Data Mining in Client Oriented Businesses

Cveta Martinovska Bande[*], Natasha Stojkovic[*], Elena Kosturanova[*], Mimoza Klekovska[**]

[*] Computer Science Faculty, University Goce Delcev, North Macedonia
[**] American University of Europe, North Macedonia

DOI: 10.29322/IJSRP.11.12.2021.p12054
http://dx.doi.org/10.29322/IJSRP.11.12.2021.p12054

Abstract- This paper presents methodologies for creating profiles of typical customers for certain types of services and products in user-oriented businesses. Because of the specific characteristics of each of the domains where these methodologies would be applied creating one solution for all purposes does not solve the problem. The emphasis of this paper is on experimenting with real data from telecommunication sector and on the construction of clustering system for user profiling. The approach for solving this problem involves analysing the characteristics of the observed data and adjusting the process of intelligent data analysis according to its specifications.

Index Terms- Data mining, User segmentation and profiling, Clustering, Evaluation measures for clustering.

I. Introduction

The development of information and communication technologies has created conditions for production and storage of huge amount of data: computers in stores write a record for each purchased product, telecommunication operators have information about the time, duration, and called number for each call, banks through credit cards have data about financial habits, amounts and locations of

The development of information and communication technologies has created conditions for production and storage of huge amount of data: computers in stores write a record for each purchased product, telecommunication operators have information about the time, duration, and called number for each call, banks through credit cards have data about financial habits, amounts and locations of different purchases, web servers in logs keep data about visits of various web pages, and thus provide information about the habits and interests of the visitors.

Hence, many decisions, choices and habits were recorded in different databases.

With the growing amount of data capabilities for their tracing and understanding are reduced. A vast number of databases created a need for powerful analytical tools that can convert stored data into useful information. Traditional techniques for data analysis generate quantitative and statistical results about the data. But these techniques cannot produce a qualitative description of the rules and classes that are not explicitly expressed in the data, nor can derive assumptions about the reasons for grouping similar samples into categories. To perform these tasks the system for data analysis should include analytical procedures from the machine learning field.

Tools for predictive analytics [1] [2] are an indispensable part of modern systems that support the process of business decision making and thus contribute to creation of qualitative, structured knowledge about various aspects of business processes.

Methods and technics for online analytical processing, data mining [3] [4] [6] [7], process mining, business performance management and predictive analytics, known under the common name Business Intelligence, nowadays are widely used [5].

As a process of extracting implicit regularities by analysing large amounts of data, data mining is used to formulate rules for the features of the observed objects that have statistical confirmation in the analyzed data. The major problem in this process is the fact that the number of eventual regularities is tremendous hence restricting their validation. Hence machine learning algorithms use different assumptions and heuristics that imply a trade-off among the number of extracted rules v.s. speed, readability and simplicity of the rules.

Data mining techniques are used in many areas. In banking and insurance these techniques are used to develop predictive models and risk assessment models for the financial institutions, such as credit rating analysis of customers or risk analysis of insured.

Then, in investments these techniques are used to predict the movement in exchange rates, anticipating the changes in the price of stocks and other securities and analysis of changes in market segments. They can also be applied for detecting fraud and predicting typical use of credit cards in shops, detecting unauthorized intrusion into telecommunications and computer networks.

The techniques of data mining are used in marketing campaigns, for directing the activities in direct marketing based on the analysis of the previous sale and analysis of all interactions with the customer. Also, they are used for analysis of web logs, analysis of navigation of web page visitors, segmenting the visitors of the web pages, automatically creating a personalized content according to

the visitor profile, finding and collecting relevant information on the web according to given criteria.

Although there are a number of programming tools present on the market as a support for the Business Intelligence methods, very few of the companies in our country use them. These programming tools cannot be applied without prior analysis and adjustment of the data. Moreover, it is necessary to choose appropriate algorithms according to the specifics of the problem domain.

Several papers discuss different clustering techniques for user segmentation and profiling [8] [9] [10]. The idea of this work is to investigate data mining techniques, primarily clustering algorithms, that would support the management team's decision making and would contribute to increasing the efficiency and profitability of the companies. Particularly, data mining techniques are used for customer relationship management, customer segmentation and profiling by analysing the habits and understanding and predicting the customers behaviour.

