Düzce University
Journal of Science & Technology

Research Article

Initial Seed Value Effectiveness on Performances of Data Mining Algorithms

Tunahan TİMUÇİN[a,*], İrem DÜZDAR ARGUN[b]

[a] *Department of Computer Engineering, Faculty of Engineering, Düzce University, Düzce, TURKEY*
[b] *Department of Industrial Engineering, Faculty of Engineering, Düzce University, Düzce, TURKEY*
* *Corresponding author's e-mail address: tunahantimucin@duzce.edu.tr*
DOI: 10.29130/dubited.813101

ABSTRACT

After 2000s, Computer capacities and features are increased and access to data made easy. However, the produced and recorded data should be meaningful. Transformation of unprocessed data into meaningful information can be done with the help of data mining. In this study, classification methods from data mining applications are studied. First, the parameters that make the results of the same data set different were investigated on 4 different data mining tools (Weka, Rapid Miner, Knime, Orange), It has been tested with 3 different algorithms (K nearest neighborhood, Naive Bayes, Random Forest). In order to evaluate the performance of the data set while creating the classification models, the data set was divided into training data and test data as 80% -20%, 70% -30% and 60-40%. The accuracy, roc and precision values was used to test the performance of the classifying data. While classifying, the effect of algorithm parameters on the results is observed. The most important of these parameters is the initial seed value. The initial seed is a value using especially in classification algorithms that determines the initial placement of the data and directly affects the result. In this respect, it is very important to determine the initial seed value correctly. In this study, initial seed values between 0 and 100 were evaluated and it was shown that the classification could change the accuracy value approximately by 5%.

Keywords:*Data mining, Classification, Credit approval, Seed value.*

Veri Madenciliği Algoritmalarının Performanslarında İlk Tohum Değer Etkinliği

ÖZET

2000'li yıllardan sonra, Bilgisayar kapasiteleri ve özellikleri artmış ve verilere erişim kolaylaşmıştır. Ancak üretilen ve kaydedilen veriler anlamlı olmalıdır. İşlenmemiş verilerin anlamlı bilgilere dönüştürülmesi, veri madenciliği yardımı ile yapılabilmektedir. Bu çalışmada, veri madenciliği uygulamalarından sınıflandırma yöntemleri incelenmiştir. Öncelikle aynı veri setinin sonuçlarını farklı kılan parametreler 4 farklı veri madenciliği aracında (Weka, Rapid Miner, Knime, Orange) araştırılmış, 3 farklı algoritma ile test edilmiştir (K nearest neighborhood, Naive Bayes, Random Forest). Sınıflandırma modelleri oluşturulurken veri setinin performansını değerlendirmek için veri seti eğitim verileri ve test verileri olarak % 80-% 20, % 70-% 30 ve% 60-40 olarak ayrılmıştır. Accuracy, roc and precision değerleri, sınıflandırma verilerinin performansını test etmek için kullanılmıştır. Sınıflandırma yapılırken algoritma parametrelerinin sonuçlar üzerindeki etkisi gözlemlenmiştir. Bu parametrelerden en önemlisi ilk tohum değeridir. İlk tohum, özellikle verilerin ilk yerleşimini belirleyen ve sonucu doğrudan etkileyen sınıflandırma algoritmalarında kullanılan bir değerdir. Bu açıdan ilk tohum değerinin doğru belirlenmesi çok önemlidir. Bu çalışmada 0 ile 100 arasındaki başlangıç tohum

Received: 19/10/2020, Revised: 27/01/2021, Accepted: 02/02/2021

değerleri değerlendirilmiş ve sınıflandırmanın doğruluk değerini yaklaşık %5 değiştirebileceği gösterilmiştir.

Anahtar Kelimeler:*Veri madenciliği, Sınıflandırma, Kredi onayı, Tohum değeri.*

I. INTRODUCTION

Nowadays, the big databases, also known as Big-Data, are made up of huge dimensions that reach the level of Petabyte or even the Exabyte. However, the data within these databases, which are produced and recorded with the growing technology, are not fully usable. The desired situation is to process and convert this data into a knowledge for a specific purpose. This conversion process is called data analysis [1]. Data analysis methods that constitute the technical part of data mining; classification, clustering, association rule can be defined as.

