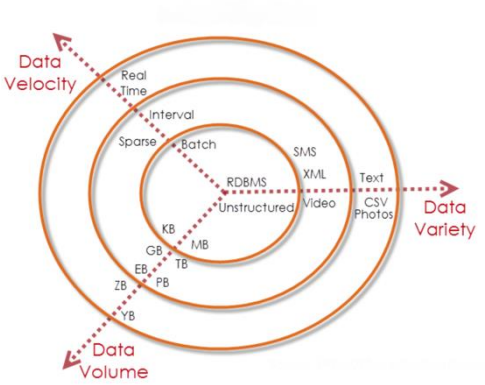


Fig. 2. Characteristics of big data

The most important one would be the volume, which means that large amounts of data can be processed at once, which brings an advantage when it comes to data analytics.

The next concept to be considered is velocity, referring to the rate at which data changed throughout an organization and being able to quickly use the available data for gaining different sorts of information can become a competitive advantage. For example, if online retailers can gather and

store a customer's history of navigation on the ecommerce site, they will be able to use that information for recommending additional purchases and this can work in their favor, inside a vying market [1].
The third most important concept refers to variety: data almost never has a homogenous form, perfectly ordered and ready to be used. This is why big data takes this raw, unprocessed form and extracts and orders its meaning, for it to be used further on, either by humans or as an input for algorithms or applications.

2 Big Data vs Traditional Databases

As shown in Table 1., there are many aspects to be considered when talking about the uses and differences of traditional databases and big data. Even if some criteria are specified, one cannot state that there is a better option when choosing the system for storing the needed data.

Table 1. Comparison between traditional databases and big data [2]

Characteristic	**Traditional Database**	**Big Data**
Data architecture	Centralized architecture, where problems are solved by a single computer, so it is costly and ineffective for large sets of data.	Distributed architecture in which the computations are done in a computer network, providing more power and improved performance.
Types of data	Structured data in fixed formats or fields that only provide insight at a small level.	Semi-structured or unstructured data, which allows the data to be gathered from a variety of sources and transformed into knowledge based information.
Volume of data	Small amount of data is stored, up to gigabytes.	Bigger amount of data, the order of petabytes.
Data schema	Fixed schema that is static and cannot be changed after it is saved.	Dynamic schema which is applied only after raw data is ready to be read.
Scaling	Very difficult to achieve, as it runs on only one server that would require a lot of power and generate high costs.	Scaled out architecture, in which the distributed computing principles employ more than one server.
Accuracy	Not all the data can be store, as this would be very expensive, so this decreases the amount of data to be analyzed, therefore the accuracy is also decreased.	Data is stored in big data systems, allowing huge amounts of data to be analyzed so the points of correlation are easily identified, providing high accurate results.

For both small and large applications that do not demand storing very big amounts of data, traditional databases are still the best option, as there are numerous things to do to optimize them and get the best out of them. It could even be considered a distributed architecture, that relies on the master-slave concept, so that data is processed fast, it is secure and can be easily recovered.
When talking about applications that store historical data, or that apply some concepts of data mining or machine learning, big data presents more advantages because of the accurate, varied data that it can store, but also because its reliable architecture and

dynamic schema.
It is obvious that both types of data storing come with their limitations and disadvantages and this is why, a thorough analysis should always be done, before choosing the best option for a specific case.

3 Big Data Technologies

When considering big data, the volumes of data that are used are way larger than the conventional ones, so powerful parallel processing in usually required. The specific architecture can be chosen regarding the needs of the application, considering which one of the three – volume, velocity, variety – is more relevant.
Cloud Dataflow is a native Google cloud data processing service integrated with simple programming model for both batch based and streaming data processing tasks.
This tool, which is a fully managed service handles the operational tasks including performance optimization and resource management. It also provides the possibility to manage the resources dynamically, in order to maintain high utilization efficiency while minimizing latency.
Cloud Dataflow provides a unified programming model method, so that programming model switching cost is no longer a issue. This method aids in batch and continuous stream processing, making it easy to express computational requirements without worrying about data source. [8]
To ease access to big stores of data, the concept of data lakes can come in handy. These data lakes represent vast data repositories that can collect data from various sources and store it in its raw, unprocessed state. Those repos differ from data warehouse, where even if data comes from different sources, it is processed and stored in a structured way
In this case, the lake and warehouse metaphors are fairly accurate. If data is like water, a data lake is natural and unfiltered like a body of water, while a data warehouse is more like a collection of water bottles stored on shelves [5].
In an enterprise, it may appear the case when they want to store data, but they are not sure yet how this data will be used, and this is when data lakes may be the best solution. For example, Internet of Things (IoT) data may be stored in data lakes, as it already has a big role in the growth and development of such storing solutions.
In the case if NoSQL databases, MongoDB is one that can be used for storing big data. Traditional relational databases cannot meet the current challenges posed by Big Data. Recently, databases of the NoSQL type are becoming more and more popular for storing large data. They have emerged from the need of companies like Google, Facebook or Twitter to manipulate huge amounts of data which traditional databases simply cannot handle.
NoSQL databases were designed to store very large volumes of data generally without a fixed scheme and partitioned on multiple servers. NoSQL databases offer flexible working modes, a simple API, and the possible consistency of a data. NoSQL databases thus become the core technology for Big Data. [9]
The main advantage of using NoSQL databases is that they allow efficient work with structured data, such as e-mail, multimedia, text processors. NoSQL databases can be seen as a new generation of databases: not relational, distributed, open source and characterized by horizontal scalability.
Another important feature of the NoSQL systems is the "shared nothing" architecture through which each server node is independent, does not shares memory or space. The architecture of a NoSQL database is shown in Fig. 4.

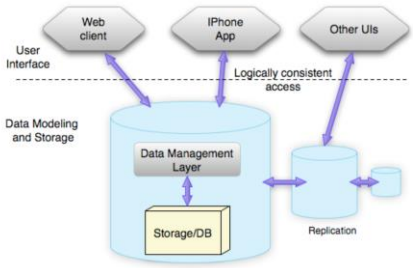


Fig. 3. NoSQL Architecture

3.1. Apache Frameworks

Hadoop

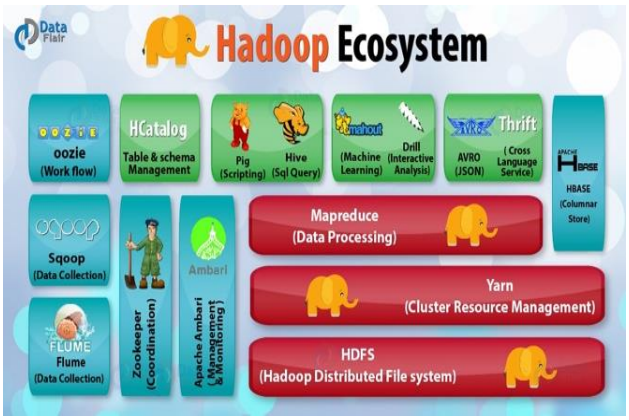


Fig. 4. Component scheme for Apache Hadoop

Apache Hadoop is an open source Java framework used for processing and querying big volume of data, on a set of large clusters. This project has been initiated by Yahoo! and its success is based on a large community of contributors for all over the world. Considering a significant technology investment done by Yahoo!, Hadoop has become a cloud computing technology ready to be used at an enterprise level and the most important framework when talking about big data.

It changes the economics and dynamics of large-scale computing, enabling scalable, fault-tolerant, flexible and cost-effective solutions [4].

Apache Hadoop consists of the following modules:

• **Hadoop Common**: contains libraries and tools needed for other Hadoop modules.

• **Hadoop YARN**: a resource management platform, responsible for cluster resource management and their use for user application planning

• **Hadoop MapReduce**: a programming model for large-scale data processing. It is named after the two basic operations that this module performs: reading data from the database, putting it in a suitable format for analysis (map-map), and performing mathematical calculations in a customer database (reduce), for example: counting men over 30 years of age.

• **Hadoop Distributed File System**: Allows you to store data in an easily accessible format in a large number of connected storage devices. A "file system" is the method used by a computer to store data so that it can be found and used. Normally, this is determined by the computer's operating system, yet a Hadoop system uses its own file system that is "above" the computer's file system itself - it can be accessed using any computer running any operating system accepted.

For data storing, Hadoop has its own distributed file system, HDFS, which makes the data available to multiple computing nodes. The typical Hadoop usage pattern involves three stages:

- Loading data into HDFS
- MapReduce operations
- Retrieving results form HDFS

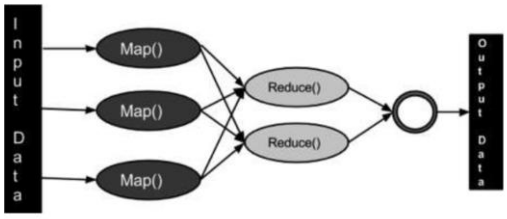


Fig. 5. MapReducer

As shown in Fig. 5., MapReducer algorithm performs two important tasks, Map and Reduce. Map is the task that takes converts an initial set of data into a new one, where elements are broken down into tuples (key/value pairs). Secondly, reduce task, takes the output from a map and uses it an

input in order to combine those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job [3].
Hadoop also offers solutions for making programming easier; working directly with Java APIs can be difficult and can expose the application to a lot of errors and restricts the usage of Hadoop to Java programmers only. Therefore, the other solutions that it offers are:

- Pig – a programming language that simplifies the common tasks of working with Hadoop: loading data, transformations, retrieving data.
- Hive – that enables Hadoop to operate as a data warehouse

Various other procedures, libraries or features have come to be considered part of the **Hadoop "framework"** in recent years, but Hadoop Distributed File System, Hadoop MapReduce, Hadoop Common, and Hadoop YARN are the most important.
The flexible nature of a Hadoop system refers to the ease of adding or modifying, for companies, the data system as their needs change, using cheap and easily accessible components from any IT service provider.
Today, this is the most widespread storage and processing system through computer groups - relatively cheap systems connected, as opposed to expensive, customized manual operations.
Some of the reasons why organizations use Hadoop is the ability to store, manage and analyze large amounts of structured and unstructured data quickly, reliably, flexibly and at low cost. The main benefits are:
• **Scalability and performance** - distributed data processing for each node in a cluster allows the company to store, manage, process, and analyze petabyte data.
• **Reliability** - Large clusters are prone to failure of individual nodes in the cluster. Hadoop is fundamentally resistant - when a node fails processing is redirected to the remaining functional nodes in the cluster and the data is automatically recreated for other failures that may follow.
• **Flexibility** - unlike traditional relational database management systems, structured schemes don't need to be created before data is stored. Data can be stored in any format, including semi-structured or unstructured formats.
• **Low cost** - unlike its own software, Hadoop is open source and runs on low-cost hardware groups.

Spark

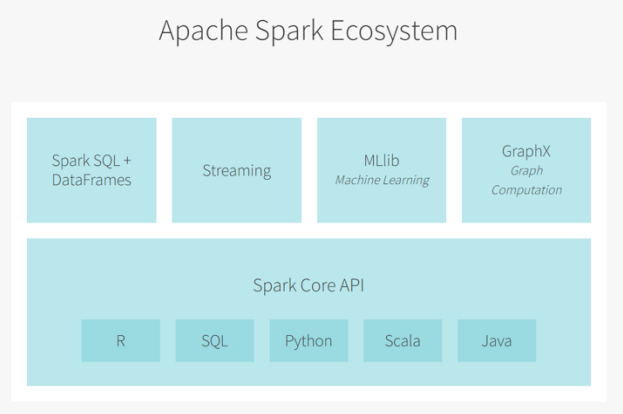


Fig. 6. Apache Spark component layout

Apache Spark is an open source cluster platform, an unified analysis engine for large-scale data processing.
Spark is seen by industry technicians as a more advanced product than Hadoop - it is newer and designed to work by processing data into pieces "in memory". This means that it transfers data from physical, magnetic hard drives to a much faster electronic memory, where processing can be done much faster - up to 100 times faster in some operations. It has proven to be very popular and is used by many large companies to store and analyze huge multi-petabyte data. This was partly because of its speed. Last year, Spark set a world record by completing a benchmark that included sorting **100 terabytes of data in 23 minutes**.

In addition, Spark has proven to be very suitable for Machine Learning applications. Machine learning is one of the growing and more exciting fields in computer science, where computers are taught to present data patterns and adapt their behaviour based on modelling and automated analysis of any task they are trying to achieve.

It is designed to be easy to install and use. To make it available to multiple companies, many vendors offer their own versions (as in the case of Hadoop) that are industry-specific or custom-tailored for individual client projects as well as associated consulting services to put them in function.

Spark uses cluster computers for its computational (analysis) power, as well as its storage. This means it can use resources from multiple computer processors connected for it analyses. With distributed storage, huge data sets gathered for Big Data analysis can be stored on multiple smaller physical disks.

It speeds up read / write operations, because the component that reads information on the disks has a smaller physical distance to migrate to the surface of the disk. Like processing power, you can add more storage space when needed, and using commonly available hardware groups (any standard hard disk for your computer) will help you maintain infrastructure costs.

The main benefits are:

• **Speed**: It can be 100 times faster than Hadoop for widespread data processing through computer memory and other optimizations. The spark is also fast when data is stored on the disk and currently holds the **world record for large-scale disk sorting**.

• **Easy to use:** has easy-to-use APIs to work on large data sets. This includes a collection of over 100 operators for data transformation and data APIs known to handle semi structured data.

• **A unified engine**: includes top-level libraries, including support for SQL queries, streaming data, automated learning and graphics processing. These standard libraries increase developer productivity and can be seamlessly combined to create complex workflows.

Ignite

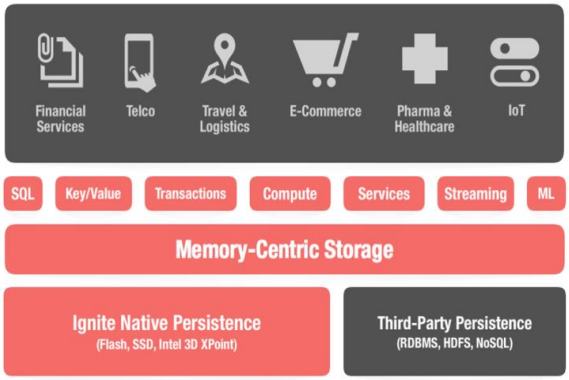


Fig. 7. Component Chart for Apache Ignite

Apache Ignite is a distributed open source, cache and processing platform designed to store and calculate large data volumes (petabyte scale) in a group of nodes. Provides a unified API that supports SQL, C ++, .NET, Java / Scala / Groovy, Node.js and more application-side access. The unified API connects cloud applications with multiple data warehouses that contain structured, semi-structured and unstructured data (SQL, NoSQL, Hadoop). It provides a high-performance data environment that allows companies to process all ACID transactions and generate valuable information from real-time and interactive queries.

Main Benefits:

• **Sustainable memory**: The durable memory component of the Ignite device treats RAM not only as a caching component but as a fully functional storage component. This means that users can turn persistence on or off as needed. If persistence is disabled, then Ignite can act as a database distributed in memory or in the data grid in memory, depending on the preference to use the SQL API or key. If persistence is enabled, then Ignite becomes a scalable distributed database that guarantees

complete data consistency and resists complete cluster errors.

• **ACID Compliance**: Data stored in Ignite is compatible with ACID both in memory and on disk, making the system consistent. Ignite operations work in the network and run on multiple servers.

• **Scalability and durability**: it's a scalable, horizontally distributed, system that accepts the addition and removal of cluster nodes on demand. Ignite also allows storage of multiple copies of data, making it resistant to partial cluster failures. If persistence is enabled, then the data stored in Ignite will not be affected by errors, they will remain consistent.

4 Software products

There are various software products that utilize the concept of big data, and in this section some of them will be presented, in order to give an overview of how anyone can find a product for addressing a specific issue or need regarding the use of big data.

Fujitsu proposes to the market a large data toolkit, generic called Big Data Software, which allows large data technology to be used efficiently in the information and mission-critical systems of companies. [10]

Practically, Big Data Software consists of four products: a parallel-distributed data processing product, a complex event processing product, a ultra-fast transaction processing product, and an in-memory data management product, all of which represent standard technologies in large data applications.

According to the company, besides the reliability and performance guaranteed by incorporating many proprietary Fujitsu technologies, which have a proven track record of good results in mission-critical systems, solutions can be easily installed and operated. They can also be combined with products from other vendors, including open source software to build ecosystems to help customers use large data solutions.

Cloudera Enterprise 4.0 Big Data is the most recent product of Cloudera and it is a management platform. [11]

This platform for managing and processing big data provides tools for deploying and managing Hadoop systems, as well as management automation of large scale clusters and an easy integration with a broader range of management tools and data sources.

The new version of Hadoop that the platform uses, offers high-availability features that eliminates the single point of failure of the Hadoop Distributed File System, increased security that allows more sensitive data to be stored in CDH, and the ability to run multiple data processing frameworks on the same Hadoop cluster.

Datameer 2.0, is a software for Big Data analytics which can combine data integration, analytics and visualization into a single package that offers a spreadsheet interface. [12] This software is offered both in enterprise edition and workgroup and desktop editions to ease the access for every user.

A Business Infographics Designer is included in this software, in order to easy creation of graphics and data visualization design control. Built on HTML5, the software provides an enhanced user interface, and also offers support for additional data sources including Facebook and Twitter. It also has a useful feature that provides improved integration with the Hive data warehouse system for Hadoop.

Another software application useful for big data LucidWorks Big Data [13], which is a cloud-based development system of open-source software for prototyping Big Data applications. This application can help businesses analyze unstructured information, the so called "dark data" - text messages, audio files, email repositories, log files and other unstructured content.

LucidWorks Big Data is good choice as it incorporates a lot of different technologies such as Hadoop, but also Apache Lucene and Solr search, which are open source applications.

Moreover, it has R programming language for developing analytical applications and supports Apache Mahout which is used for building scalable machine learning algorithms.

5. Case Study Apache Ignite vs Spark

In order to compare the two frameworks, a multi node setup was used. Each node consists of a virtual machine that runs Ubuntu Linux 16.04 LTS. These VMs were provisioned with 1 virtual CPU, 2048 MB of RAM and a network card bridged in a Local Area Network. Each node had been assigned a static IP in order to easily start and stop them. In order to monitor the cluster, Elasticsearch and Kibana were used on a laptop that collected all the data received from each node. Each node sent data to Elasticsearch through metricbeat.

To properly asses the performance of these frameworks, different scenarios were tested, but the main calculation that was performed was a matrix multiplication of square matrices. The purpose was to assess how these frameworks behave in terms of time of execution, resources used and scalability. With these benchmarks in mind, this testing plan took shape:

For each framework, a matrix multiplication algorithm would be developed in their native APIs programming language:

Scala for Spark and Java for Ignite, and these algorithms are ran across a cluster of nodes. In order to asses the way these frameworks use computing resources, how much time they need to do matrix multiplication and in order to see how scalable they are, the need to vary some parameters arised.

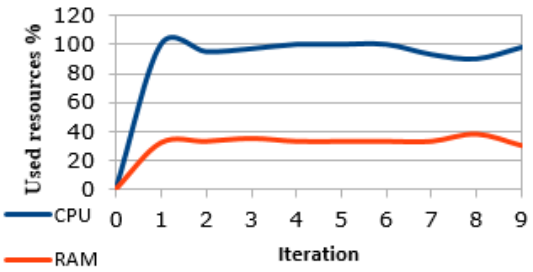
The parameters that were varied were the number of rows and columns each matrix had and the number of nodes that would take part in the computation. So, for each framework, the algorithm was ran for 9 times, varying the matrix dimension from 1000 rows and columns to 2500 and then finally to 5000, and by varying the number of nodes from 2 to 4 and a maximum of 6 nodes.

Below, we can see the way they behaved in each of these scenarios.

Spark

Table 2 – Spark Results

Iteration No.	No. of nodes	Matix dimension	Time (mm:ss)
1	2	1000x1000	00:12
2	2	2500x2500	00:23
3	2	5000x5000	00:31
4	4	1000x1000	00:08
5	4	2500x2500	00:14
6	4	5000x5000	00:20
7	6	1000x1000	00:05
8	6	2500x2500	00:12
9	6	5000x5000	00:18



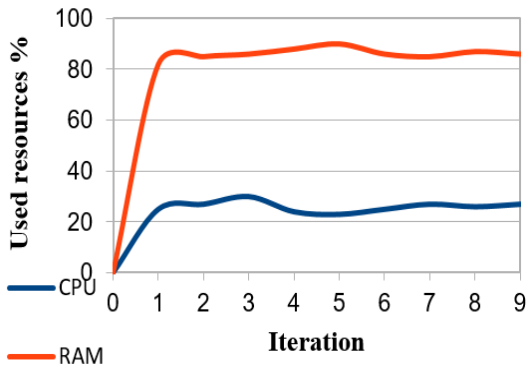
Spark used almost always 100% of CPU resources and about 33% of RAM resources

Ignite

Table 3 – Ignite Results

Iteration No.	No. of nodes	Matix dimension	Time (mm:ss)
1	2	1000x1000	00:11
2	2	2500x2500	02:56
3	2	5000x5000	28:04
4	4	1000x1000	00:05

5	4	2500x2500	01:25
6	4	5000x5000	24:09
7	6	1000x1000	00:05
8	6	2500x2500	01:52
9	6	5000x5000	27:10



Spark used almost always 100% of CPU resources and about 33% of RAM resources

6. Conclusions

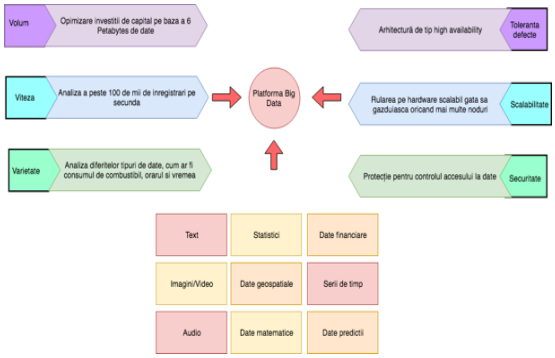


Fig.8. Summary of Big Data

As shown in Fig.8., there are various things to consider when talking about big data, no matter if it is about implementing or using such a deposit. There are the three most important requirements to be met, but also other criteria to be considered, as big data is not only about the technical part, i.e. how it is implemented, but also about the enterprise needs of such storing method.

When talking about key requirements, one should consider some main points. An application needs to handle and process huge amount of data, for example in the case of a financial application, it may need to optimize capital investments based on 6 Petabytes of information. Due to this amount, the first key requirement, volume, comes up.

Next, a vital element of many successful businesses is represented by the clients. A common need could be analyzing 100k records per second, in order to address customer satisfaction in real time, by giving suggestions or special offers. This is a good example when velocity is needed.

For the last key requirement, variety, the need to optimize shipping logistics could be considered, which demands the analysis of various types of data, such as fuel consumption, schedule, and weather patterns.

Apart from these requirements there are also some enterprise criteria that are important in every field of business. No matter if it is about finance, tourism, or statistics, a high failure tolerance is needed, which means a high availability architecture to support hardware or application failure.

Furthermore, big data should be a good option for expanding enterprises, so the hardware should be scalable and always ready to accommodate new nodes for processing more data. Security is also a must, as big data, like any other database solution, requires protection for granular data access control. The last element to be considered in that it needs to analyze data in native format and not demand a specific, standardized one, so that different types of data, such as text, images, statistics or time series can be all processed in their raw form.

Considering the presented case study, it can be stated that choosing the right technology or framework is strongly related to the needs of the specific company and to what kind of data is to be processed.

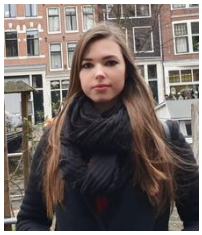
In terms of time spent computing, Spark is the clear winner by far. In terms of resources used, the two frameworks are very different. In terms of scalability, it appears that Ignite is not scalable because, as the number of nodes rises, the time does not shorten.

This issue might be because of the limitations of the LAN speed that might have slowed down the communication between the master node and the slaves.
The main conclusion to be drawn about big data is that it can be a very good solution for storing huge amounts of data, without having to put it in a specific form. There are many advantages for using it, as it provides volume, variety and velocity, and there are various technologies and software products that can address all the needs a company has when it come to storing data and processing it in order to obtain valuable information.