II. USER SEGMENTATION AND PROFILING

In today's competitive business environment, the economy is focused on the user and seeks to increase the level of service that is offered. The long-term strategy of competitive advantage cannot be built at the level of the product but by values that are difficult to replicate and the most important advantage of this type is knowledge. In user-oriented economy the knowledge about the clients that use the services represents a basic condition for gaining and maintaining market share.

Data mining comprises techniques for finding regularities in the behaviour of clients, enables better understanding of their motives and prediction of the future behaviour.

Segmentation and profiling of customers are used to solve problems that are important for marketing and sales departments, such as customer retention, identification of frauds, trend recognition in the customer behaviour, cross selling and up selling. Segmentation or clustering separates users into homogeneous groups based on common attributes (habits, tastes, etc.).

While profiling is building a customer description according to certain attributes such as age, gender, income and lifestyle. In user profiling following characteristics have to be considered: geographical background, cultural and ethnic differences, economic conditions, income, affinities and value system, living standard.

III. MATERIALS AND METHODS

A. Data Collection and Preparation

For each call made in the telecommunications network, some information is stored as a call details record. The number of such records is huge in the databases of telecommunication operators and contains enough information to describe the important characteristics of each call. The minimum information that each entry contains are the details of incoming and outgoing calls, the date and time of the call, and the duration of the call. Call details cannot be used directly for data mining, since the purpose of data applications is to gain knowledge at the consumer level, not at the level of the individual call [11] [12].

The data used in this paper are taken from the telecommunication operator in North Macedonia. The database contains 4000 entries and 22 attributes. Twelve of the attributes represent call details, while the others are attributes contain information about the customer, as for example gender, education, payment method, age, marital status, month of contract, etc.

Data preparation process is performed in several steps: selection of data, cleaning and formatting. Complete and correct examples are selected for construction of the user model. The number of attributes may be reduced based on their information value, or based on their weak predictability, or based on a high degree of correlation with a more powerful attribute.

There are various techniques to reduce the number of attributes starting from joining more attributes with particular linear or nonlinear transformation, to statistical techniques such as the evaluation of the correlation attributes. By purifying the examples data quality is improved. Normalization of data is commonly used as purification technique, and rejection and compensation of attributes are used as a treatment for undetermined values.

The selection of attributes used in our experiments is based on previous examination and the attributes suggested in the literature [13] [14] [15] [16]. A list of attributes that can be used as a customer description based on the calls they make and receive for observed time period comprises of: average call duration, average number of calls received per day, average number of calls originated per day, percent of daytime calls (9am - 6pm), percent of weekday calls (Monday - Friday), percent of calls to mobile phones, average number of SMS received per day, average number of SMS originated per day, percent of international calls, percent of outgoing calls within the same operator, number of unique area codes called during observed period of time and number of different numbers called during observed period of time.

B. User Profiling

User profiling is performed after the selection and purification of data and reduction of the number of attributes and examples. At this phase, methods of machine learning are used to analyse data and to find hidden regularities. The phase of data modelling covers a selection of technique for modelling, defining the test examples and construction of the model. The properties of the problem should be considered from different perspectives, depending on

whether it is classification or association problem, descriptive or predictive type of problem and what is the level of uncertainty, for better modelling of its regularities.
We experimented with several clustering algorithms, such as k-means, agglomerative clustering [17] and BIRCH algorithm. Unsupervised learning algorithms as segmentation techniques, reveal a global data structure or make groups of samples based on similarity.

C. Validation

Evaluation phase consists of an interpretation and an assessment of the model. The choice of the evaluation measure for the classification models depends on the characteristics of the observed problem and the way of its application. Standard measures for evaluation of the model are accuracy, sensitivity, specificity, responsiveness, precision and ROC analysis. Typical evaluation methods for assessing the classification models are the evaluation based on testing set, the cross-validation method and "leave one out" method. Evaluation of the clustering algorithms is not as simple as finding the number of errors or the accuracy of the classification algorithm.
Two main approaches can be distinguished to determine the exact number of clusters in the data:

- Start with a sufficiently large number of clusters and successively reduce this number by combining clusters that have the same properties.
- Cluster the data in different number of clusters and verify the accuracy of the obtained clusters with validation measures.

For implementation of the second approach, various validation measures [9] are proposed, but none of them is perfect by itself. Therefore, in this paper, several measures are considered: Silhouette Coefficient (SC), Calinski-Harabasz Index (CHI) and Davies-Bouldin Index (DBI).