Clustering; this is the grouping process, which is the result of similarities between the data [2,3]. In other words, clustering refers to the distributions of the data group obtained through observations or data elements to the groups that they create without any training process [4-6].

Association Rules; is one of the first techniques used in data mining [7,8,9]. Today, the Association Rules [10] Analysis is known as the “Recommended Engine".

Classification is the assignment of a data set to one of the different and predetermined classes. Classification algorithms learn which data to assign to which class from the given training set. It then tries to assign test data to the correct classes. Values that indicate the classes of data are called labels.It is one of the most used fields of data mining [11-13]. This type of analysis is also known as supervised analysis. A certain part of the data obtained from the database is used to help the system learn how to behave against possible new situations. The remaining part is used as an experimental set after the training process and the success of the system training is checked.

In this study, the study structure of the classification method, one of the data analysis methods, was examined by various algorithms and the values obtained from the applied data set were compared [14-16].

II. OPEN SOURCE DATA MINING TOOLS

There are many tools developed in Data Mining. While some of these tools are commercial, some are open source. For this reason, data mining tools are divided into two groups as commercial and open source. For example, MATLAB, SAS, DataMelt, SPSS and Oracle's modules are developed for this purpose. Open source tools are Orange, RapidMiner, WEKA, Keel, Knime and Tanagra can be given as an example [17, 18, 19].This section provides an overview of the open source and free available Knime, Orange, RapidMiner and Weka tools.

A. WEKA

Weka is the name of the tool developed in the University of Waikato for the purpose of machine learning and consists of the initials of the "Waikato Environment for Knowledge Analysis". Today, most of the widely used machine learning algorithms and methods are included. Thanks to the fact that it was developed in Java and its libraries are in the form of jar files, the ability to integrate easily into projects written in Java has further expanded its use [20].

Weka [21] is a Java-based open source data mining tool that is currently under development. It provides access to SQL databases using JDC. It includes machine learning algorithms. The tool is able

to read file structures with the extension * .arff (Attribute Relationship File Format). The following operations, which are basically data mining methods in data mining, can be performed [22];

- Classification
- Clustering
- Association

B. KNIME

The name KNIME consists of the abbreviation Konstanz Information Miner. KNIME is an open source and cross-platform integration, data analysis, reporting platform. KNIME includes different components for data mining through the data line concept, and these tools are called "nodes". For visualization, modeling and data analysis (ETL), basic data preprocessing nodes can be used without writing any code in a user graphical interface.

For 12 years, KNIME has been used in pharmaceutical research [23] in CRM analyzes, business intelligence and financial applications.

KNIME [24] had been begun development in January 2004 by a software team at the University of Konstanz. Michael Berthold, a member of the original developer team, came from a company that provided the pharmaceutical industry software in the valley of silicon. Their first objectives were to create a modular highly graphical and measurable and open data processing tool that was allowed for easy integration of different conversion analysis, data processing and visual research modules that did not focus on any area. The tool was designed as a collaboration and research tool to serve integration analysis and other data processing projects.

Eclipse is a software developed on the Rich Client Platform. Besides the file types that Weka and RapidMiner can read, KNIME can also read data directly from * .txt files. It also supports data reading in XML based language called PMML (Predictive Model Markup Language), which allows data transfer between statistical applications [25].

C. ORANGE

In addition to being an open source data mining program, Orange [26] draws attention with its easy-to-use interface that it provides to users. With the help of Orange program, users are user-friendly with many features such as data preparation, data modeling, exploration. In addition, it contributes to python and visual programming issues as well as features designed for data mining. In addition, it is an advanced program with components in many different areas such as machine learning, text mining, bioinformatics. Orange, which has advanced data analysis features, is a machine learning-enabled software package that contains books on coding, data analysis and visualization thanks to its powerful and flexible structure.

Graphical end-user interface, which is built on a cross platform, Orange can also be used with various programming languages (Python, C ++). The program, which can be obtained free of charge with the General Public License, includes many components such as modeling, evaluating the obtained models, and discovering. Supported by many operating systems such as different versions of Windows, Apple, Linux, Orange was developed by the University of Ljubljana.