7. References

[1] O'Reilly Media Inc., Big Data Now: 2012 Edition, Kindle Edition
[2]https://www.projectguru.in/publications/difference-traditional-data-big-data/
[3]https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm
[4] Vignesh Prajapati, Big Data Analytics with R and Hadoop, Packt Publishing, ISBN 9763-78236-328-2
[5] https://www.datamation.com/big-data/big-data-technologies.html
[6] Mike Barlow, Real-Time Big Data Analytics: Emerging Architecture Kindle Edition
[7]http://www.imexresearch.com/newsletters/Newsletter2_22.html
[8] https://cloud.google.com/dataflow/
[9] Guy Harrison, Next Generation Databases: NoSQL, NewSQL, and Big Data, Apress Publishing, ISBN 978-1-484213-30-8
[10]http://www.clubitc.ro/2017/06/02/solutii-fujitsu-pentru-recolta-big-data/
[11]https://blog.cloudera.com/blog/tag/big-data/
[12] https://www.datameer.com/product/
[13]https://lucidworks.com/2013/01/29/getting-started-with-lucidworks-big-data/
[14] https://databricks.com/spark/about
[15]https://opensource.com/business/15/4/guide-to-apache-spark-streaming
[16]https://apacheignite.readme.io/docs
[17]https://www.baeldung.com/apache-ignite

Mădălina GHEȚU is a second year master's student at University of Economic Studies of Bucharest. She graduated from University Politehnica of Bucharest, Faculty of Engineering in Foreign Languages, holding a computer science degree. She is an instructor, teaching fundamentals of PHP programming and MySQL databases, but also a web developer, working with various technologies, including PHP, JavaScript and PostgreSQL.



Adriana BĂNUȚĂ is a second year student at University of Economic Studies of Bucharest at "Database - Business Support" Master. She is working as quality assurance engineer in an eCommerce company. She is familiar with PHP, PL/SQL and MySQL.

Cătălina DRAGOMIR is a graduate of the Faculty of Entrepreneurship, Business Engineering and Management, of the Politehnica University Bucharest in 2017. She is currently a student at the Academy of Economics Studies of Bucharest, and she is working as an IT Consultant at a small company in Bucharest.

Open Standards for public software used by a National Health Insurance House. A study of EU vs USA standardization approaches

Antonio CLIM, Răzvan Daniel ZOTA
The Bucharest University of Economic Studies
antonio.clim@csie.ase.ro, zota@ase.ro

Information technology improves reliability, innovation, and efficiency in the medical care sector by assisting in coming up with electronic health records. Looking into the interoperability of software and databases is relevant from the perspective of electronic health records. The standardization of processes in the European Union and the United States is diverse, which makes it all the more important to discuss open standards. Software systems create patient-centric medical care services and a platform for management. Thus, they facilitate the formation of functional health information networks and the exchange of information. Therefore, this improves the value proposition for all stakeholders involved. Open-source standards have been found to be developed independently of any single party. They do not have any legal or technical closest that prevent any party to use them. Similarly, they do not have extensions or components with a dependency or being based on preparation standards. Additionally, they are available for full public assessment without any form of constraints. This paper discusses these open standards and how best they have been deployed in the United States and the European Union — understanding that advantages and disadvantages of open standards are also imperative.

Keywords: *User as Developer (UaD), Free And Open-Source Software (FOSS), cost-effectively software, EHR, interoperability, software, open standards, databases, security*

1 Introduction

The benefits of information technology in the healthcare sector include higher efficiency, innovation, and reliability. Electronic health records (EHR) are currently being adopted in most healthcare facilities to store information related to patients' health. These information systems enable processing, mass storage, secure transmission, and accessibility to multiple stakeholders. Moreover, these information systems are designed to support continued and integrated healthcare services [1]. The information systems are additionally expected to assume a longitudinal form, i.e., to contain long-term patient data while delivering comprehensive functionalities regarding the management of health events and care from diverse service providers and institutions.

The healthcare sector consists of a diverse set of stakeholders, ranging from private hospitals, independent clinics and individually licensed practitioners, to the Government, research institutions of regulatory agencies. A requirement of contemporary Health Information Systems (HIS) is that they should integrate the operations of these stakeholders in a way that guarantees efficiency and security [13]. This effort requires a guiding framework which can be used by individual players to create their information systems and software. It can be expected that such coordination will facilitate greater integration between all stakeholders, be they large or small.

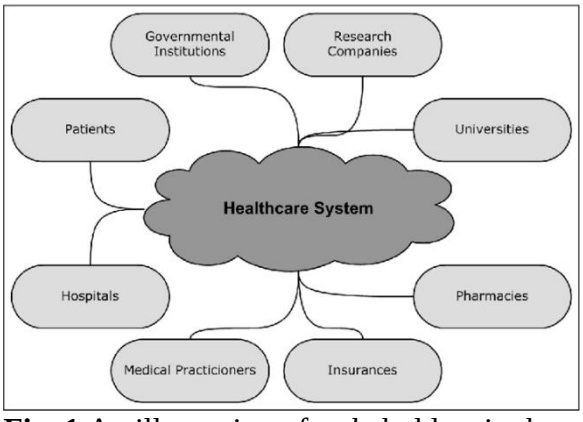


Fig. 1. An illustration of stakeholders in the

healthcare sector [10]

The National Health Insurance House (NHIH) is one of the organizations tasked with regulating the quality and access to healthcare. The organization regulates an industry comprised of a diverse group of service providers in terms of scale and types of health services they offer. It thus requires an information system and technologies that can consolidate the data derived from a broad number of channels. Other agencies with different roles within the healthcare sector often have their independent information systems. However, it is possible that the information collected and produced by the NHIH may also be useful to these organizations. The NHIH may also need to acquire information from the other agencies within the healthcare sector. The increasing connectivity between these agencies validates the debate regarding standards that should be employed in designing the software used by the NHIH and how such standards should be managed. Adopting open standards is the most appropriate approach, as open standards are managed by the community of users, developers, and stakeholder groups.

2 Public software used by aNHIH

In the contemporary operating environment, information systems are a critical component for any organization. The NHIH has a public-facing role that requires extensive environmental awareness and access to diverse data sources. For an information system to be fit for use in such areas of operation, it must be integrative in nature and contain elements that promote and maintain user engagement. These software systems must also maintain a high degree of transparency within the organization, while deriving support from the agency's management structures [8]. Moreover, the modeling strategy used in these information systems must also be accurate and clear. In most cases, the challenges with the design and implementation of public software in such institutions also include budget and time overruns, as well as failure to deliver features that fully comply with required specifications.

The public software used by the NHIH is required to present comprehensive and updated sets of information to users. This requirement includes the facilitation and support of best practices within and outside the organization. These software applications are required to effectively help users and other stakeholders within the healthcare sector in comprehensively delivering their services and decision-making processes [18]. It can also be expected that the success of the public software used by the NHIH is dependent on its ability to flexibly serve the emerging needs of its users. The systems are required to accommodate future updates at minimal costs in order to serve the future needs and stakeholders who may come into the current operating environments.

Compatibility is a major requirement for software systems used by the healthcare agencies such as the NHIH. To enhance connectivity, stakeholders within the healthcare segment implement information systems that will require interfacing with systems such as those used by the NHIH. Similarly, the agency may also need to receive data feeds from other public and private information sources. This requirement can be effectively delivered by designing information systems that are compatible with the other systems [18]. The achievement of this objective is dependent on the ability of the organization and the stakeholders to use a common framework for design and system implementation.

3 Standardization Strategies

3.1Using standards

Software systems are essential for the creation of patient-centric healthcare services and management. These systems are required to facilitate the formation of

effective health information networks and exchange of information among independent agencies, as well as stakeholders in the healthcare sector. Organizations such as the NHIH are required to create population databases that contain personal health records that can be used for effective health surveillance. Such systems have a broad range of applications from the delivery of social services to providing defense against bioterrorism. This segment of the paper evaluates the importance of standards in health information systems, the processes used in creating them, as well as how the management of such standards can be effectively achieved.

At the highest level, standardization of systems is considered an important step towards the creation of interoperable infrastructure. The NHIH is expected to handle large volumes of diverse data. This information will be subsequently applied to policy-making, partnerships, governmental and institutional decision-making, as well as accountability to the public. The nature of these applications demands that public software systems have interfaces through which partner and third-party systems can be connected without the need for further design reviews [11]. Interoperability allows other stakeholders in the healthcare sector to innovate and deliver services based on the data provided by the NHIH, while giving the agency an avenue to connect to other regulatory agencies and independent organizations in order to obtain data [5]. This level of communication is only possible if the associated systems operate on the same set of protocols and standards.

The design standards used in the software system have the potential to define how individual elements are designed and implemented. All concepts and applications in which data is used within the NHIH can thus be defined using a specific list of attributes which can be replicated at a future date. From these basic definitions and the design rules outlined in a standard, the higher-order data structures, process templates and documentation processes can be defined. Affiliated stakeholders can then follow the same standards to design, build, and implement information systems that seamlessly connect to the software system used by the primary agency such as the NHIH. The use of standardized design processes creates the room for future innovations and development of the information systems.

3.2 Creating a standard

The creation and implementation of standardized systems is based on the operational imperative to interface systems made by affiliated stakeholders. The introduction of common protocols and operational procedures has been recognized and implemented in many industries, with its most evident benefit being the creation of diverse business opportunities. The standards to be created must have a direct benefit to the stakeholders involved in its creation [7]. The creation of standards requires the availability of technical expertise within the affected industry. In the healthcare sector, expertise on information systems is likely to be derived from the software engineering community, healthcare managers and regulatory agencies. The standards are thus expected to meet the information sharing and privacy requirements while delivering the highest levels of availability and security.

The creation of open standards requires a majority support within the community of stakeholders that the standards affect. The standard must effectively remain neutral in its technical features and must be passed through an open acceptance process. This process opens up the standards for additional support and implementation by stakeholders. Generally, applications and systems that require data sharing and communication are enhanced in efficiency through the implementation of a common standard. The four basic ways through which a standard

can be created are outlined below:

- A community of interested stakeholders convenes and designs an ad-hoc standard to be used to build varied independent systems.
- The government or a regulatory agency may conceptualize, design, and enforce the implementation of a specific set of standards.
- The economic aspects of an industry, such as competition and technological evolutions, may lead to a natural adoption of a specific set of standards.
- Standards can also be created through a formal consensus process between stakeholders through the mediation of a regulatory committee or agency.

4 Open Standards management

In the contemporary information systems, open standards are protocols for system design, implementation and documentation that are formulated and maintained by a majority of its users and stakeholders. There are open standards that are applied to hardware systems as well as those that apply to software products. The results of using these standards include the elimination of barriers to interoperability and the reduction of switching costs to the end users. Reduced switching cost means that users have a greater number of choices in their product selection [15]. Open standards consist of rules and design methodologies that are created or adopted by a majority of market stakeholders due to their popularity or efficiency. Therefore, there is no need for a governing body since changes to the standards are entirely based on the community consensus. Open standards are available to all the stakeholders free of charge and their modification is based on the community consensus [6].

4.1 User as developer (UaD)

Open standards in the software development process eliminate restrictions on each of the stakeholders. This allows standardized methodologies and procedures to be implemented in the software product (Fig. 2) without any prior advantage to any of the other stakeholders.

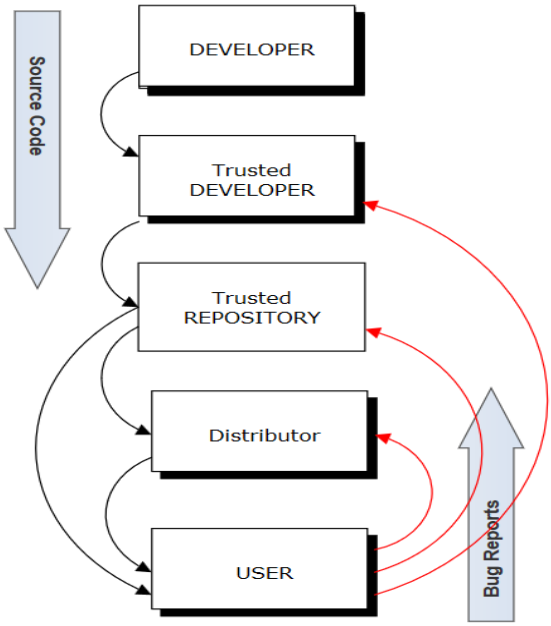


Fig. 2. Improvements (as source code and evaluation results for UaD diagram [2]

Thus, the key features of an open standard are listed below:

- The standard must be available for a full public assessment without placing any constraints on any party.
- It does not have components as well as extensions that are based or have dependencies on proprietary standards, i.e., components that do not meet the open standard requirement.
- It does not have technical or legal clauses that prevent its utilization by any party or environment.
- Further development and management of the open standards are done independent of any single party in a process that is fully inclusive of third parties and competitors.
- The open standard may be available in multiple complete implementations or one complete implementation to all the interested parties.

Open standards are promoted with the aim of fostering competition in the business

environment and to allow a diverse range of applications to interact. Open standards in the software development process enable sharing of data and functions in large systems regardless of the vendor or technical solution used in the individual components [11]. Open standards are required to be minimalistic, i.e., they should be as simple as possible. The minimalist nature of the open standards allows a large number of implementers and other stakeholders to participate in future developments and in the management of the standard. This is in contrast to proprietary standards that often include large numbers of features and components that are rarely used.

Software applications built using open standards, i.e., open source software (OSS) are those that any user can gain access to and modify the source code. The OSS include the license agreement that allows its users to access its source code, make modifications on it, and make new machine code, change any set of areas of applications, and redistribute the products [15]. This is in contrast to proprietary standards where a user can only execute the machine codes while source codes remain hidden or their modification being forbidden in the license agreements. Stakeholders in the software industry are advocating for the adoption of open standards through the Free Software and Open Source Software Movements. These movements are campaigning with a goal of providing developers and users with access to source codes and the freedom to redistribute software built on these standards.

The primary goal of the open standards movement is to shift the focus of competition from platforms towards their diverse implementations. This is based on the assumption that the open standards eliminate the barriers that prevent the entry of new players in any product or service segment. By focusing on the implementation of the same standard, the quality of the output, i.e., the product or service is anticipated to grow continuously. Additionally, the competing players are limited to the implementation of the same sets of standards, hence they focus on elements such as efficiency and cost considerations. In this regard, open standards lay a foundation for sustainability in design and implementation of systems.

4.2 Open Standards as key feature

Open standards hand the purchaser the capacity to create a strong position. This is because the development process is focused on the user needs and availability of demand. The software development process under open standards can be accomplished by the users themselves or by the developers who wish to use it in their own creations. In this model, users do not have to depend on experienced developers who are likely to demand high costs for such a process. The accessibility of open standards means that individual users, enthusiasts, and inexperienced developers may learn and effectively build the intended product. Product specifications can simply be adapted to meet users' financial capacity and needs.

The use of an open standard eliminates the necessity of an existing market demand for the finished product as is often the case in proprietary systems. This allows for the development of information systems purely for strategic purposes without regard to its immediate financial returns to the organization. The open standards have the characteristic of involving users in the development processes. Developers can further engage the user community in determining the elements of a minimum viable product that can be used to test the usability, reliability, and even the security features of the information systems. Contributions from a community of users, developers, and regulators lead to a superior development process. In such a process, the quality of code and the security of information systems can be checked fast and cost-effectively. The use of an open standard attracts highly motivated individuals and groups to contribute towards system

improvement at zero cost in most cases.

The use of open standards in the development of information system software taps into a large pool of community knowledge. Users, vendors, and other stakeholders have the capacity to suggest feature improvements and report errors in the operations of the product [9]. The large number of individuals and stakeholders vetting the source code and the product means that the operator of a system based on open standards has a large repository of quality feedback that can be used to strategically improve it over time. The cumulative learning process associated with communities working on open source projects leads to faster learning and improvement of the target systems or products. This is evident in some of the most popular software products and systems such as the Linux operating system and Android which have gained popular usage across the globe.

5 Advantages of Open Standards to the NHIH

The development of information systems often involves the allocation of large amounts of resources and the development of such systems is often time-intensive. It can thus be expected that the use of open standards and the developer communities that contribute towards their creation will save on costs and the time required to deploy the required system. The use of open standards will encourage the development of complementary and competing systems which further motivates innovation in the sector [19]. The net impact will be that users have superior information systems and access to efficient and reliable services. The use of open standards will also enable NHIH to choose components from a diverse range of competing implementations and, thus, guarantee that the most effective selections.

Healthcare agencies such as the NHIH have large numbers of stakeholders and roles that are focused on delivering optimal service to the public. Software systems that are used to facilitate operations within the agency and communication to public stakeholders are required to be efficient and error-free. The use of open standards enables individual users and experts to identify software errors as well as possible fixes that may optimize its operation. Such a process using proprietary systems may be nearly impossible as users and the developer communities do not have access to source codes. Users and developer communities may also suggest new features that effectively improve the quality of service NHIH delivers to the public.

The use of open standards reduces the costs associated with developing and maintaining the software. The components and procedures used in developing open source software are continuously updated by the developer and user communities. Using the community feedback and input, the NHIH can reliably use a smaller team of core developers to build and improve the key features of the public software. The open standards also have forums in which vulnerabilities to software components are reported and the mitigation measures are proposed. The use of open standards will guarantee that the NHIH is consistently providing optimal security to its information systems especially with regard to personal data that it may need to collect from its users.

Open standards often have the requirement that the code base remains unchanged in form or structure. This sets a foundation for long-term continuity as technology evolves and users need change. NHIH is expected to serve its users for the foreseeable future. Therefore, the agency needs information systems that will readily incorporate emerging users' needs and technologies in the future. Evidence suggests that open standards have a high degree of flexibility regarding the incorporation of new innovations and components [3]. The updated documentation that accompanies

open standards also mean that the NHIH can deploy other teams of engineers to work on the platform in the future without dependence on the initial developers. The open standards also ensure that the NHA will have continuous access to quality feedback from the user and developer communities to develop a reliable and efficient information system.

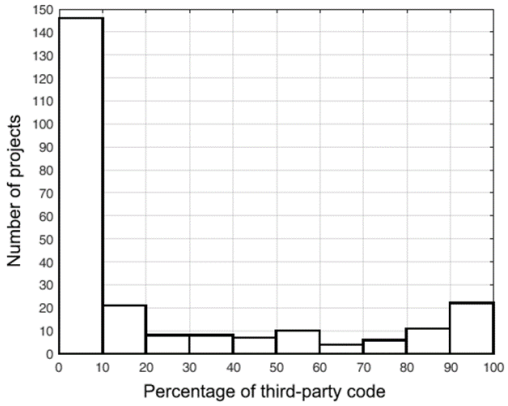


Fig. 3. Graphical analysis showing a high degree of compliance to open standards by stakeholders [14]

Open standards often have a high-level modularity which allows for the incorporation of new features and user needs. A public software used by an organization such as the NHIH is likely to be used by other stakeholders in the healthcare sector to deliver products and services. An example of this is the emerging popularity of connectivity and smart devices used in the healthcare sector. The service providers in these emerging segments, as well as traditional players which are currently digitizing, may need to interface their platforms to the NHIH's public information system [4]. The use of recognized and community accepted standards stands out as an important initial step towards ensuring that the NHIH can partner with the other organizations effectively in delivering healthcare services. The ease of interfacing to other information systems creates the opportunity for the NHIH to explore other commercial and noncommercial partnerships within the healthcare sector and beyond.

One of the benefits of open standards is the large repository of information that is created and shared by the developer and user communities. The extensive volumes of information that are often available regarding the implementation of open standards reduce the learning time required by developers and users [12]. This effect also contributes towards increasing the amount of qualitative feedback received from the various stakeholders. The NHIH can also use the information on open standards to implement best practices, an approach that minimizes its cost of operating public software. These are features that align with the organizational aim of contributing towards improvements in patient care, enhancing efficiency, and minimizing costs related to its operations.

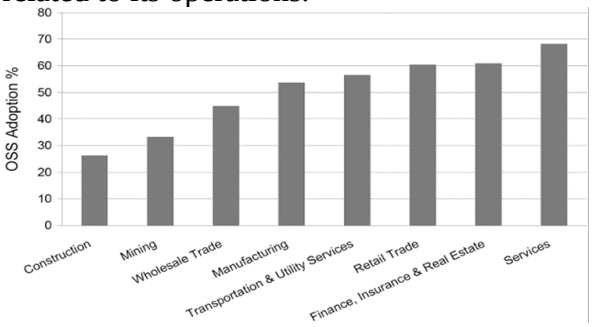


Fig.4 The general trend towards the adoption of open standards in various industries [16]

6 Disadvantages and barriers to Open Standards implementation

A major barrier to the adoption of open standards in public software within healthcare institutions such as the NHIH is the minimal understanding and familiarity with open source platforms. Healthcare institutions have traditionally depended on proprietary systems developed by contracted firms. Private organizations that offer software development services often use proprietary processes that do not conform to the open standards implemented in the broader industry. The implementation of the open standards in the NHIH will thus require the organization to develop its own capacity in terms of human resource or integrate into developer communities that

are working in no-related industries. However, this barrier is likely to be eliminated in the near future as open standards are gaining popularity in many segments of the information technology sector.

There has been a minimal commitment from the government in support of open standards adoption across its institutions and agencies. As a government agency, the NHIH is expected to follow the precedence set by other government departments and agencies. A majority of these institutions depend on proprietary systems both for internal use and in public services. It can be expected that there are powerful lobby groups comprising of proprietary system developers and vendors that depend on large government contracts. The fact that these proprietary system vendors have sufficient capacity and an established track record also minimizes the appeal of adopting open standards where human and resource capacity will have to be developed from the ground-up.

Critics have also argued that the total cost of ownership for a software product based on the open standards is higher than that of proprietary systems. The costs are associated with security considerations and investments in training [12]. The arguments citing security issues in the open source systems are based on the fact that all the stakeholders including hackers have full access to the source code of the software. These malicious agents can thus identify the points of vulnerability and exploit them before they are effectively fixed. The liability exposure of the organization may thus be increased by adopting a public software based on the open standards. However, counter-arguments have indicated that open standards encourage early detection of vulnerabilities and implementation of mitigation measures because of large communities of contributors that evaluate and test all the aspects of the source codes.

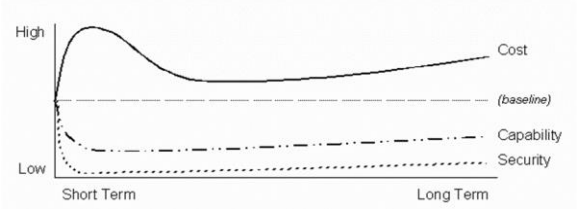


Fig. 5 An analytical illustration by the U.S Department of Defense showing an increase in long-term costs associated with proprietary software [17]

7 Conclusions

Health institutions across the European Union have a responsibility to upgrade their information systems to meet the emerging demands of the current environment. In this age of increasing integration, institutions such as the National Health Insurance House have to develop, implement, and maintain software systems that deliver information to the public and other stakeholders in the healthcare sector. The NHIH has the responsibility to share relevant information with the public and service providers such as hospitals, pharmacies, emergency health services, policlinics and even the General Practitioners. These service providers also need the means to provide the NHIH with updated service and patient information all of which comprise the holistic service experience that the citizens of the E.U. need in the 21st century.

Information in the healthcare sector is currently generated from diverse sources by many stakeholders. Some of these stakeholders primarily operate in the digital space such as mobile health applications and telemedicine services. These digital services are mostly developed based on open standards and interfacing them to the NHIH information systems will require the adoption of a compatible development approach. The primary advantage of the open standards is interoperability. Innovators, developers, and service providers will have the access needed to develop services that align with the tasks that the NHIH expects to accomplish. The adoption of open standards not only aligns the organization to the general trend of

development in the current information systems industry but also actively promotes innovation in the sector that has both short and long-term benefits.