D. Determining the Optimal Number of Clusters

The disadvantage of the proposed clustering algorithms is that the number of clusters has to be assigned in advance. The optimal number of clusters should be obtained using the methods given in the previous section. We used the method called Elbow criterion to find the optimal number of clusters. The Elbow criterion says that one should choose such a number of clusters that adding another cluster does not give better model of the data. Particularly by presenting a graph of the percentage of variance explained by the clusters (ratio of the between group variance to the total variance) versus the number of clusters, the first clusters will add a lot of information (explain the variance), but at some point the marginal gain will decrease, making an angle in the graph (elbow).

IV. EXPERIMENTAL RESULTS

This section presents the results for the validation methods that are used to determine the number of clusters. Following figures are obtained using Python and Scikit-learn machine learning library.

A. Results of K-means Clustering

Figure 1 shows the values for the CHI for the k-means clustering algorithm, using the Elbow criterion. The potential numbers of clusters are those values of k for which the angles are formed. There are elbows for the following values of k: k = 4, k = 6, k = 8 and k = 10.

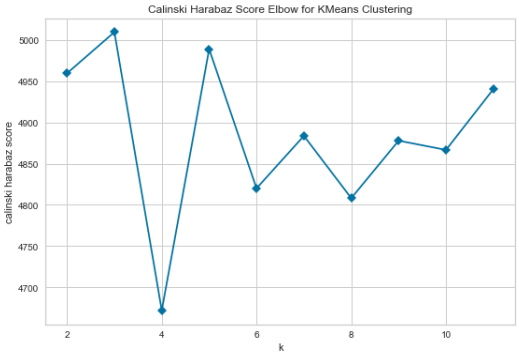

Figure 1. Calinski-Harabasz index for k means clustering using Elbow criterion

Figure 2 shows the results of applying the Silhouette Index measure to the k-means clustering algorithm with the Elbow criterion.

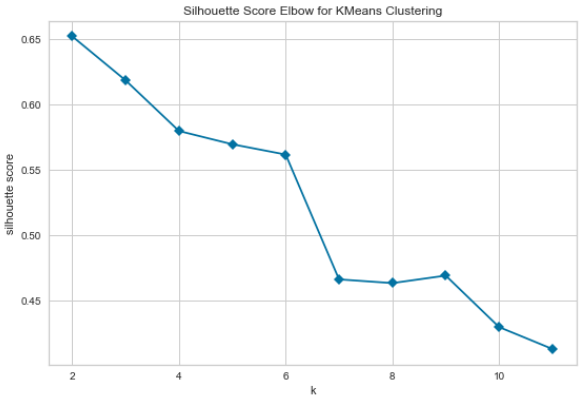

Figure 2. Silhouette score for k-means clustering using Elbow criterion

There are elbows for the number of clusters 4, 7 and 10. The common value of the elbow for k with Calinski Harabasz is for k = 4, and this value represents the optimal number of clusters. Table 1 shows the values of all validation scores with k-means algorithm depending on the number of clusters.

Table 1. Values of validation measures with k-means clustering

c	4	5	6	7	8
SI	0.57	0.57	0.56	0.47	0.46
CHI	4988.7	4988.69	4813.78	4884.29	4842.76
DBI	0.72	0.69	0.76	0.76	0.81

B. Results of Agglomerative Clustering

It is also possible to define the optimal number of clusters for agglomerative clustering by this method. To illustrate this, the results of agglomerative clustering are shown for the values of the Silhouette index and Calinski-Harabasz validation measures (Fig. 3).

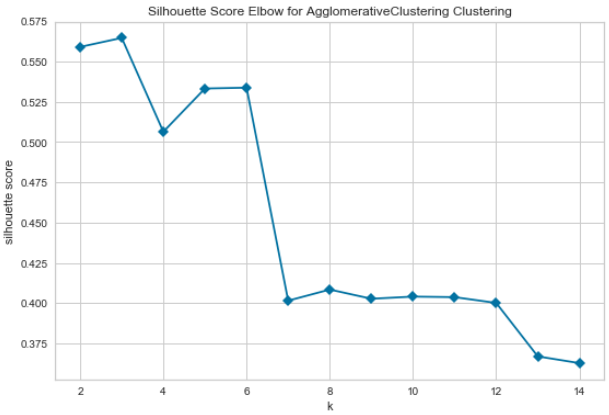

Figure 3. Silhouette score for agglomerative clustering

Figure 3 shows that the elbow is at the values of k=4, k=7, k=13, for the values of the Silhouette Index for agglomerative clustering. In Figure 4 the elbows are at the values k=4, k=7, k=13. Which means that the value k=4 in this case also proves to be the most appropriate for this clustering algorithm.