Because of its data analysis workflow, Orange comes to the forefront with its tabular features such as frame features, comparison and visualization of many learning algorithms. The program that supports * .tab, * .tsv, * .arff, * .csv, * .txt files is missing at the point of visualization of these files.

D. RAPID MINER

It is written using Java programming language. It is preferred by users because it reduces the need to write code to almost zero. The reduced need for coding reduces the risk of errors to a minimum. RapidMiner features data mining and deep learning, including data conversion, data preprocessing and visualization, statistical modeling and predictive analytical and evaluation.

RapidMiner is a software platform developed by the company with the same name as the software for machine learning and data mining needs. Can use client / server architecture and work on a cloud structure as Software as a Service (SaaS). It focuses on research and education in general. In this sense, it is possible to characterize RapidMiner as a community software. It can be widely used commercially as it can be used for purposes such as rapid prototyping and application development. RapidMiner Studio, which meets the needs with RapidMiner Studio, RapidMiner Server, RapidMiner Radoop and RapidMiner Cloud products, is free to use for academic research and for academic research.

RapidMiner [27] is Java-based data mining tool developed by Ralf Klinkenberg, Ingo Mierswa and Simon Fischer. They work in Artificial Intelligence Unit of the Dortmund University of Technology. Unlike other data mining tool, it can read and process data in 22 file formats. Similar to Weka tool, it has many algorithms. It supports PostgreSQL, MS SQL, MySQL, Oracle, Access databases assoc. On the other hand, Office is able to connect with excel files and visualization is one of the rich tool with graphical interface [28].

III. DATA MINING CLASSIFICATION ALGORITHMS

In this section, information about algorithms of data mining, KNN (K nearest neighborhood), Naive Bayes and Random Forest algorithms are provided to make big data into meaningful strings.

A. KNN

KNN algorithm; is an algorithm that determines that when a space that class quality is not specified, is added to the elements whose class quality is specified, this example must be included in the closest class. It uses a k variable to determine the closest class. This variable k specified the number of k elements that are the closest to the sample. The following 3 distance functions are commonly used when calculating the distance between the sample and the elements. These are the Euclidean, Minkovski and Manhattan.

Summarizing the algorithm with a few items;

1. Add a new instance to a space of elements of a class.
2. Determine the K number.
3. Calculate the distance between the new sample and the elements (using one of the distance functions)
4. Examine the class quality of the nearest k elements.
5. Include the sample from the nearest k-element in the class whose number is more than the others.

B. NAÏVE BAYES

The basis of the Naive Bayes classifier [29] is based on Bayes' theorem. Lazy (lazy) is a learning algorithm, it can also work on unstable datasets. The way the algorithm works calculates the probability of each state for an element and classifies it based on the highest probability value. It can do very successful works with a little training data. If a value in the test set has an unobservable value in the training set, it returns 0 as a probability value, that is, it cannot estimate. This is commonly

known as the Zero Frequency. Correction techniques can be used to resolve this situation. One of the simplest correction techniques is known as Laplace prediction. Examples of usage areas are real-time forecasting, multi-class forecasting, text classification, spam filtering, sensitivity analysis and suggestion systems.As with any classification problem, the aim of the algorithm is to create an education from the information given by using a vector with multiple characteristics and to classify the new data correctly as a result of this training. The higher the number of data taught, the more precise it is to determine the actual category of the test data.

The Bayes' theorem allows the calculation of P (a | y) from P (a), P (y) and P (y | a) as posterior probability. The Naive Bayes classifier specified that the value of an estimator (y) on a given class (a) is independent of the values of other estimators. This situation is called class conditional independence. In Equality given 1, P (a | y) is the posterior probability of the (class) given class, P (y | a) is the probability of being the likelihood of the given class, P (y) is the previous probability of the estimator and P (a) is the primary probability of the class.

$$P(a|y) = \frac{P(y|a)P(a)}{P(y)} \quad (1)$$

C. RANDOM FOREST

Random Forest [30] is easy-to-use and a flexible machine learning algorithm. Collective classification methods are learning algorithms that produce multiple classifiers instead of a classifier and then classify new data with votes from their predictions. The most commonly used batch classifiers are Bagging, Acceleration and RandomForest. RandomForest can be defined as a collection of tree-type classifiers. It is an improved version of the Bagging method with the addition of randomness feature. Rather than dividing each node into branches using the best branch among all variables, RandomForest branches each node using the best among the randomly selected variables in each node. Each data set is produced interchangeably from the original data set. Then trees are developed using random feature selection. Developed trees are not pruned. This strategy makes the accuracy of RandomForest unique.