The NHIH needs to maintain efficiency in conducting its operation with the diverse group of service providers both in terms of scale and types of health services. It thus requires an information system and technologies that can consolidate data derived from a broad number of channels. Other agencies have differentiated roles within the healthcare sector and often utilize independent information systems. However, it is possible that the information collected and produced by the NHIH may be useful to these organizations. The NHIH may also need to acquire information from the other agencies within the healthcare sector. This mutual dependence and increasing connectivity illustrate the need for the open standards that should be used in designing software used by the NHIH. Such standards should be managed by communities of contributors and developers.

The development and use of open standards benefit from the fact that there is a large community of experts, users, and partner institutions that can contribute in the form of feedback and advice. This aspect of the open standard systems has been cited to result in systems that optimally meet the user needs and continuously improve to cover the emerging demands. It can be expected that the NHIH would have to evolve its services in the future to meet new demands from its stakeholders including hospitals, the public, and academic researchers. The organization can effectively prepare for this future by adopting open standards which are highly flexible because of the large communities that continuously contribute to developing them. The widespread availability of information relating to open standards also means that the organization will not have to invest large volumes of resources in developing new standards for their software solutions or the carry out extensive research that is usually needed to continuously improve them.

References

[1] Balgrosky, Jean A. “Essentials of Health Information Systems and Technology.”,*Jones & Bartlett Learning*, 2015, pp. 172-178.

[2] CIO U.S. DOD. "Open Source Software FAQ". *Dodcio.Defense. Gov*, 2018, http://dodcio.defense.gov/Open-Source-Software-FAQ/ [Accessed 17 Apr 2018]

[3] CIPPM, “An Analysis of the Public Consultation on OPEN STANDARDS: OPEN OPPORTUNITIES FLEXIBILITY AND EFFICIENCY IN GOVERNMENT IT”,Centre for Intellectual Property Policy & Management (CIPPM), Bournemouth University for the Cabinet Officeof UK HM Government,Centre2012, pp. 1-83.

[4] DeNardis L. “Opening Standards: The Global Politics of Interoperability”,*The Information Society Series.* MIT Press, 2011, p. 75.

[5] Garde, Sebastian et al. "Towards Semantic Interoperability for Electronic Health Records". *Methods of Information in Medicine*, vol 46, no. 03, 2007, pp. 332-343. Schattauer Gmbh, doi:10.1160/me5001.

[6] Hammond, W. Ed. "The Making and Adoption of Health Data Standards". *Health Affairs*, vol 24, no. 5, 2005, pp. 1205-1213. Health Affairs (Project *Hope*), doi:10.1377/hlthaff.24.5.1205.

[7] HIQA. “Developing National Ehealth Interoperability Standards for Ireland: A Consultation Document.”*Health Information and Quality Authority*, Dublin, 2011, pp. 1-27. [Accessed 15 Apr 2018]

[8] HM Government. “Open Standards Consultation: The Government Response.” UK HM Government, London, 2012, pp. 1-20. [Accessed 15 Apr 2018]

[9] Jakobs, Kai. "Standardization Research in Information Technology"*Information Science Reference*, 2008, pp. 49-65.

[10] Meier, Andreas. "Open Standards for Data Exchange in Healthcare Systems - Semantic Scholar". *Semanticscholar.Org*,2007, https://www.semanticscholar.org/paper/Open-Standards-for-Data-Exchange-in-Healthcare-Meier/4cca0cbb504341d276d9e8abc9249fbdabb1abaa. [Accessed 17 Apr 2018]

[11] Moahi, Kgomotso H et al. "Health Information Systems and The Advancement of Medical Practice in Developing Countries",*IGI Global*, 2017, pp. 33-77.

[12] Reynolds, C.J, and Wyatt J.C. "Open Source, Open Standards, And Health Care Information Systems". *Journal of Medical Internet Research*, vol 13, no. 1, 2011, pp. 1-13.doi:10.2196/jmir.1521.

[13] Russell, A. L. "Open Standards and The Digital Age."*Cambridge University Press*, 2014, p. 278.

[14] Shah, A. and Abualhaol. I. "License Compliance in Open Source Cybersecurity Projects". *Timreview.Ca*, 2016, https://timreview.ca/article/966. [Accessed 17 Apr 2018]

[15] Sittig, D. F. and Wright A. "What Makes an EHR "Open" Or Interoperable? Table 1:" *Journal of The American Medical Informatics Association*, vol 22, no. 5, 2015, pp. 1099-1101. *Oxford University Press* (OUP), doi:10.1093/jamia/ocv060.

[16] Spinellis, D. and Vaggelis G. "Organizational Adoption of Open Source Software". *Journal of Systems and Software*, vol 85, no. 3, 2012, pp. 666-682. Elsevier BV,doi:10.1016/j.jss.2011.09.037.

[17] The MITRE Corporation. "Use Of Free And Open-Source Software (FOSS) In The U.S.Department Of Defense". *Terrybollinger.Com*, 2003, http://www.terrybollinger.com/dodfoss/dodfoss_html/index.html. [Accessed 17 Apr 2018]

[18] Wager, K. A. et al. "Health Care Information Systems"*Wiley*, 2013, pp. 312-340.

[19] Weston S. and Kretschmer M. "Open Standards in Government IT: A Review OfThe Evidence"UK HM Government, Bournmouth, 2012, pp. 1-62. [Accessed 15 Apr 2018]

Răzvan Daniel ZOTA has graduated the Faculty of Mathematics – Computer Science Section at the University of Bucharest in 1992. In 2000 he has received the Ph.D. title from the Bucharest University of Economic Studies in the field of Cybernetics and Economic Informatics. His research interests include Business Informatics, Computer Networks and Smart Cities. Currently he is full Professor of Economic Informatics within the Department of Computer Science in Economics at Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies. He is the author of more than 25 books and over 75 journal articles in the field of software quality management, software metrics and informatics audit. His work focuses on the analysis of quality of software applications

Antonio CLIM has graduated the Faculty of Food Science and Engineering at the University of Galati in 1996. In 2017 he has been admitted as Ph.D. candidate at the Bucharest University of Economic Studies in the field of Economic Informatics. His research interests include Business Informatics, Computer Networks and Smart Cities. Currently he is Teaching Assistant at Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies. His work focuses on the analysis of machine-learning applications for health services.

Improving the Customers' In-Store Experience using Apriori Algorithm

Ioana DAVID
Department of Economic Informatics and Cybernetics
Bucharest University of Economic Studies
ioana.david4@yahoo.com

The 21st century is the era of technology and digital development. That's the reason why mobile applications nowadays became an important support for businesses and a significant part of our daily activities. The usage of smart devices and the ease of access to technology lead to obvious changes in consumer behavior. Therefore, as it has been remarked a decrease of in-store shopping, improving the shopping experience in traditional stores has become of high interest for many retailers. In this paper, we propose a mobile application which helps people to optimize the time they spend inside of hypermarkets and which suggests an optimal placement for aisles in a store.
Keywords: *in-store experience, aisles placement, mobile application development*

1 Introduction

Grocery shopping is a continuous activity in everyone's life. One of the effects of technological evolution is the desire to optimize daily activities by using mobile phones. Thereby, online shopping becomes more popular than in-store shopping. As online shopping is focused mainly on non-perishable products, it gives room to traditional stores to maintain or even grow their businesses. Due to the wide range of products, hypermarkets allow customers to buy everything that can satisfy their routine shopping needs.

The process of buying products for the current consumption became a challenge both for the clients and for the stores. The time spent searching products is the most important problem that people take into consideration when deciding to visit a hypermarket. On the opposite side, retail stores are preoccupied with strategies to optimize products positioning in order to improve sales and revenue.

Providing a large variety of grocery and household products, hypermarkets accentuate in-store environmental stimuli, such as shelf-space allocation and product display. These strategies can lead to positive effects like maximizing profit, improving stock control or improving customer satisfaction. [1] The problem of deciding how to stock products among the aisles of a store can be resolved by extracting valuable information from the store's transactions history. Different analytical tools and algorithms help companies to study cutomer purchasing behaviour.

A well-known technique to understand the way customers make decisions is the *Association rule learning* which helps to discover relationships between items bought toghether. This is an algorithm that helps companies to study and to identify purchasing patterns that can be used in order to establish certain marketing actions or strategies. [2]

The application proposed in this article combines companies' necessity to position products in the most effective way and clients' interest in optimizing the time spent inside hypermarket. For these reasons, the application has a preliminary module in which an algorithm is applied to place products based on antecedent transactions. The main module consists in a mobile application with a user-friendly interface, where people can view details about many hypermarkets or can organize their future shopping sessions.

2 Market Basket Analysis

The concept of *affinity analysis* is a data analysis and data mining principle, which

presents the coexistence relationships between different entities. It is used also in the retail industry, where it is known as *Market Basket Analysis*, which extracts rules that associate products based on the frequency of purchasing them together in sets of different sizes. When using the term of *market basket*, we only refer to item sets purchased by a customer during a single shopping session. The process of frequent item sets analysis begins from the premise that no consumer takes isolated decisions, so he rarely buys a single item and he always inclines to buy numerous products from distinct categories. Consequently, analysts' purpose is to discover the products which lead certainly to buying other products. Revealing this information allows managers to develop methods of influencing the shopping behavior, increasing the demand for some products through promotional offers. In addition, companies focus on optimizing prices of products in order to achieve higher cross-selling rates and on high-margin products, because these are the main prices that must be adapted to purchasing patterns. [2]

Depending on a high volume of data which involves all the transactions realized in a hypermarket, market basket analysis can be considered a difficult operation which is based on the analysis of data sets which exceed the processing capacity of traditional software. This challenge is correlated to the concept of *big data*, which regards the predictive analysis, the human behavior analysis and extraction of information from high amounts of data.

The systematization of the extracted data from a hypermarket's transactions and the analysis of sold products, both from the point of view of the associations and from the point of view of prices and quantities, help in the process of product placement. Once the whole analysis is completed, a hierarchy f~~or~~ of the aisles from a store department can be established in order to join products with a high affinity probability.

The most used algorithm for estimating associations is the *Apriori algorithm* proposed in 1994 by three researchers who defined the problem of market basket analysis: Rakesh Agrawal, Tomasz Imieliński and Arun Swami. [3] This algorithm returns the associated itemsets using a high complexity in time and space, because it searches all the possible associations (2^n - 1).

3 Application Design

The application fulfills two main goals, one reffering to an optimal layout for hypermarkets' aisles and the other one to minimize the time spent by customers at shopping, whithout decreasing the company's profit.

First of all, the application suggests a specific placement for all the aisles from a store and identifies each aisle from the system with the position indicated after applying the Apriori algorithm. At this step, user interaction with the application is not necessary, because the store organizing process takes place before the application ~~to~~ be available.

Second of all, the application is ready to be used by hypermarkets' customers during theirs shopping sessions. Users have access to many functionalities designed to improve every shopping experience. Among these it can be mentioned the possibility to set and modify a favourite store on the basis of which are developed all the future actions, such as an interactive map with the hypermarkets' layout, a list with all the available products, a search engine to find a particular product or a particular category of products, a module to view all the information regarding a specific hypermarket and an option to locate a store using Google Maps. In addition, users can define their own shopping list which can always be modified and which can contain products from all the stores available on this platform or they can add products to a list with their favorite products. To ensure the

transparency of all prices and sales amounts during the last year, the application provides a module named *Statistics* which presents different graphs and hierarchies based on users' filters.

A user's interaction with the system can be visually represented in a use case diagram which helps to create a technical and functional perspective over the entire application. Thus, the diagram illustrated in Figure 1 presents the system's functionalities.

Fig. 1. Use case diagram

In the next phase - database design - the entities needed to implement the desired functionalities were identified: *Hypermarket, Department, Aisle, Category, Product, Receipt, Product History, Favorites List* and *Shopping List.*

Based on the relationships between entities, the structure of the database can be defined. Thereby, between these entities there are different kinds of relationships and cardinalities. For example, this model contains multiple one-to-many relationships, but there are also two many-to-many relationships. Whereas relational database paradigm does not accept the existence of many-to-many logical connections, it is necessary to transform this relationship into an additional table. This design pattern provides a consistent database model for future data operations and ensures the clarity of the system. The many-to-many relationships identified inside of this model are those between *Aisles* and *Products*, respectively the one between *Products* and *Receipts.*

Regarding *Favorites List* and *Shopping List* entities, it is not necessary to store the corresponding data inside the general database, because these contain information about a specific user, individualized by his own smartphone. Thus, data about favorite products and the shopping list will be stored locally on user's device.

The final database structure is shown in Figure 2, which represents the database schema, including all the attributes.

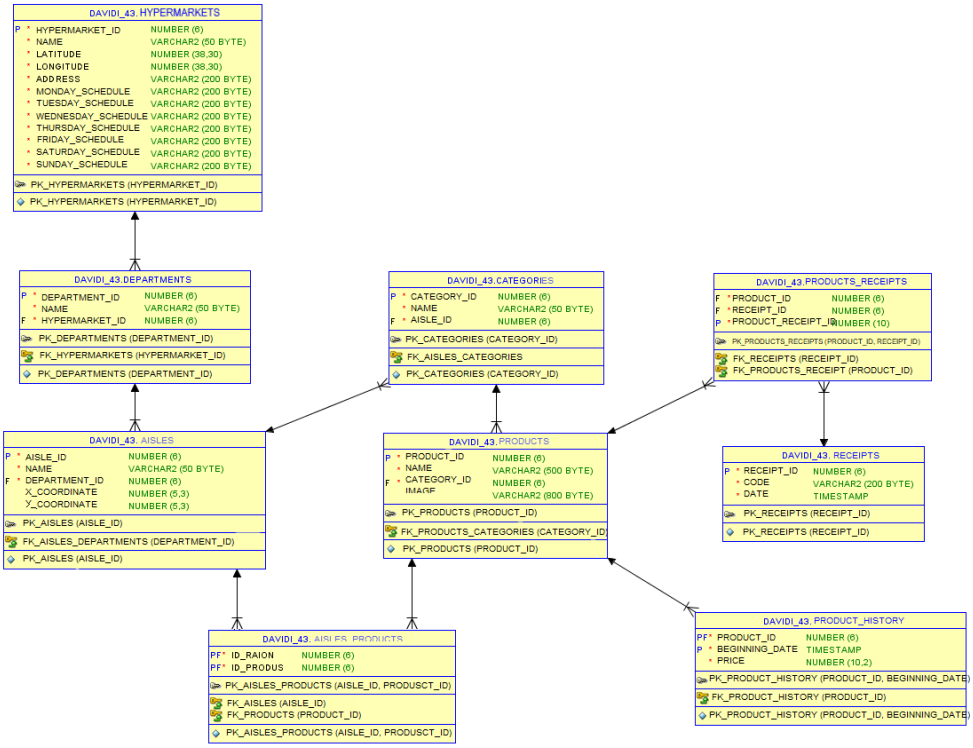


Fig. 2. Database schema

4 Software technologies

The development of this application is based on several technologies, such as Python language, Flask micro-framework, Amazon Web Services Cloud9 integrated development environment, SQL and PL/SQL languages, Java language and Android Studio integrated development environment.

Python is an open-source programming language, distinguished by its simplicity and its applicability in many areas. [4] That represents one of the reasons why Python was used both in data scraping and in developing the application's server. In order to perform the in-store experience analysis and optimization, the development of this application included the process of collecting data from a hypermarket's website. For Web scraping two Python libraries were used: BeautifulSoup and Selenium-Webdriver, which allow to automatically parse the data in HTML containers into Comma Separated Values files, which represent an intermediate stage before inserting it into an Oracle database. The functions used to extract data implied the launching of a virtual browser from which the elements from the Document Object Model (DOM) can be loaded. The DOM represents an interface which exposes an XML document as an arborescent structure. Python was also used to insert data into an Oracle database through the cx_Oracle library, which allows to establish a database connection and to execute queries with the help of a cursor.

Flask is a Python micro-framework used to develop the application's backend. It provides the possibility to develop a REST web service, with an obvious distinction between the client and the server. Through the web services developed in the REST API, create, read, update and delete operations (CRUD) can be executed using the four main standard methods of the HTTP protocol: *POST*, *GET*, *PUT* and *DELETE*. The server represents the middleware between the mobile application and the database, and the data is passed between the endpoints in JSON format.

Amazon Web Services Cloud9 is an integrated development environment in the cloud which enables the writing, compiling and debugging of programs within a web

browser. [5] It supports the Python language and provides access to a terminal. The API server was developed using this platform because it provides an external IP which is needed to be able to access to the web services from anywhere.

SQL and PL/SQL languages were used through the Oracle SQL Developer tool and helped to transform the database structure defined in the analysis process into a relational database. The high volume of data existent for storing all the products from a hypermarket leads to the necessity to model an efficient method for accessing and processing. The creation of the database was done using SQL's Data Definition Language to define and describe the nine tables and the relationships between them. During the development and the use of the mobile application, the SQL's Data Manipulation Language was used to insert data or query the tables. Additionally, the PL/SQL language was used to define sequences and triggers.

The application was designed for devices running the Android Operating System because it is the most popular operating system for mobile devices, covering 88% of the mobile devices as of the end of 2018. [6] From the point of view of mobile application development, Android is easier than other mobile operating systems, because it provides open-source tools and it uses the Java language.

The Java programming language is recommended in developing Android applications due to the variety of API tools and libraries which are available. For the development of the application the Java language and the Android Studio IDE were used thanks to the possibility to use multiple interactive instruments and to the intuitive development platform. Although the standard Java and Android libraries provide a lot of classes and methods, such as data structures, visual components and data processing methods, the use of some external open-source libraries extends the possibility to implement and simplify the effort to build an application. The libraries and APIs that were used are: OkHttp, Picasso, MPAndroidChart, ClickableAreasImages, ArcLayout and Google Maps API.

5 Input data depiction

Containing the products sold by hypermarkets, input data is relevant also for the analysis process and for the final solution. Being impossible to access a retail store database, the solution was to collect data from a hypermarket's web site. For this purpose, the chosen hypermarket was Carrefour, because its web site presents the entire architecture of the store, from departments to product categories allocated to an aisle. Thus, the three hypermarkets that can be found in this application have the same organizational structure as Carrefour.

Using the Python programming language and the two mentioned libraries the data in each page of the store was extracted. The *Selenium* library provides a *web driver* which allows duplicating a web page to parse data. Two distinct situations of data parsing were found: one with crossing data in a single page and one with crossing multiple pages. The second situation takes place when the chosen category has more than one page with products. It was necessary to separate these cases because their URLs do not observe the same rule: in the first case the URL contains only identifying names, while the URL from the second case has the number of the current page appended at the end. Because web sites implement some security restrictions, it was required to apply a function which enforces the scraping method to wait some seconds before the next processing, in order not to overcharge the site and to block the access. After this step, the function *findAll* from BeautifulSoup library authorizes the access to the HTML structure of the page. This way, parsing data is realized navigating through the tree structure of the page, choosing the useful data from different element types as: *class, div* or *src* and

storing data temporarily into *.csv* files. The method of accessing the entire set of products from a page is presented below and it is based on selecting data from an item list whose class attribute is named "product isProduct product-item ng-scope".

```
url = link
browser = webdriver.Chrome
                    ('chromedriver')
browser.implicitly_wait(60)
browser.get(url)
html = browser.page_source
soup = BeautifulSoup(html,
                    'html.parser')
products = soup.findAll('li',
        attrs={'class':'product
        isProduct product-item
        ng-scope'})
```

Another important phase of creating input data consisted in the process of adding records to the database. Mostly it was fulfilled using the Python language to insert the data from the .csv files into an Oracle database through cx_Oracle connection and its cursor. To automate the process of parsing products the *os* module - which allows the communication with the operating system of the computer - and the *glob* module - which finds all the files' paths that correspond to *.csv* extension - can be used.

```
path = 'C:/Users/user/Desktop/licenta
        /Python/Scraping/Carrefour'
extension = 'csv'
os.chdir(path)
files = glob.glob('*.{}'
        .format(extension))
```

In the same function that inserts products into the database, prices variations are also defined by inserting random historical changes with an oscillation of 1 to 4 percent at an interval of 15 days. The chosen probability for a price to be changed is 30% and it is assured by generating a random number between 1 and 100.

Another essential function generates fictive shopping receipts for a period longer than one year, with a generating frequency of 30 minutes. This function takes into consideration the hour on which a transaction is done, ensuring that it doesn't take place outside the time slot 8-22. The selection of products included on a receipt is also randomly performed, using a random number of products and selecting random IDs of products.

6 Algorithm implementation

Aisles placement inside hypermarkets is accomplished relying on the transactions previously performed and applying the Apriori algorithm (as in Figure 3). The volume of the analyzed receipts depends on the system's performance, because the running time of the Apriori function is directly linked to its high complexity of $O(2^n)$.

To organize the aisles, it is necessary to know the available positions for each department before applying the algorithm. The coordinates of the available positions are considered input data for the algorithm, along with the ID of the current department and the maps dimensions.

The positioning function itself (Figure 4) calls another function which creates the array of tuples with aisles associations after extracting all the transactions and after applying the *apriori* function from *Mlxtend.frequent_patterns* library. To establish the coordinates for each aisle, it is defined an array with the aisles sorted descending from the ones with the strongest affinity to the ones with the weakest affinity. So, the main purpose is to associate aisles with an as high as possible association relationship. On the opposite side, the coordinates array is sorted first using the coordinates from the Ox axis of the map and secondly using the coordinates from the Oy axis. Finally, using the two arrays, the coordinates of an array position correspond to the aisle at the same position of the

second array. The values added into the database are estimated as percentages of the map by dividing each coordinate to the corresponding map dimension.

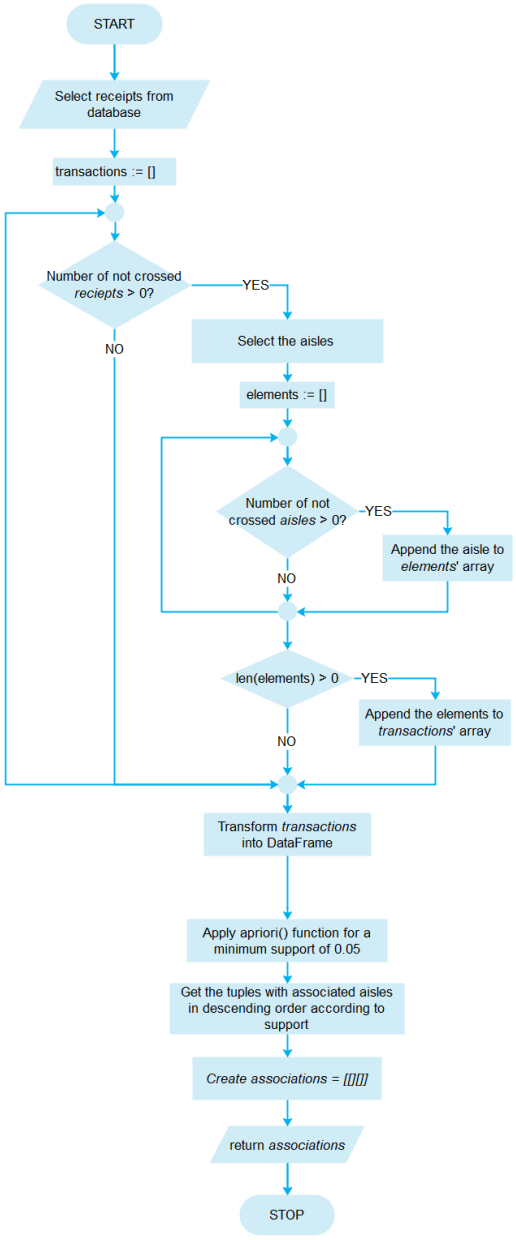


Fig. 3. Logical Schema for applying the Apriori function

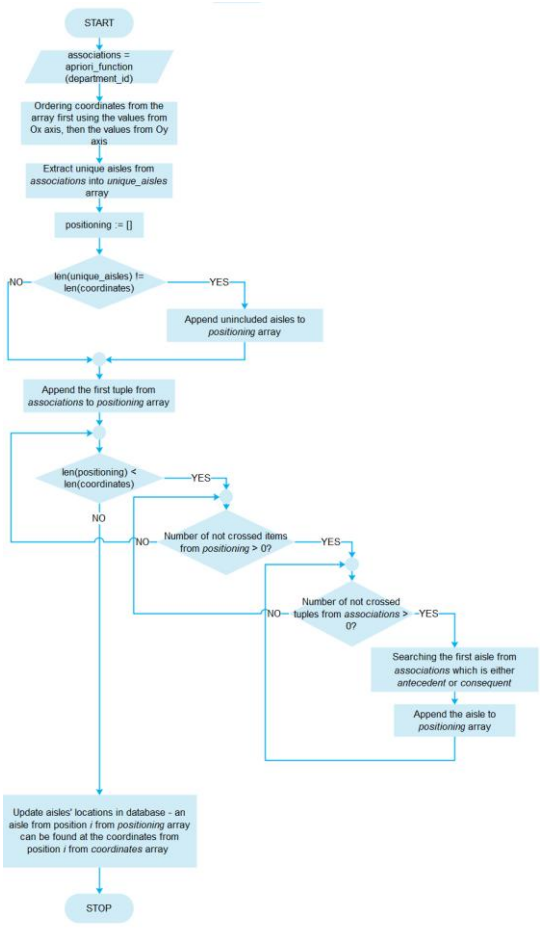


Fig. 4. Logical Schema for positioning function

7 Application interface description

The mobile application starts with the home screen that presents a series of six buttons circularly arranged and a menu at the bottom of the screen which is kept in all the screens (as in Figure 5). The six buttons have the role to link the application main functionalities to the corresponding application screens. The menu allows the users to access four essential screens from every point of the application. In the upper part of the screen it can be remarked an information icon which redirects to the application's description.