Table 2 shows the values of all agglomerative clustering measures.

Table 2. Values of validation measures with agglomerative clustering

c	4	5	6	7	8
SI	0.51	0.53	0.53	0.40	0.41
CHI	4024.08	4138.52	4092.10	4081.68	4149.26
DBI	0.83	0.78	0.72	0.74	0.80

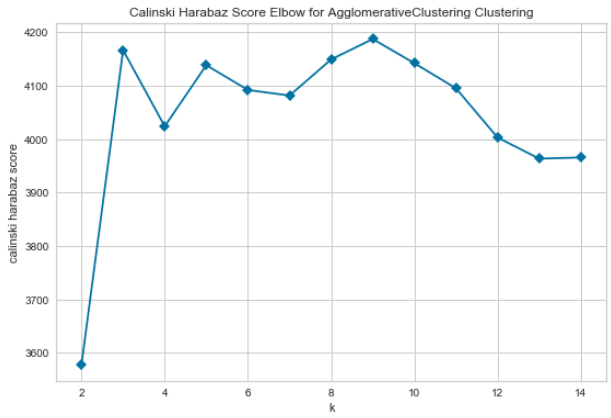

Figure 4. Calinski Harabasz score for agglomerative clustering

C. Results of BIRCH algorithm

We obtained the same results for the BIRCH clustering algorithm as for agglomerative clustering, as can be seen in Figures 5 and 6.

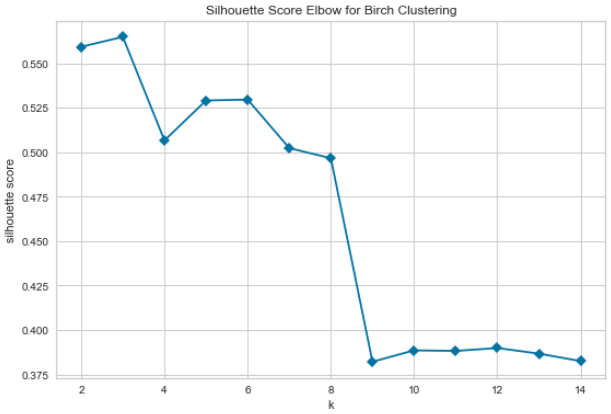

Figure 5. Silhouette score for BIRCH

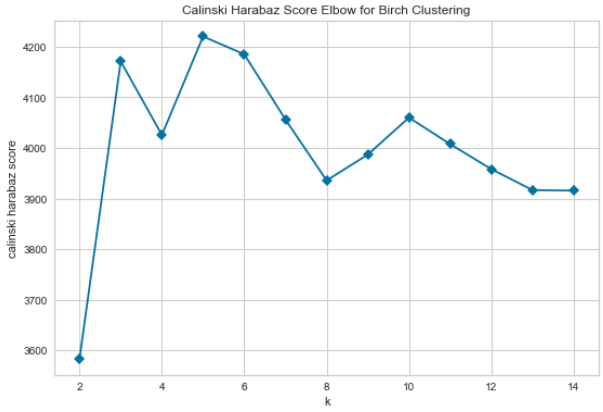

Figure 6. Calinski Harabasz score for BIRCH

D. Comparison of clustering algorithms

It is evident from the previous section that the optimal number of clusters can be determined by validation methods. Validation measures are also used to compare different clustering methods. The optimal number of clusters

obtained by previous experiments is c = 4. The validation measures for c = 4 of all clustering algorithms are summarized in Table 3.

Table 3. Values of validation measures for c=4

	SI	CHI	DBI
K-means	0.5663	4673.2201	0.7189
Agg. Clust.	0.5064	4024.0851	0.8291
BIRCH	0.5067	4025.6388	0.8291

From the validation measures for k=4 shown in Table 3 agglomerative clustering has the best results. The Figures 7, 8 and 9 show the cluster diagrams for each of the clustering algorithms.