RandomForest is also very fast, resistant to over-sleeping, and the more trees you work with, the more you want.An important advantage of random forest is that it can be used for both regression and classification problems, which make up the majority of available machine learning systems. Figure 1. shows a Classic Random Forest Model.

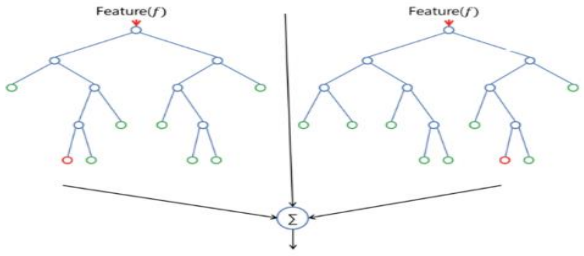

Figure 1.*Classic Random Forest Model [31]*

D. DATA SETS

In this article, 2 data set are used.

The first of these datasets is Quinlan and Mason et al. [32,33] the Credit Approval data set used in the UCI Machine learning repository [34]. This data set is related to credit card applications. Data set; It includes features such as employment status, gender, age, marital status, place of residence, monthly payment, and payment opportunity. All property names and values have been replaced with

meaningless symbols to protect the privacy of the data. It consists of 15 input features and 1 result. These 15 properties are classified and aimed to identify those who are eligible for the loan. This dataset is so appropriate for classification because there is a good mix of attributes; continuous data, nominal with larger numbers of values and nominal with small numbers of values. There are also a few missing values.

The second data set consists of data recorded by 12 users in a motion capture environment by wearing gloves with sensors attached for 5 types of hand postures. Although there is missing data due to resolution, most of the data is preprocessed. Data Uci is the "MoCap Hand Postures" data set on the Machine Learning data set. It consists of 37 attributes and 1 class [35].

IV. APPLICATION AND RESULTS

Credit Approval data set on Weka, RapidMiner, Knime and Orange tools, Naive Bayes, K-nn and Random Forest Algorithms 80% -20%, 70% -30% and 60% -40% (Training data - Test data) rates has been applied.

Table 1, shows the results obtained from 4 different tools where the Credit Statement data set is allocated as 80% -20% (Training data-Test data). In this partitioning, similar results are obtained in RapidMiner and Orange. This situation can show that RapidMiner and Orange have same coding structure and use same features. Compared to all tools, the best result was always obtained by the random forest algorithm. The random forest algorithm is an algorithm suitable for this data set and for many data sets thanks to the creation of forests according to all properties. Random Forest, showed an average success rate of over 85% in the case of this partitioning.

Table 1.*Results obtained (%80 Training-%20 Test)*

#	**Weka** *(%)*	**Rapid Miner** (%)	**Knime** *(%)*	**Orange** *(%)*
Naive Bayes	76.33	79.39	84.70	78.64
Knn (IBk)	77.86	70.23	58.77	70.48
Random Forest	**83.96**	**89.31**	**87.78**	**89.54**

Table 2, shows the results obtained from 4 different tools platforms where the Credit Statement data set is allocated as 70% -30% (Training data-Test Data). The very unstable results in the K-nn algorithm show that the K-nn algorithm is not an appropriate algorithm for this data set. Different results are caused by various effects in different tools for K-nn Algorithm. Some of these reasons; the re-coding of the algorithm in open-source tools (also understood from the change of the name of the algorithm. For example, in Weka =IBK, others=K-nn), in Weka, the search algorithm can also be selected in the Ibk or Knn algorithm (eg Linear NN search), it can't in other platforms. These cases indicate that better results are obtained by applying the re-encoded algorithm with IBK name. The results show that the success increased by around 10%.