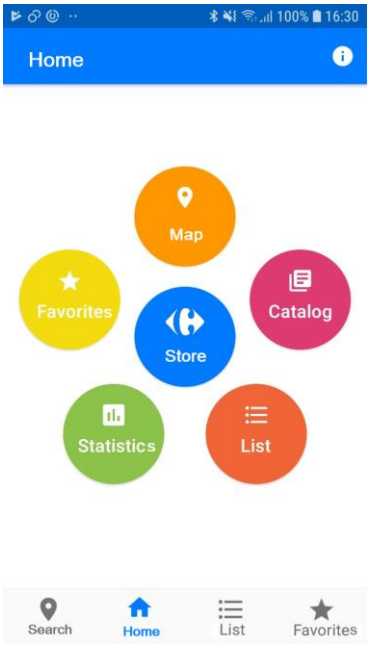


Fig. 5. Home screen

Whereas the application's main idea is to set a favorite hypermarket and to perform some actions based on it, a useful screen is the one that is opened after selecting the *Store* option from the home screen. This screen presents information about the favorite store and contains two buttons designed to locate the store on Google Maps and to change the favorite store as in Figure 6.

Fig. 6. Store

In addition, the users can access the *Map* section where they can navigate on the map of the favorite hypermarket. Each area can be selected and after that a message will be displayed at the bottom of the screen to indicate the department.

Accesing the *Catalog* screen gives the possibility to observe all the available products from a specific category selected by the user. The filter is applied after the gradual selection in three different screens which illustrate departments, aisles and categories. Each product from can be added either to the shopping list by selecting the associated button or to the favorites list by activating the star icon. At the same time, by selecting a product, it will be located on the hypermarket's map.

The screen named *Shopping list* allows users to keep a register of the products they need to buy during the next shopping session. In this screen, the user can see products from all the stores, not only from the favorite one, by expanding the list from the left side. The list calculates the total payment taking into consideration the quantities from each product. During the shopping session, the user can check the products already put in the basket, in order to be removed from the list. Furthermore, to optimize the time spent to search products in the hypermarket, the user can locate every product just by clicking on it (as in Figure 7).

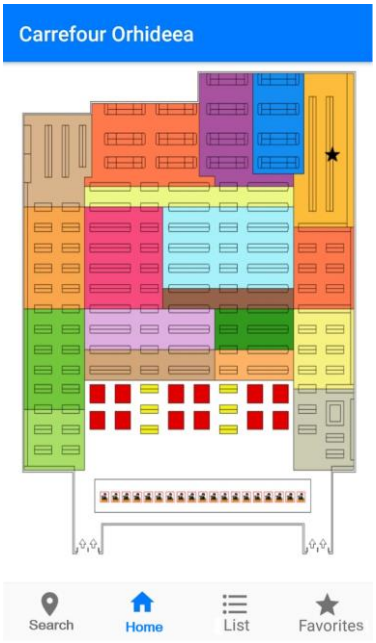


Fig. 7 Locate a product

Fig. 8 Statistics

The Favorites section offers the possibility to view the products selected as favorites to ease the process of adding product to the shopping list. The products can be eliminated from the list by disabling the star icon or added to the shopping list.

To search as fast as possible for a specified product from the favorite store, users have access to a search engine that returns the products which correspond to their requirements.

Another functionality is included in the *Statistics* screen (as in Figure 8) where prices and sales evolutions are presented in the form of a bar graph, a line graph and a hierarchy. When accessing this screen, the standard statistics are referring to all the products from the store, but users can personalize the results by either just mentioning the aisle or mentioning both the aisle and the category. Following that, the bar graph illustrates the amounts of sales in the recent 12 months and the line graph illustrates prices evolution in the last 12 months. The hierarchy presents the top three cheapest products from the selected category or the three most sold products.

8 Conclusions

The developed application represents a system which incorporates facilities for all the parts involved in a buy and sale process. Its main purposes are to increase retail stores' sales and to improve customers' in-store experience.

The application can be further extended. The main two directions in which the application can be improved are: adding an application module to be accessed by companies to refresh the database and implementing a complex process of positioning products inside aisles, taking into consideration the prices, the profit margin and other marketing strategies.

References

[1] G. Aloysius, "An approach to products placement in supermarkets using PrefixSpan algorithm," *Journal of King Saud University - Computer and Information Science*, vol. 25, pp. 77-87, 2013.

[2] M. Kouzis-Loukas, "Analysing Customer Baskets - A Business-to-Business Case Study," Erasmus School of Economics, Rotterdam, 2014.

[3] "Apriori algorithm," Wikipedia,

[Online]. Available: https://en.wikipedia.org/wiki/Apriori_algorithm..

[4] "Python Website," Python Software Foundation, [Online]. Available: https://www.python.org.

[5] "Amazon Web Services Cloud9," Amazon Web Services, Inc., [Online]. Available: https://aws.amazon.com/cloud9.

[6] "Global Market Share Held by Smartphone Operating Systems," [Online]. Available: https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems.

Ioana DAVID (b. February 4, 1998) has graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies with a bachelor's degree in 2019. She studied Mathematics and Computer Science at the National College of Computer Science "Tudor Vianu" from Bucharest.

Waterative Model: an Integration of the Waterfall and Iterative Software Development Paradigms

Mohammad Samadi GHARAJEH
Young Researchers and Elite Club, Tabriz Branch,
Islamic Azad University, Tabriz, Iran
m.samadi@iaut.ac.ir; mhm.samadi@gmail.com

Software development paradigms help a software developer to select appropriate strategies to develop software projects. They include various methods, procedures, and tools to describe and define the software development life cycle (SDLC). The waterfall and iterative models are two useful development paradigms, which have been used by various software developers in the last decades. This paper proposes a new software development methodology, called waterative model, which applies an integration of the waterfall and iterative development paradigms. In this model, the iterative model is embedded into the waterfall model to use the advantages of both models as an integrated one. It, in the most cases, is appropriate for large software products that need a long-term period of time for the development process. Experimental results demonstrate that the customer satisfaction score could be high by using the proposed model in various software projects.
Keywords: *Software Engineering, Software Development, Waterfall Model, Iterative Model, Waterative Model*

1 Introduction

Software is formed by using a collection of executable programming codes, associated libraries, and essential documentations. Software product is developed for a specific requirement, which is composed of various phases such as system analysis, coding, and testing. Software engineering is, in fact, the development process of software products using well-defined methods, tools, and procedures. It includes some of the important challenges such as scalability, cost, and quality management that should be considered by software development teams, precisely. The development process of software products is not possible without considering such challenges [1-6].

A software developer should choose appropriate software development models in the development process. A new software development model, called waterative model, is proposed in this paper. This model uses an integration of the waterfall and iterative software development paradigms. Since the waterfall and iterative models have useful advantages in the software development process, their integration could be very useful compared to some of the existing models. This model, in the most cases, can be used to develop large software products.

The remainder of this paper is organized as the follows. Section 2 represents the elements of software development life cycle. Section 3 presents a literature review on some of the software development paradigms. Section 4 describes various steps of the proposed model. Section 5 includes the evaluation results of some software projects that have been developed by the proposed model. Finally, the paper is concluded by Section 6.

2 Software development life cycle

Software development life cycle (SDLC) [7] is a well-defined and structured sequence of various stages in software engineering to develop various types of software products. SDLC framework includes various steps that are entitled by Fig. 1. The remainder of this section describes a short description about each stage.

Communication is the first step of SDLC, which indicates the user's requirements for a

desired software product. Requirement Gathering leads the software development team to conduct the project's requirements. Requirements consist of user requirements, system requirements, and functional requirements. They can be conducted by studying the desired software, referring to the database, or collecting answers from questionnaires. Feasibility Study investigates whether a software project can be designed to fulfill all of the user's requirements or not. Moreover, it analyzes the financial, practical, and technical feasibilities of the project. System Analysis assists developers to determine a roadmap about their plan and also define an appropriate software model for the project. It considers some of the features such as product limitations, problems identification, and the effect of project on organization. Software Design brings down the user's requirements and any other knowledge about development process to design the software project. It is composed of two designs: logical design and physical design. Moreover, this step can be conducted by using various tools such as data dictionaries, logical diagrams, and use cases. Coding is also known as programming step. It implements the software product via writing the program codes by a suitable programming language and developing an error-free executable program. Testing determines the acceptance rate of a software product, which is done by testing team and, later, by the customer. It can be managed while writing the codes by the developers and at various levels of the coding step such as module testing, programming testing, and product testing. Integration step can be used to integrate the software product with libraries, databases, and other programs. Implementation means installing the software product on the user's machines. It can be used to test the portability and acceptability features of product. Maintenance determines the efficiency and error-free rate of software product. Furthermore, it aids to train the users by using some of the required documentations. Software product can timely be maintained by updating the program code based on the changes taking place in users, environment, and/or technology.

Fig. 1. An overall view of the SDLC framework

3 A literature review on software development paradigms

Waterfall model [8, 9] is the simplest development model from among a list of available software development paradigms. All steps of the SDLC framework are conducted one after another through a linear manner. That is, the second step will start only after the first step is finished and so on. This model considers that every process is perfectly conducted as planned in the previous step without any need to think about the past issues. Therefore, if there are some of the issues that are left from the previous step, this model will not work smoothly. This model is more appropriate for development teams when they have already designed and developed the same software in the past so they are aware of any development conditions. Fig. 2 shows an overall viewpoint of the model.

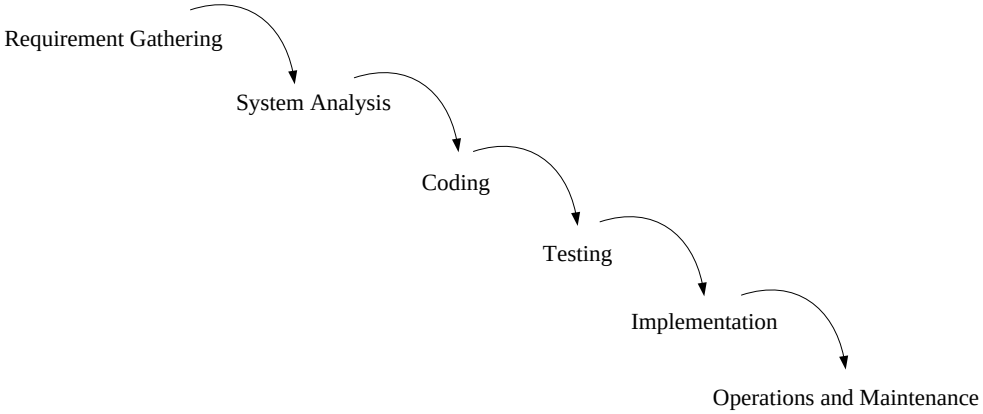


Fig. 2. Elements of the waterfall model

Iterative model [10] works based on iteration in the software development process. It conducts the development process via a cyclic manner so that every step is repeated after another. Firstly, software is developed on a small scale so that all of the steps are followed sequentially. Afterwards, more features and modules are designed, tested, and appended to the project on next iterations. In fact, the software at every iteration has more features and capabilities than the previous iterations. It is worth noting the management team can investigate risk managements of the project after completing each iteration in order to prepare the next iteration. Fig. 3 illustrates elements and processes of this model.

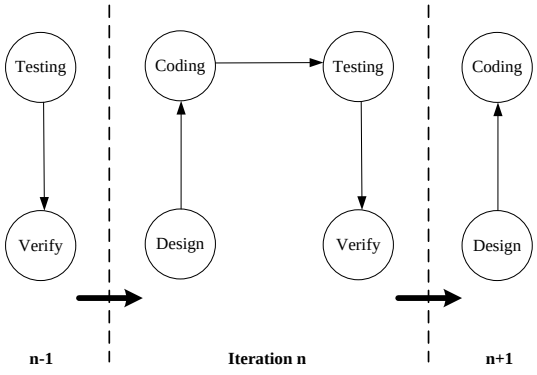


Fig. 3. A schematic of the iterative model

Spiral model [11-13] is a combination of iterative model and one of the SDLC models. In addition to any other process, it considers risk management that is often neglected by the most development models. In the first phase, it determines objectives and constraints of the software product at the beginning of iteration. In the next phase, the model conducts prototyping of the software and risk analysis. Afterwards, one of the SDLC models is applied to develop the software. The plan of next iteration is prepared in the final phase. Fig. 4 depicts a schematic of the spiral model.

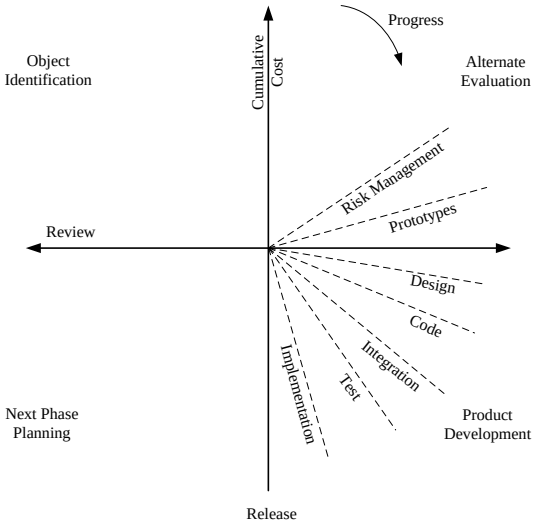


Fig. 4. An overall view of the spiral model

V-model [14, 15] facilitates the testing capabilities of software at each step via a reverse manner. It can solve the major drawback of waterfall model in which every step should be started only after the previous step is finished, without any chance to go

back. In this model, test plans and test cases are created at every step to verify and validate the software product based on the requirements of that step. This process leads both verification and validation to be conducted simultaneously. It is worth noting this model is also known as verification and validation model. Fig. 5 shows all elements and steps of the v-model.

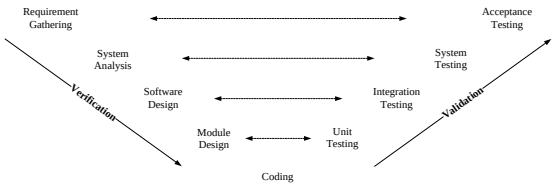


Fig. 5. An overall description of the v-model

Big bang model [16, 17] is the simplest model compared to the other development paradigms. It can be conducted by little planning, lots of programming, and lots of funds. This model is, in fact, similar to the process of universe in which lots of galaxies, planets, stars are created after an event. If the lots of programming and funds are put together, the best software product maybe will be achieved. A small amount of planning is required for this model. There is not any especial process to conduct the model as well as customers are not sure about the current and future requirements of product. In this model, the input requirements and conditions are arbitrary. It is worth noting this model is not appropriate for large software projects, but it is suitable for the learning and experimental purposes. Fig. 6 depicts a brief schematic of the big bang model.

Fig. 6. An overall schematic of the big bang model

4 Waterative model

Waterative model is formed based on an integration of the waterfall and iterative software development paradigms. It can be useful in software projects because it uses the advantages of both waterfall and iterative models. However, some of the new steps are involved in the proposed model as well as some of the steps are merged together. Since large software projects require a long-term development time, this model is suitable for such projects. Fig. 7 shows an overall view of the proposed development model. The most steps of this model have feedback to previous steps in order to solve the problems of software product at any step.

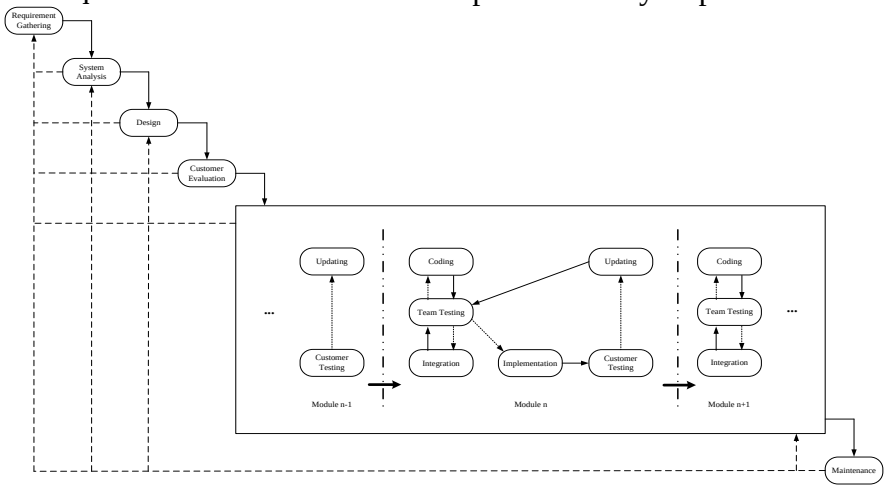


Fig. 7. Overall view of the proposed waterative model

Requirement Gathering includes both communication and requirement gathering of the SDLC framework. The service provider receives the request from the customer about developing a new software product. Afterwards, the development team manages the requirements according to user requirements, system requirements, and functional requirements. This step is a fundamental and essential step in this development model.

System Analysis involves both feasibility study and system analysis of the SDLC framework. The development team analyzes the feasibility and useful features of software product based on the user's requirements. The developers can use various algorithms to evaluate the financial, practical, and technical feasibilities of the product. Afterwards, the development team specifies a suitable roadmap to develop the software product according to requirements and feasibilities. It can consider development constraints, apply some of the learning systems, determine the scope, schedule, and resources of the project, and analyze other specifications of the project.

Design is, in fact, the software design step of SDLC. The software product is designed at this step based on the user's requirements and system analysis. The logical and physical designs of the product are also carried out in this step. They can be conducted by using some of the logical diagrams, data-flow diagrams, and pseudo codes. These designs should be organized carefully because the customer's initial evaluation and the team evaluation will be done by using meta-data and diagrams of this step.

Customer Evaluation is used is the proposed model before going on coding and the other development process. The reason is that the most problems of product at the system analysis and design steps should be solved by the customer's initial evaluation. This process leads further problems at the customer testing step of every module to be reduced considerably and, thereby, the most modules will be created and delivered to the customer during an acceptable period of time.

After the customer confirmed the correctness of development process until the design step, the software product will be divided into several modules to develop every module one after another, sequentially. Every module is composed of Coding, Team Testing, Integration, Implementation, Customer Testing, and Updating. All of the modules can be delivered to the customer sequentially to reduce fault rate of the whole system and decrease the project delivery time. After the code is programmed by programmers, all units of the code will be tested by tester team. Afterwards, the source codes are integrated with each other and the integration testing and system testing will be conducted by testers. Finally, the module will be implemented on the customer's machine to perform an acceptance testing by the customer. If the module has any problem according to the user's requirements then the module will be updated and tested by the team; otherwise, the next module will be created until the whole product will be delivered to the customer.

Maintenance is the final step in the proposed development model. In this step, the code could be updated based on any changes in the customer's requirements, technology, and platforms. It can also solve some of the challenges from the unpredicted problems and hidden bugs.

5 Evaluation results

Table 1 represents the experimental results related to some of my software projects in the last decade. In the most cases, I have attempted to use the proposed development model to accomplish various steps of these projects. However, this model has been progressed during a long-term period of time based on my experiences in developing various software projects. Customer

satisfaction score indicates that the customers have highly been satisfied by the development process.

Table 1. Experimental results

#	Project subject	Number of updates	Customer satisfaction score
1	Financial accounting software	7	85/100
2	Management software for civil engineering	34	95/100
3	Accounting software for shops	5	90/100
4	Thesis management software for universities	4	100/100
5	Educational scheduling software for educational institutes	6	100/100

6 Conclusions

A software developer should use a proper software development paradigm to develop software products from the requirements phase to the maintenance phase. All phases of the software development life cycle (SDLC) are defined and expressed by every development model. The developer can use various methods, tools, and procedures of any model to facilitate the development process of software products.

This paper proposed a new software development methodology, called waterative model. It uses an integration of the waterfall and iterative software development paradigms, which uses all phases of the SDLC framework through an efficient process. The iterative model is embedded into the waterfall model to use advantages of both models. In the most cases, large software projects can be developed by this model via a long-term development process. Evaluation results indicated that the customer satisfaction score was enhanced by using the waterative development model in various software projects.

I have used the proposed model in various software projects in the last 15 years. My experiences have demonstrated that this model could be useful in the most projects, especially large software products. Therefore, I decided to present this model as a new development platform to could be used by other software development teams.

References

[1] R. Mall, Fundamentals of software engineering. Delhi: PHI Learning Pvt. Ltd., 2014.

[2] N. Fenton and J. Bieman, Software metrics: a rigorous and practical ap-proach. Boca Raton, Florida: CRC press, 2014.

[3] F. Tsui, O. Karam, and B. Bernal, Essentials of software engineering. Burlington, Massachusetts: Jones & Bartlett Learning, 2016.

[4] E.J. Braude and M.E. Bernstein, Software engineering: modern ap-proaches. Long Grove, Illinois: Waveland Press, 2016.

[5] B. Fitzgerald and K. -J. Stol, "Continuous software engineering: A road-map and agenda," Journal of Systems and Software, vol. 123, pp. 176-189, 2017.

[6] M. Solari, S. Vegas, and N. Juristo, "Content and structure of laboratory packages for software engineering experiments," Information and Soft-ware Technology, vol. 97, pp. 64-79, 2018.

[7] D.J. Mayhew, "The usability engineering lifecycle," CHI'99 Extended Ab-stracts on Human Factors in Computing Systems, New York City, New York: ACM, pp. 147-148, 1999.

[8] Y. Bassil, "A simulation model for the waterfall software development life cycle," The International Journal of Engineering & Technology, vol. 2, no. 5, pp. 742-749, 2012.

[9] S. Madgunda, U. Suman, G.S. Praneeth, and R. Kasera, "Steps in re-quirement stage of waterfall model," International

journal of computer & mathematical sciences, vol. 4, no. 7, pp. 86-87, 2015.

[10] S.C. Ahluwalia, D.B. Bekelman, A.K. Huynh, T.J. Prendergast, S. Shreve, and K.A. Lorenz, "Barriers and strategies to an iterative model of advance care planning communication," American Journal of Hospice and Pallia-tive Medicine, vol. 32, no. 8, pp. 817-823, 2015.

[11] B.W. Boehm, "A spiral model of software development and enhancement," Computer, vol. 21, no. 5, pp. 61-72, 1988.

[12] B. Boehm, J. Lane, S. Koolmanojwong, and R. Turner, The Incremental Commitment Spiral Model. Boston, Massachusetts: Addison Wesley, 2014.

[13] B. Boehm, J.A. Lane, S. Koolmanojwong, and R. Turner, The incremental commitment spiral model: Principles and practices for successful systems and software. Boston, Massachusetts: Addison-Wesley, 2014.

[14] S. Mathur and S. Malik, "Advancements in the V-Model," International Journal of Computer Applications, vol. 1, no. 12, pp. 29-34, 2010.

[15] G. Whyte and A. Bytheway, "The V-model of service quality: an African case study," European Journal of Marketing, vol. 51, no. 5/6, pp. 923-945, 2017.