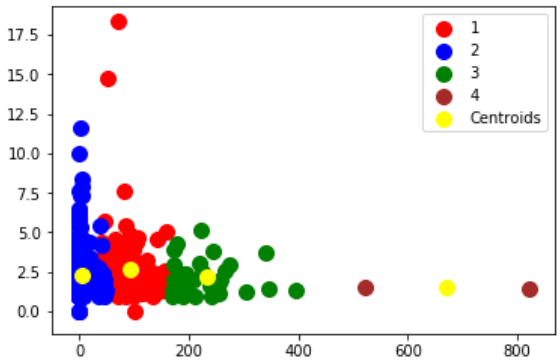

Figure 7. Clusters created using k-means algorithm

The points in clusters are displayed with different colors. From the diagrams is evident that the best distributed clusters with similar number of points has agglomerative clustering, while k-means and BIRCH have one cluster with significantly smaller number of data points (users) and the center of that cluster is located far away from the centers of the other three clusters.

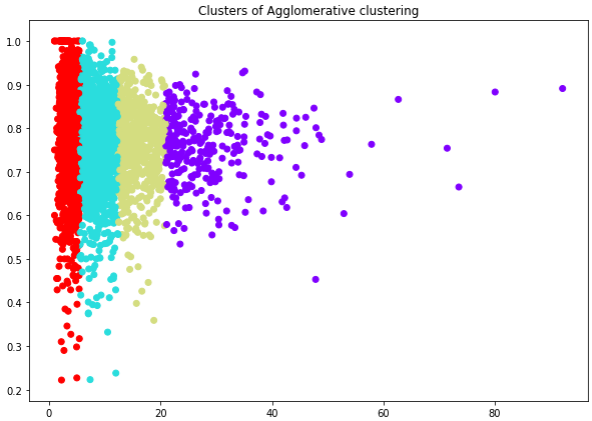

Figure 8. Clusters created using agglomerative clustering

We tried to interpret the results obtained from the clustering algorithms based on the attribute values of the clients assigned to the clusters:

Cluster 0 consists of the users who make calls to the same operator more than the mean value of this attribute, but at the same time have values for the attribute outgoing call and unique codes below the mean value, which means that in this cluster are users that make calls to the same operator.

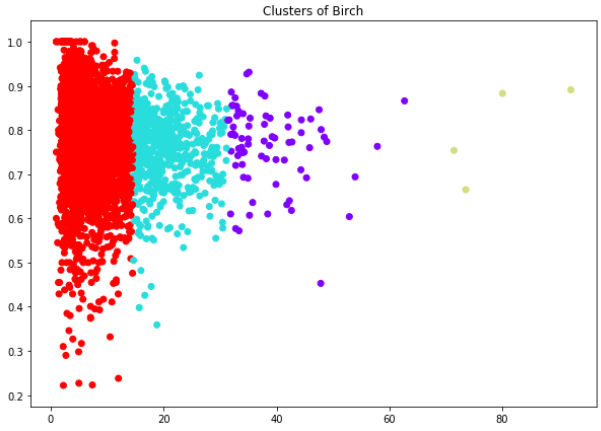

Figure 9. Clusters created using BIRCH algorithm

Cluster 1 is created of users who have extremely high values, above the average, for the feature duration of the call as well as a value above the average for sent SMS messages. The value for outgoing calls is below average, which means this is a cluster of users with long conversations who send many SMS messages.

In cluster 2 we have a value lower than the mean for the first attribute, call duration, but higher than the mean for outgoing calls and calls to the same operator. These users do not talk in different area codes. This is a cluster of users who make outgoing calls in the same area code and operator.

Users in cluster 3 make outgoing calls and send SMS calls to the same operator, with values above the mean for these attributes. The call duration is very short compared to the average value for this attribute.

V. Conclusion

The goal of this paper is to best describe the differences between the various types of users served by a telecommunication operator. This would provide the operator with insights on how to improve its telecommunications services and packages and offer appropriate services to specific user profiles.

The segmentation of users is performed with several clustering algorithms, such as k-means, agglomerative clustering and BIRCH, and the number of initial clusters is determined by the Elbow criterion.