Table 2.*Results obtained (%70 Training-%30 Test)*

#	**Weka** *(%)*	**Rapid Miner** *(%)*	**Knime** *(%)*	**Orange** (%)
Naive Bayes	78.06	78.57	84.18	77.46
K-nn (IBk)	79.08	70.41	61.73	71.02
Random Forest	**81.63**	**89.80**	**87.75**	**88.41**

Table 3, shows the results obtained from 4 different tools platforms where the Credit Statement data set is allocated as 60% -40% (Training data-Test Data). The number of Seed is the feature that affects

all tools in every situation. The Seed number is set default to 10 in the Weka tool, while the other tools can only be activated with the "use random seed number" feature. Rapid Miner has a Laplace transform by default for the Naive Bayes algorithm. Because Naive Bayes Algorithm does not have many different parameters, the results are obtained in close values in all tools.

Table 3.*Results obtained (% 60 Training-% 40 Test)*

#	**Weka** (%)	**Rapid Miner** *(%)*	**Knime** (%)	**Orange** *(%)*
Naive Bayes	77.01	78.93	82.82	79.21
K-nn (IBk)	79.69	66.67	64.50	74.62
Random Forest	**82.75**	**89.27**	**85.87**	**87.38**

A. COMPARISON OF SOFTWARES

Data mining software are computer programs that can have different algorithm parameters, different coding infrastructure, and enable us to search for correlations that can make predictions about the future from large data heaps [36].

In this study, by using the same algorithms on the sample data set, the reasons for the difference of the results were investigated.Table 4 shows the reasons arising from the different dynamics of the software and which may affect the results.

Table 4.*Conditions effecting results in data mining softwares*

Weka	➢ The Knn algorithm is referred to as the Ibk algorithm under the Lazy methodology in Weka. This indicates that the algorithm is re-encoded and affects the result obtained. ➢ In Weka, the search algorithm can also be selected in the Ibk or Knn algorithm (e.g. Linear NN search), it can't in other platforms.
Orange	○ Distance calculation method can be selected in Knn algorithm in Orange platform. But, it comes by default on other platforms.
Rapid Miner	❖ Unlike other tools, RapidMiner offers conversion for the Naive Bayes algorithm (e.g. Laplace). ❖ In addition, because of the forced type conversion (e.g. numeric to nominal) in RapidMiner, decreases the number of instances and it changes the results. ❖ Finally, in RapidMiner, the random forest algorithm provides the option of selecting one of the best split methods (Gini_index, information gain, gain_ratio) when applying to the data set.
Knime	▪ Knime is a free model-based data mining software. The value conversions that need to be made in the established models may cause a decrease in the data set and affect the results.

A. EFFECT OF SEED NUMBER

The initial seed number is critical for knowledge discovery. The Seed number determines the starting sequence for dataset. In general, the initial seed value is used in classification algorithms. There are several methods to find seeds. Two of the most known methods are Random and Partial [37]. Adrian et al., have graphically demonstrated the importance of seeding in the fuzzy-c means algorithm [38]. Rahman and Islam have demonstrated the effect of the initial seed value on the biomedical dataset [39]. Mahdi et al., showed the change in clustering status by applying different initial seed values on multiple sets of data [40].

This study shows that the initial seed value is an important factor in the success of the classification algorithms. Random forest algorithm was applied to 2 different data sets. While running the algorithm, in order to observe the effectiveness of the initial seed value, only the seed value has been changed while all other parameters are constant. The number of seeds thought to change the results was tried by assigning all integers from 1 to 100 and the following results were obtained.

Accuracy, Roc Area and Precission values were used for comparison. The results obtained revealed the importance of determining the initial seed value correctly.