[16] A. Finkelstein and J. Kramer, Software engineering: a roadmap. New York City, New York: ACM Press, 2000.

[17] J. Ludewig, "Models in software engineering–an introduction," Software and Systems Modeling, vol. 2, no. 1, pp. 5-14, 2003.

Mohammad Samadi Gharajeh received ASc in Computer Software on 18 February 2005, BSc in Engineering of Computer Software Technology on 18 February 2009, and MSc in Computer Engineering – Computer Systems Architecture on 3 February 2013. He has already developed various software programs, simulations, intelligent systems, and research projects. His research interests include software engineering, artificial intelligence, soft computing, intelligent control, and embedded systems. He was a Technical Program Committee member and a Reviewer in some of the international conference proceedings. Besides, he is an Editorial Board member and a Reviewer in some of the international scientific journals, a Lecturer of university, and an IEEE Member now.

Organizational development through Business Intelligence and Data Mining

Denis-Cătălin ARGHIR, Ioana-Gilia DUŞA, Miruna ONUŢĂ
The Bucharest University of Economic Studies, Romania
arghir.denis@gmail.com, ioanagd94@gmail.com, miruna.onuta@gmail.com

The article presents the concept of Business Intelligence and their influence on decision making. Examining Business Intelligence systems was accomplished by theoretically comparing of four systems: Microsoft Power Bi, IBM Cognos, Oracle BI, and SAS, focusing on "functionality", "performance", "usage" and "cost" criteria. Functionality testing was done through the Power BI system using a HORECA industry dataset, namely a café retailer. On this dataset has been applied data mining concepts as cluster analysis, KNN classification analysis, and association study, to determine the frequently encountered templates, to categorize buyers into various key categories, and to help the business thrive.
***Keyword**: Business Intelligence, Power BI, Data Mining, Apriori Algorithm, Cluster Analysis, KNN Analysis.*

1 Introduction

The roots of the business intelligence (BI) concept dates back to the nineteenth century when the term BI was originally invented by Richard Millar Devens in the paper "Encyclopaedia of Commercial and Business Anecdotes" published in 1865.
According to Devens [1], the concept was used to describe how Sir Henry Furnese's banker was able to make a profit by receiving information about the banking environment and how he acted before his competitors.
The capacity to collect information and to react on the basis of these denotes the ability showed by Furnese and which underlies the concept of Business Intelligence today.
The next century's development expanded and refined business by first introducing the term "Business Intelligence" in 1989 by Howard Dresner from Gartner Group, which defined this concept, according to [2], as "a method of improving decision-making through the use of fact-based support systems".
Although it is related to enterprise applications, Business Intelligence is not a product or a system, it is a concept that shelters architectures, applications, and databases. Its purpose is to access user's data from an organization as easy as possible by interactive, in real-time access of databases, and also manipulation and analysis of them.
By analysing historical data, BI performs a valuable insight into business activities and business situations, and managers are actually assisted in making decisions through essentials information, including those behavioural and of forecasting.
In the current sense, this term denotes a set of concepts and methods used to improve the quality of business decision-making process and represents a platform for presenting information in a correct, useful and capable way to support the daily activities and decisions of every person in management positions in order to choose the most efficient alternatives.
Business Intelligence is made up of a series of applications and technologies that help gather, store, query, report, and analyse large volumes of data, and provide access to necessary data in company decision-making processes by obtaining analyses and reports.
With today's BI solutions, managers can analyse data directly without needing help from IT staff and without waiting for complex reports to run. This democratization of access to information helps users make informed decisions based

on concrete facts - not on suspicions and instincts.

2. Fields of application of BI

The effects of using a BI system are stunning, because it produces the needed information, at the time it is necessary, providing one of the prerequisites for business success. BI is the art of knowing and harnessing information, gaining competitive advantages.

BI can provide answers to the core issues of an organization, helping it making good decisions to resolve it. Finding answers is based on analysing and comparing historical data, both created within the organization, and data from external sources.

The providing benefits of the BI system, unconcerned of the field of activity, are varied, for example: a producer can quickly find out the need for materials, raw materials or stock based on past sales; a sales manager can create more profitable sales plans following the evolution of the previous period; a distributor can find out the most profitable distribution channels; a service provider has the possibility to anticipate and identify loyalty programs.

3. The architecture of a BI system

The architectural model of a BI system can include the following components:

- Data source - can be extracted from various sources or systems, such as Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), relational databases, Excel spreadsheet files, Comma Separated Values (CSV) or text files (TXT); Once the data has been extracted from external sources and has been transformed according to operational needs, the data is loaded to achieve the final goal, this process is called *Extract Transform Load* (ETL);
- Data processing - once the data has been loaded to the final target, either Data Warehouse or Data Mart can be processed, can be added a series of new data, can be maintained records in a form of data logging;
- Analysis and data presentation - there are various analysis tools, analyses can be made using applications such as:
  - OLAP – useful for dynamic data analysis, fast access to a large amount of data, synchronization of data sources from multiple databases, historical analysis based on time series;
  - DATA MINING – useful for analysing large data sets in order to identify models and relationships for establishing future tendencies - clustering, association, classification;
  - DASHBOARD - useful for a quick view of performance indicators relevant to a business process.

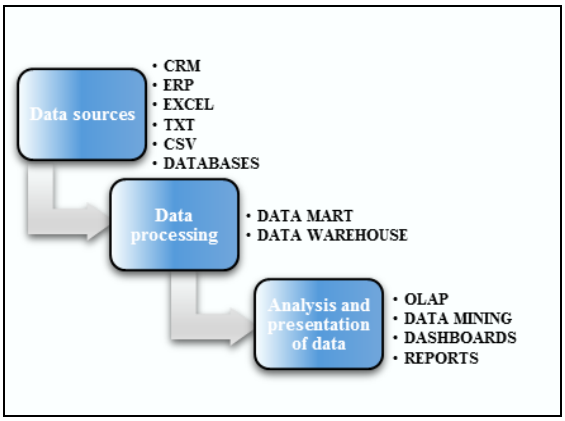


Fig. 1. The architecture diagram of a BI system

In a survey conducted in 2018 [3] on a sample of 600 companies from various industries interested in business intelligence software, has been achieved the top of the most wanted functionalities, which is as follows:

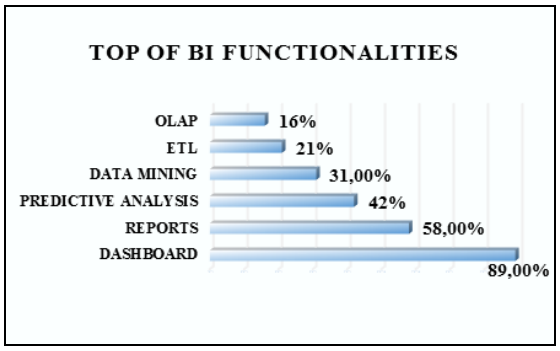


Fig. 2. Top of the most wanted BI features

4. Business Intelligence tools

By using BI tools can be obtained useful information that allows the user to understand at a glance the current state of some relevant business indicators. The tools are varied and start from simple *spreadsheets*, *Online Analytical Processing* (OLAP) - dynamic reporting solutions that allow users to interact with complex structures as time series, data trend; *Dashboard systems*; *Data Mining* – the process of extracting patterns from a large volume of data by combining statistical and artificial intelligence methods with those from database management; *Decision Engineering* – provides the framework that unites a number of good practices for organizing decision-making process; *Data Warehouse* - a data repository designed to facilitate an organization's decision-making process; *Process mining* – extracting knowledge from events recorded by the informatics system; *Exploratory Data Analysis* (EDA) – the exploration of a data set, without a strong dependence on assumptions or models, the objective being to identify patterns in an exploratory manner; *Business performance management* – is a set of managerial and analytical processes that allow the performance management of an organization to achieve one or more purposes.

Among the most popular tools can be listed Microsoft BI & Excel, Oracle BI, IBM Cognos, SAS, Qlik, Tableau, SAP Business Objects, and in terms of accessing business intelligence solutions, the most used by companies are, according to [4]:

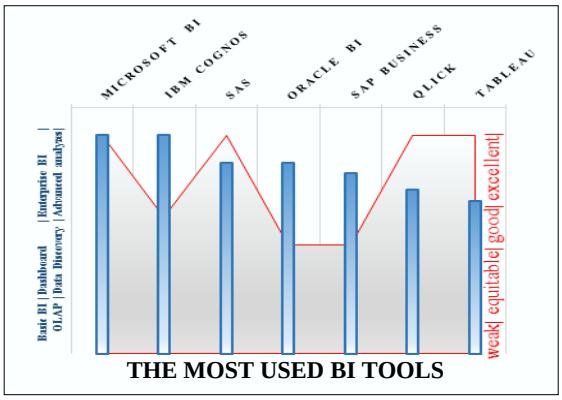


Fig. 3. Most popular BI tools

5. The purpose of using Business Intelligence

Business Intelligence has become a strategic tool to help a company to lead, optimize, discover and innovate to change the landscape of its organization. Business Intelligence systems are useful to modern businesses because they have the ability to provide a continuous flow of information and the capabilities of such a system that implements the BI concept allow employees:

- to align day-to-day operations with overall goals and strategies;
- to identify and understand the relationships between business processes and their impact on performance;
- to access relevant information for analytical responsibilities specific for the analysis;
- to analyse data from documents and to develop them very easily;
- to monitor vital business indicators, such as the current financial reports, the effectiveness, and profitability of sales departments or other relevant measurement indicators.

Business Intelligence represents the ability of an organization to think, plan, predict, solve problems, abstract thinking, understand, innovate, and learn in ways that enhance organizational knowledge, inform in the decision-making process, to

allow effective actions and to help establish and achieve business goals. The role of business intelligence is to create an informational environment in which operational data collected from transactional systems and external sources can be analysed to discover strategic business dimensions.

This information should help organisations to respond to business key issues, make predictions, and act on real-time data to improve the quality and speed of the decision-making process.

Expected benefits do not always justify investing in business intelligence technology.

This means that the development of business intelligence capabilities can only provide information-based decisions, but they cannot implement them.

An analysis of the impact of business intelligence should not focus on the impact at a given point in time, but it should be longitudinal to determine how and why it varies over time. Similarly, to enterprise resource planning systems, there are case studies that have examined long-term success or failure of business intelligence.

On a more theoretical side, in order to substantiate business intelligence research within the research information systems, rigorous preparation based on theory and impact analysis is required.

Given that the rapid evolution of technologies and managerial methods is a significant challenge for theoreticians, most of the previous business intelligence research has not had a theoretical foundation.

In this analysis, the organizational theory of information processing is used to analyse how new business intelligence technologies can favourably enhance information processing capabilities.

Business intelligence clearly reduces decision-making risk and directs operational and marketing activities to generate real value and which can be capitalized, with minimal resources.

Business intelligence projects are not meant to teach managers how to make the right decisions, but instead, help them make decisions based on facts and figures rather than assumptions.

6. The elements that turn BI into a viable business solution

By using a Business Intelligence solution, business people have access to current and quality information, highlighted in a visual and effective way.

Many organizations implement business intelligence systems, but their long-term impact on the quality of the decision-making process and of the performance consequently varies greatly.

An analysis of the factors that influence the continued use of these systems is required and is usually focused on the need for information processing in the continued use of business information and the factors that influence these needs.

Business complexity means, today, that a company needs to perform regularly complex analysis with vast amounts of data.

Many businesses are now implementing business intelligence systems to get timely information about organisational processes and company environments that combine information about past circumstances, present events and projected in future actions to answer to questions that solve various problems.

Business intelligence solutions have seen unprecedented growth over the past decade, and companies that offer them have seen spectacular growth despite all the vicissitudes of the economic environment.

Functionalities offered to users have become increasingly varied covering a wide range of needs ranging from simple tabular reports or graphical reports to the ability to track the organization's main performance indicators in a synthetic and concise manner.

To the extent that Romanian companies want to survive the pressure of European

competition, business intelligence solutions can provide them the necessary means to do this. The only remaining problems are those related to the wish of companies and those time-related because the market already has solutions for any budget.

The benefits of a business intelligence system are obvious - the analysts are optimistic, showing that in the coming years, millions of people will use day-to-day visual analysis tools and BI. The market is already saturated with the range of analytical applications available, which can carry out all sorts of analysis to support decision-making process at all levels.

Other benefits to be taken into account are:

- Reducing downtime spent with periodic reporting activities (collection of reports, consolidations, and various adjustments);
- Reducing time spent with repetitive activities;
- Reduce the role of the IT department in generating reports in favour of the end-user;
- Reduce the time needed to make a decision.

Given that the decision will be even better documented due to the quality of the information provided, we will finally be able to talk about an organization prepared to face any changes in the market, no matter how abrupt they are.

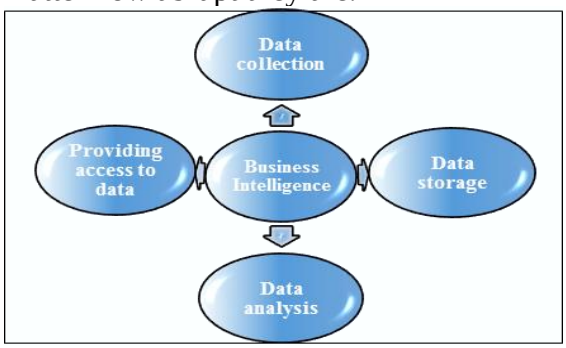


Fig. 4. Benefits provided by BI solutions

7. The Influence of Business Intelligence on Decision Making

Information from business intelligence needs to be integrated into the business process of an organization. This can be achieved by building a decision-making system. The results obtained from this business information are used by operational managers in the form of recommended actions. A prerequisite condition for efficient use is the availability of high-quality data, including good data management, covering:

- identifying users' needs;
- unification of data;
- clearing data;
- improving data quality control.

Most business intelligence initiatives have therefore focused on developing a high-quality business intelligence asset that is used instead of classical reporting systems.

The value of Business Intelligence derives from the ability to extract specific data and to adapt it from a variety of heterogeneous sources.

Managing an enterprise requires efficient data management in order to monitor activities and to evaluate the performance of different business processes.

Nowadays, there are a number of changes in the world of analysis, even for the long term, where BI is beginning to struggle to adapt to new trends. The information has become a profit centre, and processes are now customer-oriented, so business people to have a say about the mode in which analyses are predicted.

With increasing expectations, BI systems are looking to constantly improve their capabilities, because the need for faster data processing has increased, especially that the most data are not in the internal system, companies using information from outside the BI environment of enterprises.

8. Advantages of using Business Intelligence systems

At this moment, the business environment has favoured the spreading of BI applications. There are industries that budget big spending on technology purchases and BI are more or less obvious initiatives that lead to improved

profitability rates. Practically, these applications help to make wise decisions.
BI applications allow the ability to summarize and aggregate by specific and detailed categories at the same time, specific to a particular analysis or process, presenting the exact information and excluding the extra elements. Thus, a decision-maker can monitor the performance variables of a business. Example: sales per region, per product, per quarter, or product return rate for various reasons, customer behaviour analysis based on specific preference analysis.

- Possibility to retrieve data from various computer systems and to carry out a detailed analysis of them for decision support;
- Identifying and adjusting defective processes, even modifying the logic of performing certain activities to meet the company's efficiency requirements;
- Development of modules specifically designed for every single requirement;
- Permanently communication with all the centres for accurate, up-to-date and easy to follow reporting;
- The ability to develop Data Warehouse solutions that support the most demanding reporting requirements;
- Real-time updating of transactional data on which to build a decision support system;
- A business BI system is simple, visual and easy to understand, giving companies the freedom to answer questions immediately as they occur;
- The possibility of creating interactive views in just a few seconds even when working with very large volumes of information.

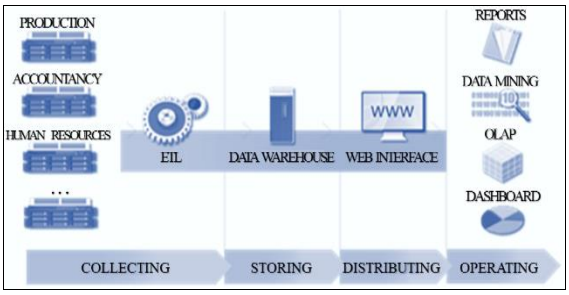


Fig. 5. Advantages of using BI systems [5]

9. Comparison between BI tools

We chose for comparison four BI tools, surprising different aspects (functionalities, performance, usage, costs) to find out which is the most suitable tool for organisational development:

Microsoft Power BI
FUNCTIONALITIES
-*exposure*: there are multiple versions: Power BI Desktop – on-premise version, Power BI service **app.powerbi.com** - Software as a Service (SaaS) online version and Power Bi Mobile - Android, iOS and Windows mobile device version; -*interactivity*: creates real-time reports, creates data analysis models, assures quality, reliability, and scalability of the reporting system; -*modelling*: dynamic and conducive environment for build comprehensive and relevant real-time reports with the possibility of automatically updating them; -*presentation*: graphics - pie, column, line, matrix, Excel reports, dashboards; -*data source*: Excel spreadsheets, Access/ SQL databases, XML files, flat text and CSV files;
PERFORMANCE
-*data integration*: creating relations between tables of type “primary key” - “foreign key”; pivot operations; -*data processing*: data can be processed from Excel spreadsheets or from a local database;
USAGE
-*interface*: depending on the type of version you are working with, data can be viewed as reports, custom dashboards, cubes - either in the form of Mobile or Desktop;
COSTS
-moderate costs;

IBM Cognos

FUNCTIONALITIES
-*exposure*: various tools: Report Studio (interactive and complex report developer tool), Query Studio (simple query and report creation tool), Analysis Studio (multi-dimensional analysis tool that provides drag & drop functionalities for exploration, analysis and comparing large data in a very short time); -*interactivity*: allows filtering, sorting, performing additional calculations, transforming values into graphs and charts; -*modelling*: provides a supportive environment for planning, budgeting, making forecasts and reliable plans in a short time; -*presentation*: export into images, Excel, PPT, PDF; -*data source*: Excel spreadsheets, XML files, flat text and CSV files;
PERFORMANCE
-*data integration*: individual queries can be joined using SQL commands; -*data processing*: data is delivered from cubes to memory for improved performance;
USAGE
-*interface*: provides a web-based architecture with advanced creative capabilities; the dashboard allows desktop customization and access to content;
COSTS
-lower costs than other traditional BI products;

SAS
FUNCTIONALITIES
-*exposure*: various visualization capabilities implemented in the product suite, interface that allows interaction with charts; -*interactivity*: users interact with a dashboard - Dashboard Builder;
-*modelling*: interaction with predefined algorithms and pre-built models for modelling datasets; -*presentation*: export as Excel files, TSV (Tab-Separated Values), CSV etc.; -*data source*: SAS datasets, Microsoft Excel spreadsheets, flat text or CSV files;
PERFORMANCE
-*data integration*: administration of server-based libraries; -*data processing*: is achieved either on the local unit or on the server;
USAGE
-*interface*: interactive, web-based reporting interface, allows the creation of basic queries and reports;
COSTS
-high costs;

ORACLE BI
FUNCTIONALITIES
-*exposure*: multiple view capabilities-basic charts, intuitive reports, diagrams; -*interactivity*: allows filtering, creating interactive parameters; -*modelling*: predictive and reporting functions for dedicated financial planning; -*presentation*: export to PDF, RTF, XML, HTML, Excel and other formats; -*data source*: supported file types are database files, XML flow, HTTP, Web services, Oracle BI analysis, OLAP cubes, LDAP server, XML files, Excel spreadsheets;
PERFORMANCE
-*data integration*: it is created using SQL commands based on table and column queries; they are specified as well as relations between them; -*data processing*: in-memory processing;
USAGE
-*interface*: web-based interface, easy to use for creating reports, with Template Builder functionality;
COSTS
-high prices for large configurations;

By comparing the four tools, we've decided to look in-depth on the features of "Power BI"-a collection of software services, applications and connectors that work together to transform unrelated data sources into a coherent and interactive perspective.

It is a simple and fast system able to create complex analyses based on various data sources, whether simple flat files like text or CSV, Excel spreadsheets or local databases.

Robust and high-quality for organizations, ready for extensive modelling and real-time analysis as well as personalized development, Power BI enables users to easily connect to data sources, visualize and discover what is important for the common good of the business. It can also

serve as an engine for analysis and decision-making for group, divisions, or entire corporations projects. More and more companies from Romania use this tool to monitor business status using dashboards that process data in real time. The primary reason for choosing this tool is that it offers the ability to work both in cloud and on-premise, that it can easily build robust and reusable models using available data, ensuring consistency between reporting and analysis within the organization, and with the Power BI web version and it can distribute various reports in just a few seconds across the organization's departments.
Also, an important feature that has convinced companies in Romania to adopt this tool is the existence of the Mobile version, so managers, the end-users, can have a view of the data anywhere and at any time. They can view custom reports and dashboard, find important information in a due time, and act immediately to reassess situations.

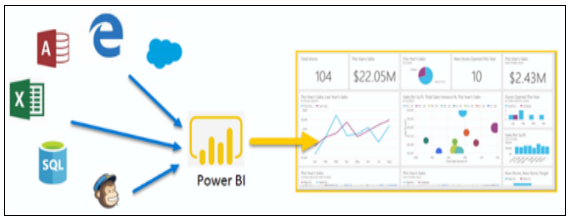


Fig. 6. Power BI Workflow

10. Case Study

Concluding that Power BI is a very useful tool for Business Intelligence analyses, we've decided to test some of the available functionalities on a specific dataset.
The analysed data is a test one and was extracted from „UCI Machine Learning Repository" (*archive.ics.uci.edu [6]*), a website that includes a structured database in various fields across the globe, with varying sizes indicators and instances, classified on different tasks for which can be used, including data mining analysis such as clustering, classification, regression, etc. We chose to study HORECA industry data, from a retailer, namely a coffee shop selling "Delicacies" products.

Why did we choose this data?

We took the data to apply business intelligence analyses, but also of data mining algorithms such as cluster analysis, KNN classification analysis and association study to extract from the dataset the most relevant information that could be used in the business environment, such as grouping data in smaller clusters in order to be able to analyse them as closely as possible, to predict the affiliation of certain classes of a test set, starting from the training set, and to track consumer behaviour on consumption of a particular category of products, in our analysis - "Delicacies", by applying different filters to see how a product sale may be influenced by another product or a mix of products.
After identifying and determining the purpose of the analysis, we imported the data we wanted in comma-separated values (CSV) format into the Power BI tool, assigning data types for each variable based on the data content.

-**Association study**: The submitted data for analysis are "Delicacies" products marketed by a café from the HORECA chain (**Ho**tels - **Re**staurants - **Ca**fes/ Coffee Shops).
For the association study, we used the "Apriori" algorithm. According to [7], by association study, it is desirable to determine consumers' consumption behaviours to find interesting and frequently encountered templates that could help the business to earn more.