The algorithms are implemented using Python and scikit-learn package. The algorithms were compared using validation measures to confirm the accuracy of the obtained clusters using the methods Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index. Experiments

have shown that the best results are achieved by agglomerative clustering and 4 clusters.

For future work we plan to create an integrated clustering environment that will offer an effective combination of methods for intelligent data analysis and for solving real business problems. Because of the different types of problems and specificities that each of them possesses we believe that the best results can be achieved with specific solutions adapted to each problem. The integrated software environment will offer methods that can be applied to various consumer-oriented businesses.

REFERENCES

[1]. Berry, M., & Linoff, G. *Data mining techniques for marketing, Sales and customer support*. John Wiley & Sons, 1997.

[2]. Berry, M., & Linoff, G. *Mastering data mining*. John Wiley & Sons.

[3]. Dhar, V. (2011) Prediction in financial markets: The case for small disjuncts. *ACM Transactions on Intelligent Systems and Technologies*, 2 (3), 2000, pp.1-22.

[4]. Han, J., & Kamber, M. *Data mining: concepts and techniques*. Morgan Kaufmann, 2001.

[5]. Kohlbacher, M. The Effects of process orientation on customer satisfaction, product quality and time-based performance. In *Proc. of the 29th International conf. of the strategic management society*, October 11–14, Washington DC. 2009.

[6]. Witten, I., H., & Frank, E. *Data mining: Practical machine learning tool and techniques with Java implementations*, Morgan Kaufmann, 2000.

[7]. Witten, I., H., & Frank, E. *Data mining*, Morgan Kaufmann, 2001.

[8]. K. Tsiptsis, K., & Chorianopoulos, A. Segmentation applications in telecommunications, In *Data mining techniques in CRM: Inside customer segmentation*, Wiley, 2009, pp. 291-332.

[9]. Weiss, G., M. *Data Mining in the telecommunications industry*, Fordham University, USA, 2001.

[10]. Singh, H., & Kaur, K. Review of existing methods for finding initial clusters in k-means algorithm, *International Journal of Computer Applications*. 68(14), 2013, pp. 24-28.

[11]. Feldman, R., & Dagan, I. Knowledge discovery in textual databases, in *Proc. 1st Int. Conf. Knowledge Discovery and Data Mining*, 2015.

[12]. Frawley, W. J., Piatetsky-Shapiro G., & Matheus, C.J. *Knowledge Discovery In Databases: An overview*, AAAI/MIT Press, 1991.

[13]. Ahola, J., & Rinta-Runsala, E. *Data mining case studies in customer profiling*, VTT Information Technology, 2001.

[14]. McDonald, M., & Dunbar, I., *Market segmentation: how to do it and how to profit from it*, Boston: Artech House, 1997.

[15]. Sotiropoulos, M. V. S. A. T. N. Constructing stereotypes for an adaptive e-shop Using AIN-Based Clustering, in *ICANNGA*, 2007.

[16]. We, C. P., & Chiu, T., Turning telecommunications call details to churn prediction: A data mining approach, in *Expert Systems with Applications 23*, 2002.

[17]. Müller, A. C., & and S. Guido, S., Agglomerative clustering, in *Introduction to Machine Learning with Python*, O'Reilly, 2017, pp. 182-187.

AUTHORS

First Author – Cveta Martinovska Bande, PhD in Computer Science, Computer Science Faculty, University Goce Delcev, ul.Krste Misirkov, 10-A, Shtip, North Macedonia
cveta.martinovska@ugd.edu.mk

Second Author – Natasha Stojkovic, PhD in Computer Science, Computer Science Faculty, University Goce Delcev, ul.Krste Misirkov, 10-A, Shtip, North Macedonia
natasa.stojkovik@ugd.edu.mk

Third Author – Elena Kosturanova, MSc in Computer Science, Computer Science Faculty, University Goce Delcev, ul.Krste Misirkov, 10-A, Shtip, North Macedonia
kosturanova1992@gmail.com

Fourth Author – Mimoza Klekovska, PhD in Computer Science, American University of Europe, bul. Kiro Gligorov, bb, Skopje, North Macedonia
mimiklek@yahoo.com

Correspondence Author – Cveta Martinovska Bande, PhD in Computer Science, Computer Science Faculty, University Goce Delcev, ul.Krste Misirkov, 10-A, Shtip, North Macedonia
cveta.martinovska@ugd.edu.mk
+ 389 78 207 763