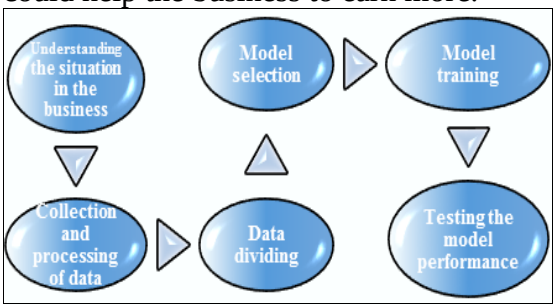


Fig. 7. Steps for implementing analysis

After understanding the need of applying the study and collecting the datasets, we switched to processing them, so, using the BI tool -we concatenated (Merge) the product type with the main ingredient; -we grouped the data set by the order_id; -we transformed the new table, containing the products grouped by the order_id, in a customized list column, with the sequence: ***(Table.Column ([Data], "Product")***; -the list we distributed on the columns (Split) after the delimiter ";" in order to be able to apply the association rules analysis.
To exemplify on a demonstration dataset, we chose 4 purchases made by coffee shop customers:

Fig. 8. Test data for the association study

By distributing the products purchased by each of our 4 customers in table form, we obtained the frequency of purchased products:

Table 1. Frequency of purchased products

ORDER	Tartă Coacaze	Apă Plată	Choux Cafea	Cafea Espresso	Desert Caise
1	1	1	1	1	0
2	0	1	0	0	0
3	0	0	0	0	0
4	1	1	0	0	1
frequency	**50%**	**75%**	**25%**	**25%**	**25%**

ORDER	Tartă Vișine	Prajitură Lămâie	Tartă Lămâie	Prajitură Mascarpone	Prajitură Căpșuni
1	0	0	0	0	0
2	0	1	1	0	0
3	0	1	1	0	0
4	1	0	0	1	1
frequency	**25%**	**50%**	**25%**	**25%**	**25%**

To determine the percentage of how often items appear together in a total of transactions, we calculated the *support* level, one of the features of the association rule:

$$\textbf{support(X)}=\frac{\textbf{count(X)}}{\textbf{N}}$$

Where:

- *X=the frequency of occurrence of articles (products);*
- *N= total number of transactions/ orders;*

Using the previous formula on the same dataset, it appears that in the case of the combination of "Tartă-Coacăze" (Blueberries Tart) and "Apă-Plată" (Non-carbonated Mineral Water) there is a frequency of 2 appearances in the four orders analysed, resulting a support of 0.5, meaning that 50% of purchases contained the 2 products.

$$\text{support(Apă-Plată \& Tartă-Coacăze)}=$$
$$=\frac{2}{4}=0,5 \quad (1)$$

In order to determine the ratio between the number of customers who buy items that appear as a rule and the number of buyers of the items that appear in the antecedent, will be calculated with another feature of the association rule - *confidence*:

$$\textbf{confidence(X}\rightarrow\textbf{Y)}=\frac{\textbf{support(X,Y)}}{\textbf{support(X)}}$$
$$\text{support(Apă-Plată)}=\frac{3}{4}=0,75 \quad (2)$$
$$\text{support(Tartă-Coacăze)}=\frac{2}{4}=0,5 \quad (3)$$
$$\text{confidence(Apă-Plată}\rightarrow\text{Tartă-Coacăze)}=$$
$$=\frac{(1)}{(2)}=\frac{0,5}{0,75}=0,(6) \quad (4)$$
$$\text{confidence(Tartă-Coacăze}\rightarrow\text{Apă-Plată)}=$$
$$=\frac{(1)}{(3)}=\frac{0,5}{0,5}=1 \quad (5)$$

We can say, according to (4), that in 60% of the cases, the buyer who bought "Apă-Plată" also bought "Tartă-Coacăze", and according to (5), 100% of the cases "Tartă-Coacăze" were bought together with "Apă-Plată".
To see to what extent the association rule is useful, we will calculate the degree of improvement - *lift*:

$$\textbf{lift(X}\rightarrow\textbf{Y)}=\frac{\textbf{confidence(X}\rightarrow\textbf{Y)}}{\textbf{support(Y)}}$$
$$\text{lift(Apă-Plată}\rightarrow\text{Tartă-Coacăze)}=$$
$$=\frac{0,(6)}{0,5}=1,33 \quad (6)$$
$$\text{lift(Tartă-Coacăze}\rightarrow\text{Apă-Plată)}=$$
$$=\frac{1}{0,75}=1,33 \quad (7)$$

We can say that both rules are relevant, since they exceed the unit value.
To apply the association algorithm across the dataset, we used the R scripting

functionality within Power BI Desktop as follows:

Having the set called "dataset" as an input, we transform it as transactions by applying the "Apriori" function to find association rules, providing the minimum support and trust levels as parameters.

The first 100 results we recorded in a variable called "output", which is the result of the query.

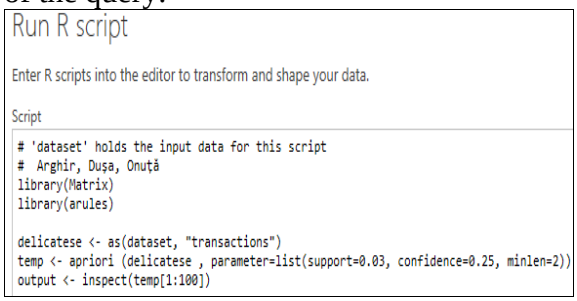
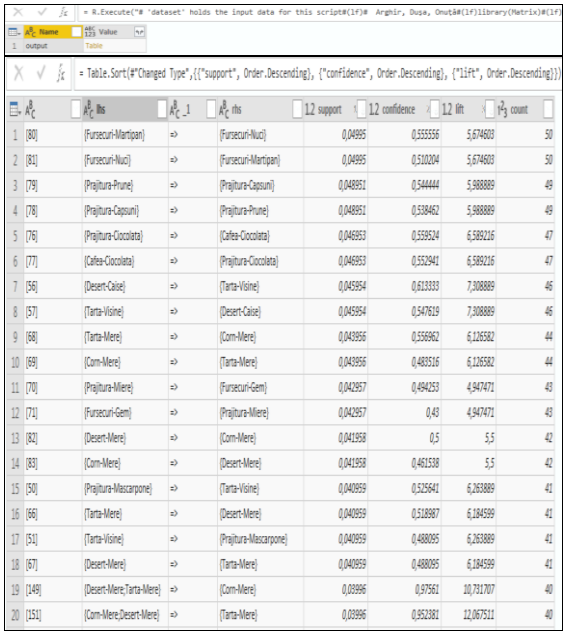


Fig. 9. The R script for obtaining the associations

Thus, the result of the query is generated in Power BI in tabular format and it can be observed the newly generated columns: "lhs" (the primary product purchased), "rhs" (the associated purchase product) and columns "support", "confidence", "lift" (degree of improvement) and "count" (number of appearances of combinations).



	lhs		rhs	support	confidence	lift	count
1	{80}	{Fursecuri-Martipan}	⇒	{Fursecuri-Nuci}	0,04995	0,555556	5,674603	50
2	{81}	{Fursecuri-Nuci}	⇒	{Fursecuri-Martipan}	0,04995	0,510204	5,674603	50
3	{79}	{Prajitura-Prune}	⇒	{Prajitura-Capsuni}	0,048951	0,544444	5,988889	49
4	{78}	{Prajitura-Capsuni}	⇒	{Prajitura-Prune}	0,048951	0,538462	5,988889	49
5	{76}	{Prajitura-Ciocolata}	⇒	{Cafea-Ciocolata}	0,046953	0,559524	6,589216	47
6	{77}	{Cafea-Ciocolata}	⇒	{Prajitura-Ciocolata}	0,046953	0,552941	6,589216	47
7	{56}	{Desert-Caise}	⇒	{Tarta-Visine}	0,045954	0,613333	7,308889	46
8	{57}	{Tarta-Visine}	⇒	{Desert-Caise}	0,045954	0,547619	7,308889	46
9	{68}	{Tarta-Mere}	⇒	{Com-Mere}	0,043956	0,556962	6,126582	44
10	{69}	{Com-Mere}	⇒	{Tarta-Mere}	0,043956	0,483516	6,126582	44
11	{70}	{Prajitura-Miere}	⇒	{Fursecuri-Gem}	0,042957	0,494253	4,947471	43
12	{71}	{Fursecuri-Gem}	⇒	{Prajitura-Miere}	0,042957	0,43	4,947471	43
13	{82}	{Desert-Mere}	⇒	{Com-Mere}	0,041958	0,5	5,5	42
14	{83}	{Com-Mere}	⇒	{Desert-Mere}	0,041958	0,461538	5,5	42
15	{50}	{Prajitura-Mascarpone}	⇒	{Tarta-Visine}	0,040959	0,525641	6,263889	41
16	{66}	{Tarta-Mere}	⇒	{Desert-Mere}	0,040959	0,518987	6,184599	41
17	{51}	{Tarta-Visine}	⇒	{Prajitura-Mascarpone}	0,040959	0,488095	6,263889	41
18	{67}	{Desert-Mere}	⇒	{Tarta-Mere}	0,040959	0,488095	6,184599	41
19	{149}	{Desert-Mere,Tarta-Mere}	⇒	{Com-Mere}	0,03996	0,97561	10,731707	40
20	{151}	{Com-Mere,Desert-Mere}	⇒	{Tarta-Mere}	0,03996	0,952381	12,067511	40

Fig. 10. Output – the result of Apriori algorithm application

With "Forced-Directed Graph 2.0.2" visualization mode in Power BI, together with the "Slicer" filter, we represented in an interactive way the relationships between the main and the related nodes, the link thickness representing the value of the support, with the possibility of selection from the drop-down list the main or associated products.

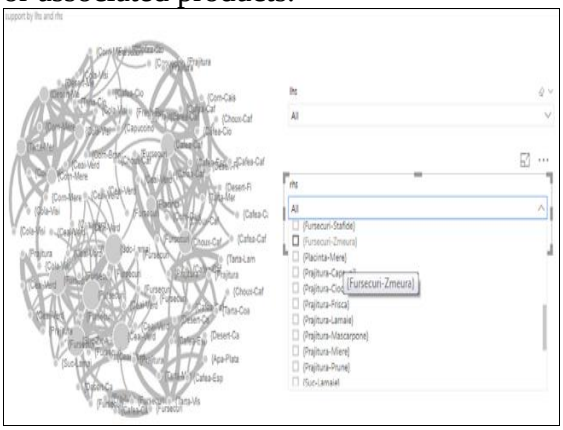


Fig. 11. Representation of associations as a graph (Forced-Directed Graph)

For example, we have selected combinations of main product associated with "Fursecuri-Zmeură" (Raspberry Cookies) and resulted in the following graph, where can be viewed at a glance which are the most preferred customer combinations of products (the most thicker lines):

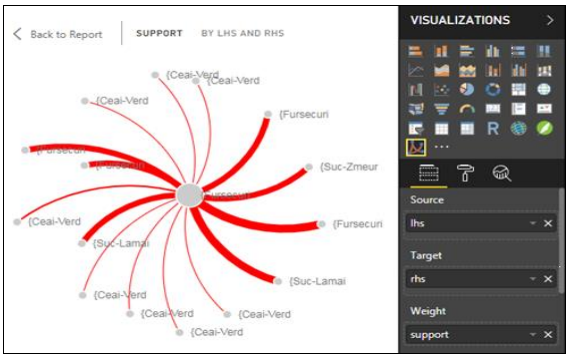


Fig. 12. Extracting all combinations of purchases that have associated the product „Fursecuri-Zmeură"

The strongest link, in the sense of the highest support, is in the case of the combination "Suc-Lămâie" (Lemon-Juice) associated with "Fursecuri-Zmeură" with a support level of 0,03, meaning that 3% of

the purchases (31 purchases) contained the 2 products.

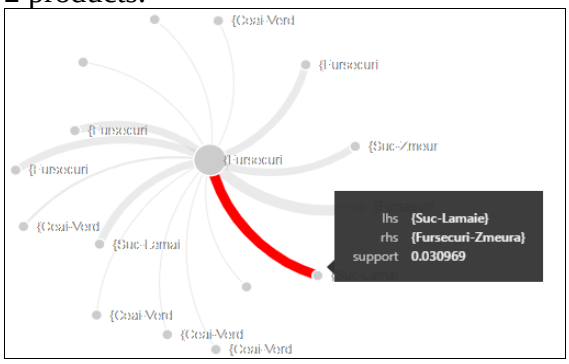


Fig. 13. Detailing the relationship with the highest support level 0,03% „Suc-Lămâie" associated with „Fursecuri-Zmeură"

Another representation can be made using the "R script visual" view:

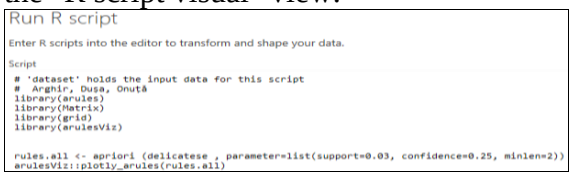


Fig. 14. R script for obtaining a scatterplot representation

Thus, the three dimensions - "support", "confidence" and "lift" can be viewed in a single graph that can be detailed with a simple click:

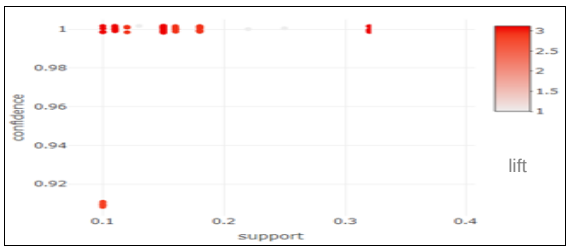


Fig. 15. Scatter-plot representation for product associations (support, confidence and lift)

Following the association study, the management of the unit can create promotional packages that could have positive results in increasing sales of "Delicacies" products.

It follows the use of another data mining analysis that is applicable to business intelligence:

-Cluster Analysis: According to [8] the purpose of clustering occurs from the need to fit, group or classify certain entities or objects in the form of categories or classes, of which delimitation must be very clear. All classification techniques are known as the theory of form recognition.

Cluster analysis is a classification technique characterized by the fact that affecting the forms or objects in clusters or groups is made progressive and without aprioristic knowing the number of classes, depending on verification of two fundamental criteria:

- the objects or forms classified in each class needs to be as similar in terms of certain characteristics;
- objects classified in a class needs to differentiate as much as possible from objects classified in any of the other classes.

The K-means algorithm is a method of dividing a dataset into a specified cluster number introduced by the user (k). This cluster analysis method aims to divide "n" observations in "k" classes where each observation belongs to the class with the closest average.

More specifically, the algorithm assigns "k" centres of the classes (centroids) in "n" points.

The steps of the k-Means algorithm:

- firstly, we chose the initial "k" group centres;
- the "k" groups are created by assigning each instance of that group to whose centre it is closest;
- we recalculate the centres in relation to the new composition of the groups;
- we repeat the algorithm with the second step if the centres have moved;

Table 2. Test data for cluster analysis

ID	PRODUCT	Fats	Carbs	Proteins	Calories	Sugars
0	Căpşuni	1.83	23.88	1.85	110	18.44
1	Lămâie	6.84	23.20	2.02	170	15.38
2	Frişcă	8.15	23.37	2.40	171	14.48
3	Mascar-	7.54	19.46	1.85	151	12.07

	pone					
4	Ciocolată	4.68	21.26	1.62	129	12.50
5	Miere	2.84	28.75	2.25	140	15.16
6	Ciocolată neagră	0.43	23.38	2.10	100	16.40

For exemplification, we chose a demonstrative dataset with 6 records, representing 6 products marketed by the café, with several nutritional features (fats, carbohydrates, proteins, calories and sugars) to promote them in 2 packages, generically called "*dietetic*" and "*caloric*", so we want to create two clusters with these features.

We have selected from the dataset the product with id "2", as having the highest values, and the product with id "6" as having the smallest values around which the two clusters will be created.

Table 3. The chosen values as initial centroids for k=2 clusters

Cluster	Product_ID	Average (centroid)
1	6	(0.43, 23.38, 2.10, 100, 16.40)
2	2	(8.15, 23.37, 2.40, 171, 14.48)

Then, we calculate the distance of each element from the two clusters, using the Euclidean distance:

$$d(p,q)=\sqrt{(p1-q1)^2+(p2-q2)^2+\cdots+\sqrt{+(pi-qi)^2+\cdots+(pn-qn)^2}}$$

$$d(p,q)=\sqrt{(0,43-1,83)^2+(23,38-23,88)^2+\sqrt{+(2,10-1,85)^2+(100-110)^2}\sqrt{(16,40-18,44)^2}}=10,31 \quad (7)$$

$$d(p,q)=\sqrt{(8,15-1,83)^2+(23,37-23,88)^2+\sqrt{+(2,40-1,85)^2+(171-110)^2}\sqrt{+(14,48-18,44)^2}}=61,46 \quad (8)$$

By comparing the distances to the two clusters, we decide that the product with id "0" is closer to Cluster 1. At this point, we need to calculate the new average (media) value of Cluster 1 (centroid):

Media:

$$(0,43+1,83)/2=1,13$$
$$(23,38+23,88)/2=23,63$$
$$(2,10+1,85)/2=1,98$$
$$(100+110)/2=105$$
$$(16,40+18,44)/2=17,42$$

Applying formulas (7), (8) and (9) for the 7 products, are obtained the following solutions:

Table 4. Categorizing products into the 2 clusters

Cluster	Product_ID	Average (centroid)
1	6	(0.43, 23.38, 2.10, 100, 16.40)
	6,0	(1.13, 23.63, 1.98, 105.00, 17.42)
	6,0,4	(2.91, 22.45, 1.80, 117.00, 14.96)
2	2	(8.15, 23.37, 2.40, 171, 14.48)
	2,1	(7.50, 23.29, 2.21, 170.50, 14.93)
	2,1,3	(7.52, 21.37, 2.03, 160.75, 13.50)
	2,1,3,5	(5.18, 25.06, 2.14, 150.38, 14.33)

Thus, we obtain the categorization of the 7 products in the two clusters: Cluster 1 *(products 6, 0, 4)* and Cluster 2 *(products 2, 1, 3, 5)*.

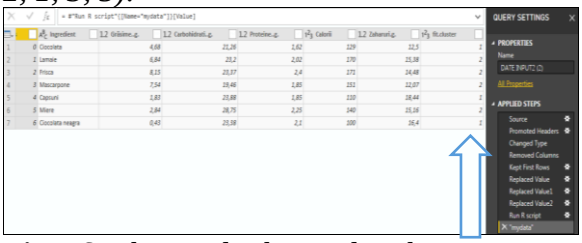


Fig. 16. The result obtained with Power BI on the test set

Returning to the whole dataset, applying the K-Means algorithm with Power BI's R scripting functionality, we have grouped the 50 products in 3 classes, which we can use to promote the products of the café in 3 packages: "*dietetic*", "*medium*" and "*caloric*".

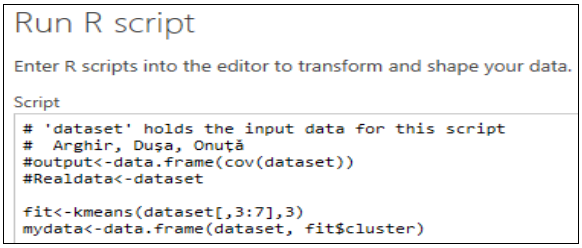
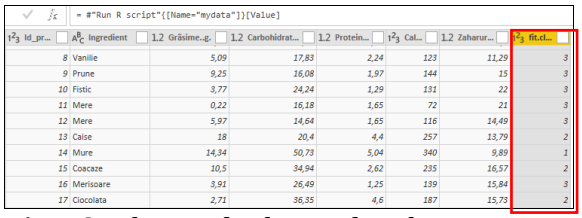


Fig. 17. R script to obtain categorisation of products ranges in 3 clusters

We obtained the categorisation of 50 products in 3 clusters:



Id_pr...	Ingredient	Grăsime...g.	Carbohidrat...	Protein...	Cal...	Zaharur...	fit.cl...
8	Vanilie	5,09	17,83	2,24	123	11,29	3
9	Prune	9,25	16,08	1,97	144	15	3
10	Fistic	3,77	24,24	1,29	131	22	3
11	Mere	0,22	16,18	1,65	72	21	3
12	Mere	5,97	14,64	1,65	116	14,49	3
13	Caise	18	20,4	4,4	257	13,79	2
14	Mure	14,34	50,73	5,04	340	9,89	1
15	Coacaze	10,5	34,94	2,62	235	16,57	2
16	Merisoare	3,91	26,49	1,25	139	15,84	3
17	Ciocolata	2,71	36,35	4,6	187	15,73	2

Fig. 18. The result obtained with Power BI on the whole dataset

We used a Treemap representation to highlight the three newly formed clusters, grouping being done by the "cluster" attribute, the details being represented by "product names" and the values being taken from the attributes "calories", "fats", "proteins" and "sugars".

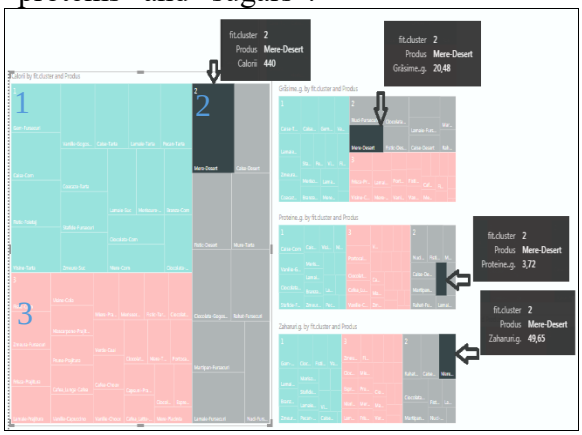


Fig. 19 Treemap representation of obtained clusters

Cluster 3 contains the products with the lowest caloric value, and can call them "dietetic", Cluster 1 is the next one, which can be called "medium" from the caloric point of view, and Cluster 2 contains the most "caloric" products.

For easier visualization of grouping in clusters, we created a "Clustered Bar Chart" for four of the product features and two "Slicer Drop Down List" to compare the average values of "calories", "fats", "proteins" and "sugars" from two clusters at the same time. For not interacting Slicers in the wrong way with all objects, we have blocked -Edit interactions: None- for charts that should remain unchanged:

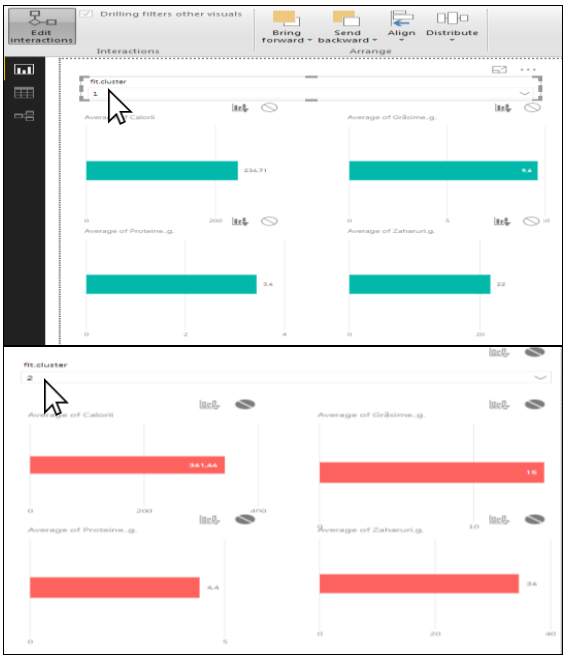


Fig. 20. Comparing the average nutrient values of products in Cluster 1 versus Cluster 2

-The classifier based on K nearest neighbours: to classify the instances, we choose the KNN (K-nearest-neighbours) algorithm, being, according to [9], a simple classification method based on placing all instances in an n-dimensional space.

As mentioned above, we have 3 groups in which the products are divided, namely packages "*dietetic*", "*medium*" and "*caloric*". The coffee shop wants to introduce 16 new products by assigning one of the three attributes according to nutritional value, taking into account the current menu that is already listed with nutrition attributes (34 products).

Thus, we want to identify the proximity of newly introduced products to existing categories.

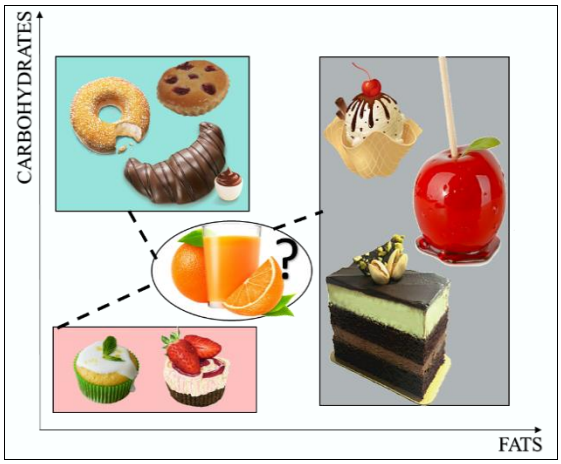


Fig. 21. Graphic representation of the distance between products

To calculate, for example, the category in which the "Fresh-Portocale" *(Orange Juice)* will be placed in the existing range, we will have to calculate the distance between the new introduced product and each products existing in the café menu, using the Euclidean metric, following the reasoning presented at (7), (8); the calculation formula for the n-dimensional case is:

$$d(p,q)=\sqrt{(p1-q1)^2+(p2-q2)^2+\cdots+(pi-qi)^2+\cdots+(pn-qn)^2}$$

By comparing the results obtained, we will choose the smallest value, and we can deduct that the newly introduced product belongs to the category in which the product is at the shortest distance.

To apply the KNN algorithm, we return to the existing dataset, grouped into the three categories "1", "2", "3". We will pick the first 34 records to form an algorithm training set, and the remaining 16 records will be used as a test set to predict belonging to one of the groups. To make the group clearer, we will change the column from numeric type into text, replacing the values: "1" = "medium", "2" = "caloric", "3" = "dietetic".

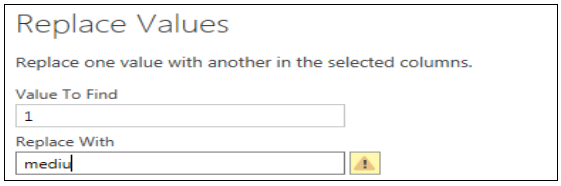


Fig.22. Replacing numerical values with text values

We will normalize the data to bring them to the same scale. For this we will use the R scripting tool, defining a function called "function (x)", which we will apply to the numerical columns in the known dataset, called "dataset":

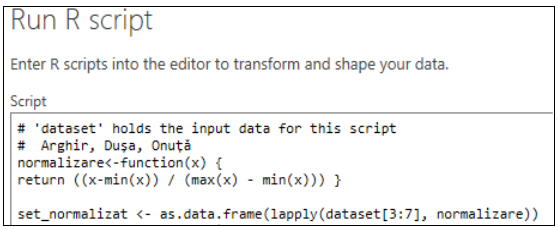


Fig. 23. R Script for dataset normalization

The "lapply" R function has two input variables, *the initial dataset*, and *the normalization function*, building a new normalized set.

The data will be divided into two - in a training set (the first 34 records) and a test set (the next 16 records) on which we will apply the prediction model. We will set prediction labels for the result as being column 8 of the dataset.

```
set_antrenare<-set_normalizat[1:34, ]
set_test<-set_normalizat[35:50, ]
set_antrenare_eticheta <- dataset[1:34, 8]
set_test_eticheta <- dataset[35:50, 8]
```

Fig. 24. R script for obtaining the training set and the test set

Once the data has been prepared, we will apply the KNN algorithm, for which is required "class" library. The R function "KNN" has as input *the train set*, *the test set*, *the prediction labels* (cl) and *the number k* - the closest neighbour for which we calculate the square root, more precisely k will take the value $\sqrt{34}\approx 6$.

```
library("class")
set_test_predictie <- knn(train=set_antrenare, test=set_test,
                          cl=set antrenare_eticheta, k=6)
output<-set_test
output$result<-set_test_predictie
```

Fig. 25. Applying the KNN algorithm based on the training set

We will have four Power BI outputs, namely: -the initial normalized dataset, -a sequence from the normalized set (the first

34 records representing the training data), - another sequence from the normalized set (the last 16 records representing the test set).

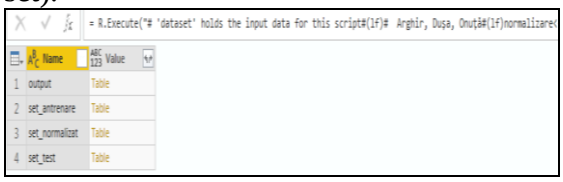
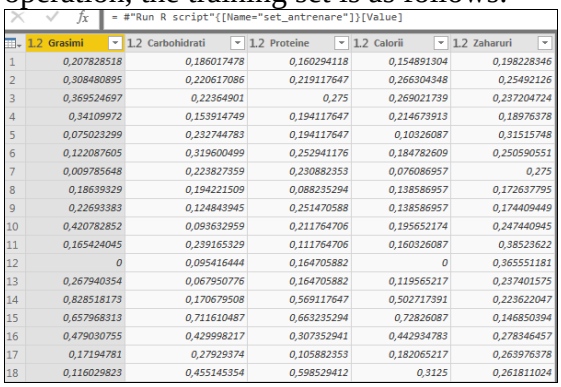


Fig. 26. Power BI Output - normalized set and prediction

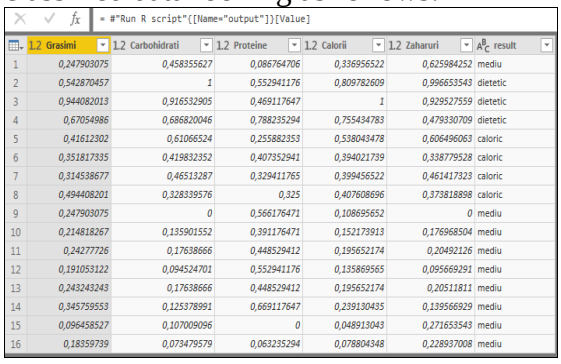
Following the data normalization operation, the training set is as follows:



	Grasimi	Carbohidrati	Proteine	Calorii	Zaharuri
1	0,207828518	0,186017478	0,160294118	0,154891304	0,198228346
2	0,308480895	0,220617086	0,219117647	0,266304348	0,25492126
3	0,369524697	0,22364901	0,275	0,269021739	0,237204724
4	0,34109972	0,153914749	0,194117647	0,214673913	0,18976378
5	0,075023299	0,232744783	0,194117647	0,10326087	0,31515748
6	0,120087605	0,319600499	0,252941176	0,184782609	0,250590551
7	0,009785648	0,223827359	0,230882353	0,076086957	0,275
8	0,18639329	0,194221509	0,088235294	0,138586957	0,172637795
9	0,22693383	0,124843945	0,251470588	0,138586957	0,174409449
10	0,420782852	0,093632959	0,211764706	0,195652174	0,247440945
11	0,165424045	0,239165329	0,111764706	0,160326087	0,38523622
12	0	0,095416444	0,164705882	0	0,365551181
13	0,267940354	0,067950776	0,164705882	0,119565217	0,237401575
14	0,828518173	0,170679508	0,569117647	0,502717391	0,223622047
15	0,657968313	0,711610487	0,663235294	0,72826087	0,146850394
16	0,479030755	0,429998217	0,307352941	0,442934783	0,278346457
17	0,17194781	0,27929374	0,105882353	0,182065217	0,263976378
18	0,116029823	0,455145354	0,598529412	0,3125	0,261811024

Fig. 27. Normalized dataset

Applying the KNN algorithm, we obtain the prediction of belonging of the 16 new products to the three defined groups; the classified data looking as follows:



	Grasimi	Carbohidrati	Proteine	Calorii	Zaharuri	result
1	0,247903075	0,458355627	0,086764706	0,336956522	0,625984252	mediu
2	0,542870457	1	0,552941176	0,809782609	0,996653543	dietetic
3	0,944082013	0,916532905	0,469117647	1	0,929527559	dietetic
4	0,67054986	0,686820046	0,788235294	0,755434783	0,479330709	dietetic
5	0,41612302	0,61066524	0,255882353	0,538043478	0,606496063	caloric
6	0,351817335	0,419832352	0,407352941	0,394021739	0,338779528	caloric
7	0,314538677	0,46513287	0,329411765	0,399456522	0,461417323	caloric
8	0,494408201	0,328339576	0,325	0,407608696	0,373818898	caloric
9	0,247903075	0	0,566176471	0,108695652	0	mediu
10	0,214818267	0,135901552	0,391176471	0,152173913	0,176968504	mediu
11	0,24277726	0,17638666	0,448529412	0,195652174	0,20492126	mediu
12	0,191053122	0,094524701	0,552941176	0,135869565	0,095669291	mediu
13	0,243243243	0,17638666	0,448529412	0,195652174	0,20511811	mediu
14	0,345759553	0,125378991	0,669117647	0,239130435	0,139566929	mediu
15	0,096458527	0,107009096	0	0,048913043	0,271653543	mediu
16	0,18359739	0,073479579	0,063235294	0,078804348	0,228937008	mediu

Fig. 28. The obtaining prediction

In order to verify the accuracy of the prediction model, we will make a matrix representation - confusion matrix. Thus, the first position shows the number of true-positive instances, in which case a "caloric" product is classified correctly as "caloric".

On the main diagonal continues with correct classified instances in which "dietetic" delicacies are predicted as "dietetic" and those "medium" caloric are classified as "medium"; only 1 product was predicted as belonging to the "medium" category, but it belongs to "calorific" category.

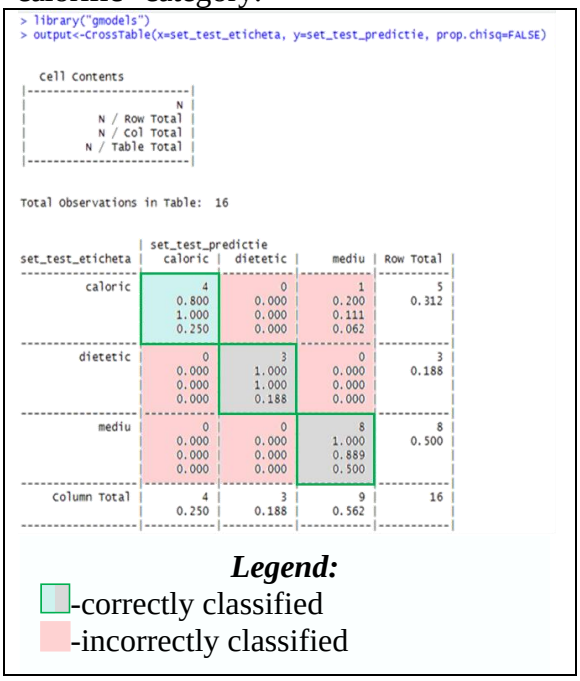


Fig. 29. The accuracy of the prediction model with the KNN algorithm

The quality of a classifier from the perspective of the correct identification of a class is measured using the information in the confusion matrix that contains:

- True Positive (TP) - a product belonging to the class was recognized as belonging to the class;
- True Negative (TN) - a product that does not belong to the class was not recognized as belonging to the class;
- False Positive (FP) - a product that belongs to the class was not recognized as belonging to the class,
- False Negative (FN) - a product that does not belong to the class was recognized to belongs to the class;

Table 5. Confusion matrix

	belongs		does not belong	
recognized	**TP**		**FN**	
	4	4	(0+1)	1
not recognized	**FP**		**TN**	
	(0+0)	0	(3+8)	11

Based on these values, a number of other measures can be calculated:

- **True Positive Rate/** Sensitivity/ Recall/ represent the number of delicacies correctly identified as positive out of total true positives:

$$TPR = \frac{TP}{(TP+FN)} \rightarrow TPR = \frac{4}{4+1} = 80\%$$

- **False Positive Rate** represent the number of delicacies erroneously identified as positive out of total true negatives:

$$FPR = \frac{FP}{(FP+TN)} \rightarrow FPR = \frac{0}{0+11} = 0\%$$

- **False Negative Rate** represent the number of delicacies erroneously identified as negative out of total true positives;

$$FNR = \frac{FN}{(FN+TP)} \rightarrow FNR = \frac{1}{1+4} = 20\%$$

- **True Negative Rate**/ Specificity represent the number of delicacies identified as negative out of total true negatives:

$$TNR = \frac{TN}{(TN+FP)} \rightarrow TNR = \frac{11}{11+0} = 100\%$$

- **Overall Accuracy** is the ratio between the correctly classified delicacies (the values on the main diagonal) and the total number of test values:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \rightarrow ACC = \frac{4+3+8}{16} = \frac{15}{16} = 91,75\%$$

- **Error Rate** is the ratio between the number of incorrectly classified delicacies and the total number of test values:

$$ERR = \frac{FP+FN}{TP+FP+TN+FN} \rightarrow ERR = \frac{1}{16} = 0,062$$

The value of global accuracy that is close to 100% and the error rate close to 0 shows that the prediction model is a very good one.

The result of the prediction compared to the actual data, we represented as a table, with help of two Slicer filters (that do not interact one with other); so we could put two filtering conditions on junction data, in the first Slicer (the real category) by selecting "caloric" and in the second (the predicted category) by selecting "medium" we get the only solution that was not correctly predicted by the model classification, namely "Corn-Ciocolată" (Chocolate Croissant), which was predicted "medium", in reality being in the "caloric" category.

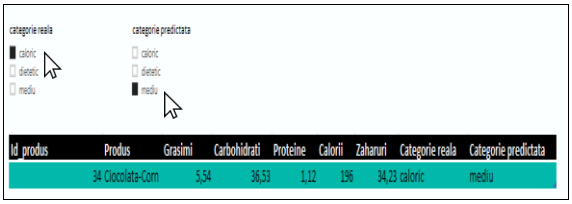


Fig. 30. Selecting an erroneous prediction using Power BI

12. Conclusions and future works

The purpose of applying data mining concepts was to extract the most relevant information and to predict on its basis the belonging of data to certain classes.

Following the application of the "Apriori" algorithm, in were we calculated the support level and the confidence level, we can conclude that in 100% of the cases, customers who bought a specific product - "Tartă-Coacăze" (Blueberries Tart), mandatory bought a second product namely "Apă-Plată" (Non-carbonated Mineral Water).

Instead, only 3% of buyers preferred the combination of "Suc-Lămâie" (Lemon Juice) and "Fursecuri-Zmeură" (Raspberry Cookies). Through cluster analysis, in which we wanted to categorize the same entities in representative categories, were obtained three consumer packages that could be addressed to a variety of buyers: "dietetic", "medium", "caloric".

These packages were obtained by analysing the components: "product name", "calories", "fats", "proteins" and "sugars".

Through the classifier based on the K-nearest-neighbours we wanted to determine in which of the three categories we could introduce a new product. This is very useful for the majority of businesses, as it manages to determine various classifications based on aprioristic knowledge.

In conclusion, we can say that the field of business intelligence is constantly developing, and by transforming raw and unstructured data, can be obtained very useful information.

This process can be accomplished through various data mining algorithms that can take the form of interactive reports that can be presented in a friendly manner, helping the businessman consolidate his future decisions.

In the near future, we want to collect a dataset of major economic interest from Romania, to be subjected to several analyses to offer a wide range of options both in the area of income and profitability and financial analysis for identifying problems and improving these vital indicators.

References

[1] Miller Devens Richard, „Cyclopaedia of Commercial and Business Anecdotes...” Nabu Press, 2011, ISBN: 978-1248003671;
[2] D. J. Power „A Brief History of Decision Support Systems, version 4.0". DSSResources.COM, 2007.
[3] https://selecthub.com/business-intelligence/key-types-business-intelligence-tools/, „What are the Different Types of Business Intelligence Tools?”, Accessed February 28, 2019.
[4] https://www.passionned.com/wp/wp-content/uploads/passionned-parabola-bi-analytics-2019.png?x18199, “References about the most used BI tools”, Accessed February 29, 2019.
[5] http://www.pentalog.ro/html/pentalog/corporate/images/corporate/schema_avantaje_bi, References about BI advantages, Accesed March 2, 2019.
[6] https://archive.ics.uci.edu/ml/index.php, “Seturi de date pentru prelucrări de data mining şi business intelligence”, Accessed March 3, 2019.
[7] http://ip.ase.ro, References about the application of data mining algorithms, Accessed March 7, 2019.
[8] http://ip.ase.ro/Ad11.pdf, References about cluster analysis, Accessed March 9, 2019.
[9] https://www.academia.edu/4424156/, „O îmbunătăţire a performanţelor algoritmului KNN în sistemele de recomandare pe web”, Accessed March 10, 2019.
[10] https://en.wikipedia.org/wiki/IBM_Cognos_Analytics, „IBM Cognos Analytics”, Accessed March 3, 2019.
[11] https://support.sas.com/documentation/onlinedoc/portal/index.html, Accessed March 3, 2019.
[12] https://comparisons.financesonline.com/microsoft-power-bi-vs-sas-business-intelligence, Accessed March 3, 2019.
[13] https://comparisons.financesonline.com/ibm-cognos-vs-oracle-bi, March 3, 2019.
[14] https://www.trustradius.com/compare-products/ibm-cognos-vs-oracle-business-analytics, Accessed March 4, 2019.
[15] https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68, Accessed March 9, 2019.

Denis-Cătălin ARGHIR has graduated the Faculty of Cybernetics, Statistics and Economics Informatics in 2017. He holds a bachelor degree in Economics Informatics and now he is following the master's programme of Database-Support for Business and Psycho-Pedagogical 2nd Module. He followed a series of practice programs at a number of top companies as Microsoft, IBM, High Tech System Software, Cegedim RX, EY Romania and governmental institutions within the Romanian Government to develop his work with databases, business intelligence, internet of things, .NET, Java, and Robotic Process Automation that represent his area of interests.

Ioana-Gilia DUȘA is a student in the final year of the master programme Database-Support for Business at the Faculty of Cybernetics, Statistics and Informatics Economics. She graduated the same faculty, the field of Informatics Economics in 2017 and she has been working as an ETL Developer for almost 3 years. Her interest domains related to computer science are: Data Warehouse, Business Intelligence, informatic systems, and databases.

Miruna ONUȚĂ graduated from The Faculty of Economic Cybernetics, Statistics and Economic Informatics (bachelor domain: Economic Informatics) and currently she is a student in the final year of the master programme: Database-Support for Business at the same university. Since 2018 she works as a business intelligence analyst and she is interested in: BI, databases, data warehouse, cloud computing platforms and services, ETL process.

Internet of Things (IoT)

Diana - Iuliana BOBOC, Ştefania - Corina CEBUC
Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies, Romania
bobocdianaiuliana@gmail.com, cebuc.stefaniacorina@gmail.com

After the World Wide Web (the 1990's) and the mobile Internet (the 2000's), we are now heading to the third phase of the Internet evolution – the Internet of Things. A new era where the real, digital and virtual converge to create smart environments that make energy, transport, cities and many other areas more intelligent. Smart is the new green and the green products and services are being replaced by smart products and services. The Internet of things means billions of smart objects that are incorporated into our everyday life, improving the social, technical and economic benefits. The Internet of Things enables anytime, anyplace connectivity for anything and anyone. However, there are many issues that need to be solved in order to reach the full potential of the Internet of things.

Keywords: *technology, Internet of things, networks, sensors, augmented behaviour, augmented intelligence, standards, interconnectivity, Web, security, privacy, IP, Internet, transfer rate, network protocols, smartphone, interoperability, smart applications.*

1 Introduction

When time is more limited and the volume of information is increasing, the Internet of things emerges and changes radically the lifestyle of people, businesses and society.

Whether it is tangible goods such as consumables, cities, cars, buildings, etc., whether it is intangible goods such as agriculture, health, tourism, energy, etc., all the things in the everyday life of people are expected to be connected to Internet, to have analytical capabilities and communication skills that significantly improve the way people live, work or interact.

This paper outlines what is the Internet of things, the history of IoT, the advantages and disadvantages of IoT, the challenges and possible solutions related to IoT, the technologies driving IoT and some example of IoT applications.

2 Definition and characteristics

Despite the global buzz around the Internet of things (IoT), there is no single, universally accepted definition for the term. However, there are many definitions that describe and promote a particular view of what IoT means such as:

- **IAB (*Internet Arhitecture Board*)** describes IoT as: "*a trend where a large number of embedded devices use communication services provided by Internet protocols. These devices, called "smart objects", are not directly operated by humans, but exist as components in buildings or vehicles or are spread out in the environment."* [1]
- **ITU (International Telecommunication Union)** has formulated the following definition: "*IoT is a global infrastructure for information society that provides advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies.*" [2]
- **IEEE (Institute of Electrical and Electronics Engineers)** definition states that IoT is "*a network of items - each embedded with sensors - which are connected to the Internet.*" [3]
- The **Oxford Dictionary** defines the term Internet of things as: "*the interconnection via the Internet of computing devices embedded into everyday objects, enabling them to send and receive data.*" [4]
- Therefore, we propose the following **definition**: "*IoT is the total number of devices interconnected through the Internet, capable of collecting data in order to monitor and control everyday things, remotely, without the need for continuous interaction between things and people.*"

According to [5], the fundamental **characteristics** of the IoT are as follows:

- **interconnectivity**: anything can be interconnected with the global information and communication infrastructure;
- **things-related services**: the IoT is capable of providing thing-related services within the constraints of things;
- **heterogeneity**: the devices in the IoT are heterogeneous as based on different hardware platforms and networks. They can interact with other devices or service platforms through different networks;
- **dynamic changes**: the state of devices change dynamically, e.g., connected or disconnected as well as the context of devices including location and speed. Moreover, the number of devices can change dynamically;
- **enormous scale**: the number of devices that could communicate with each other will be at least an order of magnitude larger than the number of devices connected now to the Internet.

As specified in [6], the phase "Internet of Things" points out a **vision** of the machines of the future: in the 19th century, the machines were taught to do, in the 20th century, they learned to think, and in the 21st century they learned to perceive - they actually sense and respond to the interaction with the environment.

Based on [7], the main **objective** of IoT is to enable objects to be connected anytime, anywhere, with anything and anyone, using any network and service. Connections multiply and create a completely new dynamic network of networks - the Internet of things. IoT is not science fiction, but is based on solid technological advances and visions of network ubiquity.

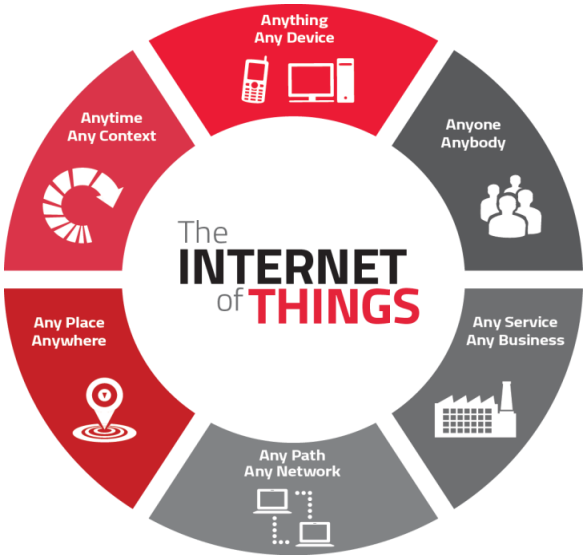


Figure 1 - Main objective of IoT [7]

As maintained in [8], the 3 Cs of IoT are:

- **communication.** IoT communicates information to people and systems such as state of the equipment (e.g., open or closed, connected or disconnected, full or empty) and data from sensors that can monitor the vital signs of a person. Usually, we didn't have access to this information before or it was collected manually and sporadically. For example, in the healthcare industry, IoT can help a hospital to track the location of anything, from wheelchairs to cardiac defibrillators to surgeons.
- **control and automation**. In a connected world, an enterprise or a consumer has the ability to remotely control a device. For example, an enterprise can remotely start or shut down a particular piece of equipment or adjust the temperature in a climate-controlled environment, while a consumer can use IoT to start the washing machine or to unlock the car.
- **cost savings**. With new sensor information, IoT can help an enterprise save money by minimizing equipment failure and enabling the enterprise to perform planned maintenance. Also, sensors have measurement capabilities such as driving behavior and speed in order to reduce fuel expense, wear and tear on consumables.

3 Evolution

3.1 IoT today and tomorrow

The Internet of Things may be a new topic in the IT industry, but it's not a new concept.

As evidenced in [8], in **1999**, **Kevin Ashton** set out guidelines for what has now become IoT at a MIT AutoID lab. Ashton was one of the pioneers who described this concept as he searched for ways that Procter & Gamble could improve business by linking physical objects to the Internet via RFID sensors. The result of his research was as follows: if all objects in everyday life would be equipped with identifiers and wireless connectivity, these objects could communicate with each other and be managed by computers.

In a 1999 article, Ashton wrote: "*If we had computers that knew everything about things - using the data they collected without any help from people - we could greatly reduce waste, loss and cost. We would know when things need to be replaced, repaired or if they are fresh or have expired. We need to develop computers with their own means of collecting information so that they can see, hear and smell the environment for themselves, without being limited to data entered by human.*" [8]

According to CISCO IBSG (Internet Business Solutions Group), IoT is the point in time when more things than people were connected to the Internet. In 2003, there were about 6.3 billion people living on the planet, 500 million devices connected to the Internet and less than one device per person. IoT did not yet exist in 2003, because the number of connected things was relatively small, given that smartphones were just being introduced. [9]

An explosive increase of smartphones and tablets has led to an increase in the number of devices connected to Internet to 12.5 billion in 2010, while the world's population has increased to 6.8 billion, making the number of devices connected per person more than 1 for the first time in history. [9]

Looking to the future, CISCO IBSG predicted that there will be 25 billion devices connected to the Internet in 2015 and 50 billion devices connected to the Internet in 2020. These estimates do not take into account the rapid technological advances and are based on the world's population, of which 60% is not yet connected to the Internet. [9]

2015 was the year when IoT gained notoriety. Companies adopt the following strategy: "*start small think big*" and invest heavily in development of IoT applications. [10]

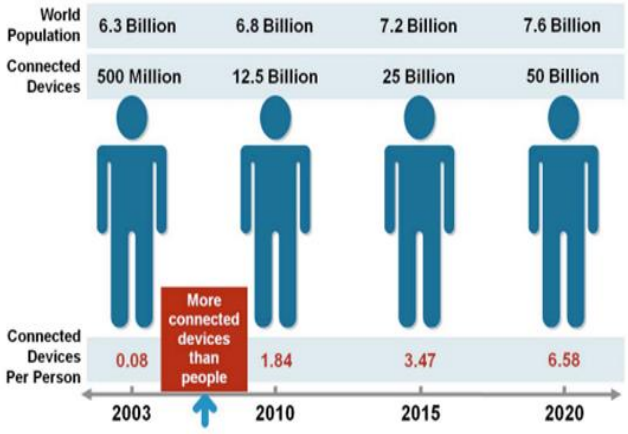


Figure 2 - Number of devices connected to the Internet [9]

In 2015, there was an explosion of new Internet of Things related job titles such as: IoT value creator, IoT chief architect, IoT technical sales engineer. This trend shows that in 2015 companies have created new IoT departments and this trend is expected to continue in the future. [10]

IoT has become part of the everyday life of people and the phase "Internet of Things" evolves continuously in content, applicability, vision and technology.

3.2 IoT: First evolution of the Internet

On the one hand, the Internet is the physical layer or a network of switches, routers and other transmission equipment, designed to transport information from one computer to another quickly, reliably and securely. On the other hand, the web is the application layer, with the role of providing an interface so that information on the Internet can be used. [9]

According to [9], the web has gone through several evolutionary stages:

- stage 1. During this stage, web was used by the academia for research.
- stage 2. This stage focused on the need of companies to share information on the Internet so that people could learn more about the products and services they offer.
- stage 3. During this stage, web have moved from static data to transactional data, where products and services could be bought and sold online. Also, companies such as eBay and Amazon.com have become very popular among Internet users.

- stage 4. During this stage, people could communicate, share information about themselves with their friends, colleagues, family. Moreover, companies such as Facebook and Twitter have become extremely popular and profitable.

Compared to the web, the Internet has been steadily developing and improving, but it has not changed much. It essentially does the same thing that it was designed to do during the ARPANET era. For example, initially there were several communication protocols, but today the Internet only uses TCP / IP. [9]

In this context, IoT has become extremely important because it is the first real evolution of the Internet, a stage that will lead to the development of applications that could dramatically improve the way people live, learn, work and entertain. IoT has already created the Internet based on sensors, which allows people to be more proactive and less reactive. [9]

4 Advantages and disadvantages

As highlighted in [11], there are many **advantages** of using IoT, which can help individuals, businesses and society on a daily basis.

For individuals, IoT can positively influence various aspects of their existence such as health, safety, financially, daily activities. IoT can also act as a tool for people to save money and time. If their home appliances are able to communicate, they can operate in an energy efficient way.

Also, the advantages of IoT spread to companies, where IoT becomes useful in various activities such as monitoring inventory, location, employees, etc. Physical devices are able to communicate with people, letting them know their condition, location, etc.

Another advantage of IoT is the ability to track the behavior of individual consumers and to identify key customers and their preferences based on information provided by devices.

The IoT has many advantages to businesses, consumers, the environment, but it also has a number of disadvantages. According to [11] and [12], among the **disadvantages** of IoT are:

- **compatibility**. Now, there is no standard for tagging and monitoring all types of sensors.
- **privacy / security**. Although there are security measures that are taken to protect personal information, there is always a possibility of hackers breaking into the system and stealing the data. Hence, all these safety risks become the consumer's responsibility.
- **complexity / over-reliance on technology**. The current generation has grown with the readily availability of the Internet and technology. However, relying on technology, making decisions based on the information it provides could lead to devastating effects. We are the users of complex software systems and we must not forget that no system is robust and fault-free. The more we depend on the Internet and technology, the more it increases the possibility of a catastrophic event if it crashes.
- **connecting an increasing number of devices to the Internet** will lead to job losses. The automation of IoT will have a devastating impact on the employment prospects of the less educated population.

5 Technologies

As described in [13], the main way to capture the processes in Mark Weiser's model is as an **Information Value Loop** with distinct but connected stages.

Through sensors, data is generated about a particular action in the physical environment, which is then communicated over the Internet and aggregated across time and space and eventually analyzed in order to be beneficial for changing future acts.

Getting information allows an organization to create value; how much value is created depends on "value drivers", which define the characteristics of the information along the loop. These factors include:

- **magnitude** - the factor that determines the amount of information that informs action;
- **risk** - the factor that determines the probability that information will create value in the manner expected;

- **time** - the factor that determines how quickly value can be created from the information. [14]

Figure 3 - The Information Value Loop [13]

For example, a sales manager wants to be able to influence customer decisions, so he needs to know what customers want "now and here". This can require information with higher frequency, accuracy and timeliness so that the manager can influence customer action in real time by offering complementary products or incentives. [14]

5.1 Sensors

Below we present the first stage in the value loop, create, as indicated in [13].
IEEE provides the following definition of the term sensor: "*A sensor is an electronic device that produces electrical, optical or digital data derived from a physical event. Then the data produced from the sensors is transformed into information that is useful in making decisions done by other intelligent devices or individuals.*"
Different sensors capture different types of information. For example, the accelerometer detects if an object moves and in which direction, a gyroscope measures more complex, multi-dimensional motions by tracking the position and rotation of an object.
Now, sensors can measure everything: motion, power, pressure, position, light, temperature, humidity, radiation and can be incorporated into everything: power sources, vehicles, smartphones, wearables, houses, etc.
There has been a rapid increase in the use of smaller sensors that can be embedded into smartphones and wearables. The average number of sensors on a smartphone increased from three (including accelerometer, proximity, ambient light) in 2007 to at least ten (including fingerprints and gesture sensors) in 2014.
The price of sensors has declined significantly over the last several years and these price declines are expected to continue in the future. The average cost of an accelerometer in the US was about 40 cents in 2016 compared to $ 2 in 2006.

5.2 Networks – RFC 7452

Below we present the second stage in the value loop, communicate, as identified in [13].
The information created by the sensors is not useful at the time and the place of creation. The signals from sensors are communicated to other devices for analysis. This involves sending data over a network. Sensors are connected to networks through network devices such as hubs, gateways, routers, bridges and switches. The first step in the process of passing data from one machine to another via a network is to uniquely identify each machine.
Network protocols are a set of rules that define how computers identify each other. There are two main categories of network protocols:

- **proprietary protocols** that ensure identification and authorization of machines with specific hardware and software components, allowing manufacturers to differentiate their offerings.
- **open protocols** that ensure interoperability across heterogeneous devices and improve scalability.

IP (Internet Protocol) is an open protocol that provides unique addresses to devices connected to the Internet.
Currently, there are two versions of the protocol: **IP version 4** (IPv4) and **IP version 6** (IPv6). IPv6 is characterized by the increasing the number of unique addresses from 232 addresses provided by IPv4 to 2128 addresses provided by IPv6. Network technologies can be wired or wireless. With continuous user and device growth, wireless

networks are useful for continuous connectivity, while wired networks are useful for transporting a large amount of reliable and secure data.

Choosing a network technology depends on the geographical range to be covered. When data is transferred over small distances (e.g., inside a room), devices can use wireless personal area network technologies (Bluetooth, ZigBee) or wired technology (USB). When data is being transferred over relatively large distances (e.g., inside a building), devices can use wired local area technologies (Ethernet, fiber optics) or wireless technology (Wi-Fi). When data is transferred over longer distances (e.g., beyond buildings and cities), devices can use WAN technologies (WiMAX, weightless, mobile technologies: 2G, 3G, 4G). Technologies such as 4G (LTE, LTE-A) and 5G are favorable for the development of IoT applications, given their high data transfer rates. Technologies such as Bluetooth Low Energy and Low Power Wi-Fi are suited for the development of IoT applications, taking into account the low energy consumption.

In the last 30 years, the data transfer rates have increased from 2 Kbps to 1 Gbps. The transition from the first to the second generation of phones changed the way messages were sent, from analog signals to digital signals. The transition from the second to the third generation of phones offered users the possibility of sharing multimedia files over high speed connections.

In 2003, the price to transfer 1 Mbps in the US was $120, while in 2015 the price dropped to 63 cents. Also, in 2018, 50% of all mobile and fixed device connections were IPv6-based, compared with 16% in 2003.

5.3 Standards

As mentioned in [13], the third stage in the value loop, aggregate, refers to data manipulation, processing and storage activities. Aggregation is achieved through the use of different standards depending on the IoT application.

International Organization for Standardization (ISO) defines a standard as follows: "*A standard is a document that provides specifications, guidelines or features that can be used consistently to ensure that products and services are suitable for their purpose.*"

There are two types of standards for the aggregation process:

- **regulatory standards** refer to recommendations on data security and privacy of data.
- **technical standards** include: network protocols, communication protocols and data aggregation standards. Network protocols define how machines identify each other, while communication protocols provide a set of rules or a common language for devices to communicate. Once the devices are connected to the network, identify each other and exchange data with each other, aggregation standards help aggregate and process data so that those data become useful.

Now, there are many efforts to develop standards that can be adopted widely. There are two types of developments: vendors (from the IoT value chain) setting standards together and standardization bodies (e.g., IEEE) collaborating to develop a standard that vendors will follow. But, it's difficult to create a single universal standard or a rule that dominates all other rules, either at the network level or at the data aggregation level.

As far as network protocols are concerned, some important players in the IT industry have had the opportunity to develop standards that IoT developers will follow in the next years. For example, Qualcomm, together with companies such as Sony, Bosch, Cisco has developed the AllSeen Alliance that offers AllJoyn platform. Also, Intel launched the open-source IoTivity platform.

5.4 Augmented intelligence

As claimed in [13], the fourth stage in the value loop, analyze, is determined by cognitive technologies and associated models that facilitate the use of cognitive technologies. These are known as augmented intelligence to highlight the idea that systems can automate, complete and improve intelligence in a way that excludes people.

The analytics stage involves a thorough search through a large quantity of confusing and conflicting data to get meaningful information that helps take better decisions. There are three

different ways in which the analytics can inform action:

- **descriptive analytics** tools answer the question "What has happened?" and augment intelligence so that it can work effectively with complex datasets that are normally hard to access and use.
- **predictive analytics** tools answer the question "What could happen?" and exploit a large and growing amount of data to build useful models that can correlate seemingly unrelated variables.
- **prescriptive analytics tools** answer the question "What should one do for a desired result?" and include optimization techniques that are based on large data sets, business rules and mathematical models.

Cognitive technologies that have become increasingly used for predictive and prescriptive analytics are:

- **computer vision**. It refers to the ability of computers to identify objects, scenes and activities in images.
- **natural language processing**. It refers to the ability of computers to process text in the same way human do, extracting the real meaning of the text or even generating text that is easy to read.
- **speech recognition**. It refers to accurate transcription of human speech.

5.5 Augmented behaviour

According to [13], the concept of augmented behaviour refers to doing a certain action, which is the result of all previous stages of the value loop, from the creation to analysis of data.

The augmented behavior, the last stage in the value loop, restarts the loop because an action leads to the creation of other data when it is configured to do so.

The difference between augmented intelligence and augmented behaviour is: augmented intelligence refers to informed actions, while augmented behavior refers to an observable action in real world.

Augmented behaviour finds information in at least three ways:

- **machine-to-machine interfaces** (M2M). These refer to the set of technologies that enable machines to communicate with each other and drive action.
- **machine-to-humans interfaces** (M2H). Based on collected data and computational algorithms, machines have the potential to convey meaningful actions to people who then decide whether or not to take the recommended action.
- **organizational entities**. Organizations include people and machines and therefore the benefits and challenges of both M2M and M2H interfaces.

Augmented behaviour is influenced by a number of factors such as:

- **low machine prices.** Lower prices of underlying technologies in IoT such as sensors, network connections, data processing tools, cloud based storage leads to lower prices of robots.
- **improved machine functionality**. Typically, robots take decisions based on programmed algorithms, regardless of situation and information availability. However, recent advances in robotics offer machines the possibility to request more information if there is insufficient information to take a decision.

6 Challenges and potential solutions

Below we present some challenges and potential solutions for IoT technology as stated in [13].

With regard to sensors, there are three main factors driving the deployment of sensor technology: price, capability and size. Even if sensors are sufficiently small, smart and inexpensive, there are some challenges such as:

- **power consumption**. The sensors are powered either through inline connections or by batteries. Inline connections are constant, but can be impractical or expensive. Batteries can be a convenient alternative, but battery life, charging and replacement, especially in remote areas, can represent significant issues.
- **interoperability**. Most of the sensor systems in operation are proprietary and are designed for specific applications. Specific communication protocols are required to facilitate communication between heterogeneous sensor systems, namely lightweight communication protocols (e.g., Constrained Application Protocol (CoAP)).

- **data security**. The use of encryption algorithms ensures data protection, though low memory capacity, power consumption concerns and sensors relatively low processing power could limit data security.

With reference to networks, there are two main factors driving the deployment of networks technology: cost and speed. Even if network technologies have improved in terms of higher data transfer rates and lower costs, there are some challenges such as:

- **interconnections**. The value of a network is proportional to the square of the number of compatible communication devices. Now there is limited value in connecting devices to the Internet. To solve this challenge, companies should connect all devices to the network and to each other.
- **penetration**. There is a limited penetration of networks through high-bandwidth technologies such as LTE and LTE-A, while 5G technology has just emerged in the information technology market. Currently, LTE-A holds 5% of total mobile connections worldwide and LTE holds 2% of the world's total mobile connections, given the investments of network providers in 3G technology over the last 3-5 years.
- **security**. With an increase in the number of sensors connected to the network, there is a need for effective solutions for authentication and access control. The Internet Protocol Security (IPsec) provides a favorable level of secured IP connection between devices.
- **power consumption**. Devices connected to network consume power and require continuous power supplies. Using protocols such as power-aware routing protocols and sleep-scheduling protocols it can improve network management. Power-aware routing protocols determine their routing decision on the most energy-efficient route for transmitting data packets, while sleep-scheduling protocols define how devices can be inactive for better energy efficiency.

For aggregation of IoT data, there is a need for solving the following challenges:

- **standard for handling unstructured data**. Structured data is stored in relational databases and queried through SQL. Unstructured data is stored in different types of NoSQL databases without a query standard. Therefore, new databases created from unstructured data can not be manipulated and used by the old database management systems that companies typically use.
- **regulatory standards for data markets**. Data brokers are companies that sell data collected from different sources. Since most data is sold offline, it is necessary to apply regulatory standards for transactions between providers and users.
- **technical abilities to use newer aggregation tools**. Although there is an upward trend in the number of people trained to use newer tools like Spark and MapReduce, this is far fewer than the number of people trained to use traditional programming languages, such as SQL.

The challenges of augmented intelligence result from data quality, human incapacity to develop a foolproof model and limited capability of old systems to manage unstructured data in real time. Even if the data and model are shipshape, there are some challenges such as:

- **inaccurate analysis due to data or model flaws**. Lack of data or the presence of outliers can lead to false positives or false negatives, thus exposing different algorithmic limitations. The algorithm might make incorrect conclusions if all decision rules are not set correctly.
- **the ability of new systems to analyze unstructured data**. Now, most analytics systems allow the management of structured data, but most IoT interactions generate unstructured data.
- **the ability of old systems to manage real-time data**. Currently, traditional analytics systems use batch-oriented processing, but IoT requires data to be processed in real time in order to get meaningful conclusions.

For augmented behavior, there are some challenges related to machines' actions in unpredictable situations and the security of information that brings to mind such judgments. Interoperability is an additional issue that occurs when heterogeneous machines have to work together in a M2M configuration. Beyond machine behavior issues, managing human behavior in M2H

interfaces and organizational entities presents their own challenges.

7 Smart applications

The evolution of applications, their economic potential and their impact in addressing the societal trends for the years after the emergence of IoT have changed dramatically. Societal trends have created significant opportunities in various areas such as: health and wellness, transport and mobility, security and safety, energy and environment, communication and e-society. [5]

Potential IoT applications are numerous and diverse, covering practically all areas of the everyday life of individuals, businesses and society.

Below we will present the main objectives and uses of IoT applications in different domains as reported in [5].

7.1 Smart Health

The main objective of smart health systems is to improve the quality of life for people in need of permanent help, to decrease barriers for monitoring important health parameters, to reduce health costs and to provide right medical support at the right time.

The uses of smart medical systems in the everyday life of people are:

- assistance for elderly people suffering from diseases such as dementia, memory loss, Alzheimer's or for people with disabilities living alone by using sensors to monitor home movements or to send notifications of the times at which certain medicines should be taken;
- monitoring vital signs for athletes in high performance centers by using sensors to measure physical exercise, walking / running steps, sleep, weight, blood pressure, etc.;
- remote monitoring of patients with chronic diseases such as pulmonary or cardiovascular disease, diabetes to obtain reduced medical center admissions or shorter hospital stays;
- sleep control by using sensors to measure small movements such as heart rate, breathing and large movements caused by tossing and twisting during sleep and to record data through the smartphone application;
- toothbrush connected with smartphone application to analyze the brushing use and habits and to display statistics to the dentist.

7.2 Smart city

A smart city is defined as a city that can monitor and create favorable conditions for all its critical infrastructures, which can better optimize its resources, plan its preventive maintenance activities, ensure security while maximizing its services to citizens.

Smart cities can change the everyday life of people as follows:

- real-time monitoring of available parking spaces in the city to identify and book the closest spaces;
- real-time monitoring of noise in crowded and central areas;
- monitoring the vehicles and pedestrians to optimize the driving and walking routes;
- intelligent and adaptive lighting according to time;
- apps on smartphone that supports QR codes to provide interesting and useful information on city sights such as museums, art galleries, libraries, monuments, shops, buses, taxis, parks.

7.3 Smart living

The smart living affects the everyday life of people as follows:

- monitoring energy, water and gas consumption to get recommendations on how to save money and resources;
- turning on and off remotely appliances to avoid accidents and save energy;
- LCD refrigerators that provide information on the products inside them, products that are about to expire, products that should be purchased or washing machines that allow monitoring of the laundry remotely and all this information is accessed via a smartphone application;
- home video surveillance and alarm systems to increase the security of people at home;
- increase personal safety by wearing jewelry that incorporates Bluetooth enabled technology and interacts with a smartphone application that will send alarms to selected

people in the social circle with information that you need help and current location.

7.4 Smart environment

The smart environment can be composed of:

- monitoring the combustion gases and fire conditions to identify alert areas;
- monitoring soil moisture, earth vibrations to detect possible landslides, avalanches, earthquakes;
- locating and tracking wild animals via GPS coordinates;
- study of weather conditions to forecast the formation of ice, rain, drought, snow, strong wind;
- use of sensors integrated in airplanes, satellites, ships etc. to control the maritime and air activities in certain areas.

7.5 Smart energy

The smart energy is characterized by:

- monitoring and optimizing the performance in solar energy plants;
- monitoring and analyzing the flow of energy from wind turbines;
- measuring the water pressure in water transportation systems;
- measuring the level of radiation in nuclear power stations to generate alerts;
- controllers for power supplies to determine the required energy and improve energy efficiency with less energy waste for power supplies of computers and electronics applications of consumers.

8 Conclusions

The Internet of Things changes everything: from users, organizations to today's society. The benefits and opportunities of IoT are endless. Everything is connected with wire or wireless through the Internet.

The Internet of Things offers some interesting applications to make people's lives easier, spanning numerous application domains: health, transport, agriculture, building etc.

However, the Internet of Things presents a number of security and privacy issues because smart devices collect personal data of users regardless of their source or confidentiality. Smart devices know when we are at home, what electronic products and appliances we use, the type of data transmission used and other data that will be retained in the databases of delivery, installation or maintenance service providers. The risk of loss of information in the hands of malicious persons is extremely high.

References

[1] International Architecture Board, *Architectural Considerations in Smart Object Networking*, 2015. Available online: https://tools.ietf.org/html/rfc7452 Accessed: May 2019;

[2] Marco Zennaro, *Introduction to Internet of Things*, 2017. Available online: https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/SiteAssets/Pages/Events/2017/Nov_IOT/NBTC%E2%80%93ITU-IoT/Session%201%20IntroIoTMZ-new%20template.pdf Accessed: May 2019;

[3] Institute of Electrical and Electronics Engineers, *Towards a definition of the Internet of Things (IoT)*, 2015. Available online: https://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_Definition_Internet_of_Things_Revision1_27MAY15.pdf. Accesed: May 2019;

[4] Oxford Dictionary, *Definition of Internet of things in English*, 2013. Available online: https://en.oxforddictionaries.com/definition/internet_of_things. Accesed: May 2019;

[5] Ovidiu Vermesan, Peter Friess, *Internet of Things - From Research and Innovation to Market Deployment*, 2014. Available online: http://www.internet-of-things-research.eu/pdf/IERC_Cluster_Book_2014_Ch.3_SRIA_WEB.pdf. Accessed: May 2019;

[6] CERP-IoT, *Vision and Challenges for Realising the Internet of Things*, 2010. Available online: http://www.internet-of-things-research.eu/pdf/IoT_Clusterbook_March_2010.pdf Accessed: May 2019;

[7] International Telecommunication Union, *The Internet of Things*, 2005. Available online: https://www.itu.int/osg/spu/publications/internetofthings/InternetofThings_summary.pdf Accessed: May 2019;

[8] Lopez Research LLC, *An Introduction to the Internet of Things (IoT)*, 2013. Available online:https://www.cisco.com/c/dam/en_us/solutions/trends/iot/introduction_to_IoT_november.pdf. Accessed: May 2019;
[9] Dave Evans, *How the Next Evolution of the Internet Is Changing Everything*, 2011. Available online: https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf Accessed: May 2019;
[10] Knud Lasse Lueth, I*oT 2015 in review: The 10 most relevant news of the year*, 2015. Available online: https://iot-analytics.com/iot-2015-in-review/ Accessed: May 2019;
[11] Cortland Suny, *The Internet of Things*, 2012. Available online: https://sites.google.com/a/cortland.edu/the-internet-of-things/home. Accessed: May 2019;
[12] Prateek Saxena, *The advantages and disadvantages of Internet of Things*, 2016. Available online: https://e27.co/advantages-disadvantages-internet-things-20160615/. Accessed: May 2019;
[13] Monika Mahto, Jonathan Holdowsky, Michael E. Raynor, Mark Cotteleer, *Inside the Internet of Things (IoT). A primer on the technologies building the IoT*, 2016. Available online:https://dupress.deloitte.com/dup-us-en/focus/internet-of-things/iot-primer-iot-technologies-applications.html?icid=interactive:not:aug15 Accessed: May 2019;
[14] Michael E. Raynor, Mark Cotteleer, *The more things change: Value creation, value capture and the Internet of Things*, 2016. Available online: https://dupress.deloitte.com/dup-us-en/deloitte-review/issue-17/value-creation-value-capture-internet-of-things.html. Accessed: May 2019

Diana - Iuliana BOBOC is a master student in Databases - Support for Business program at the Faculty of Economic Cybernetics, Statistics and Informatics from the Bucharest University of Economic Studies. She is a C# Programmer and has a background in database programming. Her interests include: C# Development, SQL Development, databases. Her hobbies are: travelling, movies, visiting, learning new things.

Ştefania - Corina CEBUC is a master student in Databases - Support for Business program at the Faculty of Economic Cybernetics, Statistics and Informatics from the Bucharest University of Economic Studies. Her interests include: databases, Android development, project management, Agile methodologies, Internet of things.