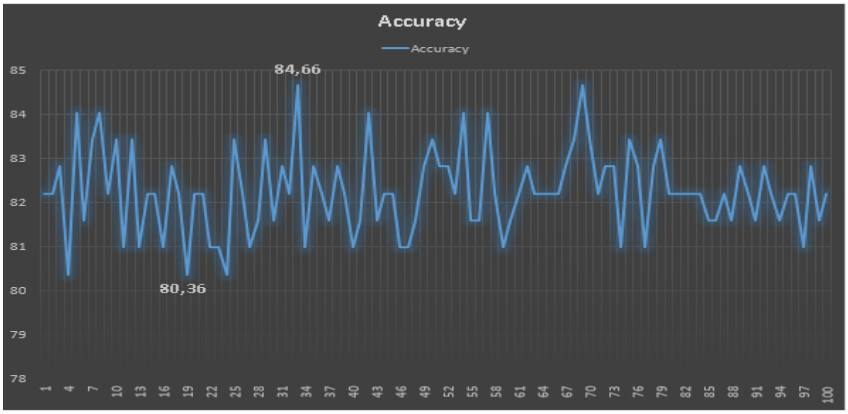

Figure 2. *Changes in accuracy value according to the number of Initial seeds.*

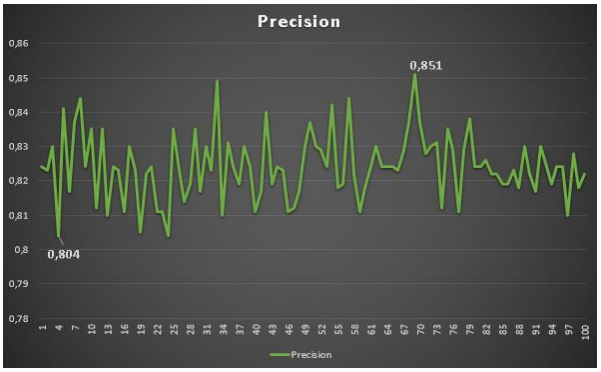

Figure 3. *Changes in precission value according to the number of Initial seeds.*

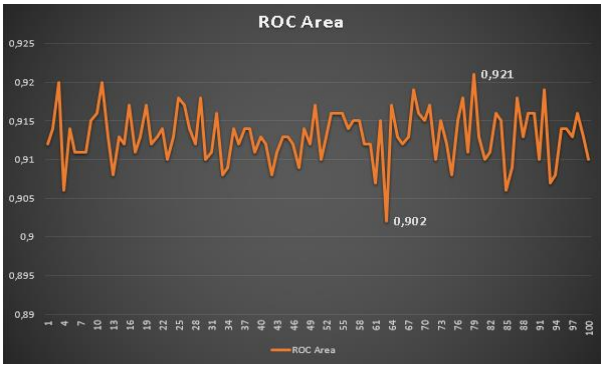

Figure 4. *Changes in Roc Area Valueaccording to the number of Initial seeds.*

The values in figure 2, figure 3 and figure 4 show that determining the initial seed value correctly can provide approximately 5% better classification in the "Credit Approval" dataset, which is the first dataset.

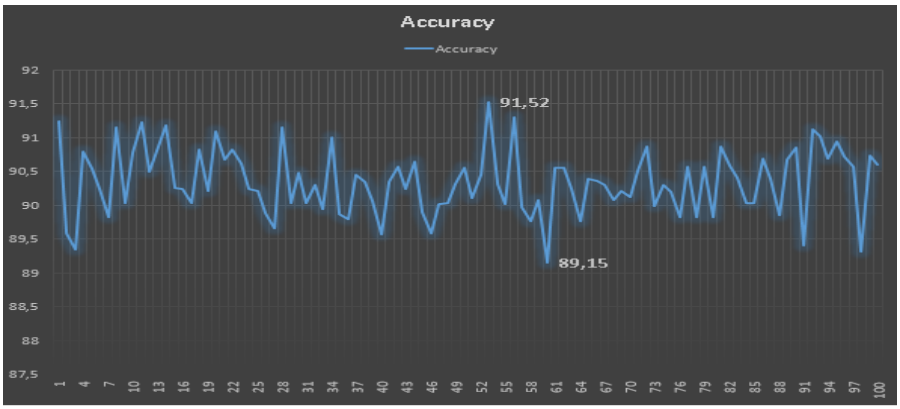

Figure 5.*Changes in accuracy value according to the number of Initial seeds.*

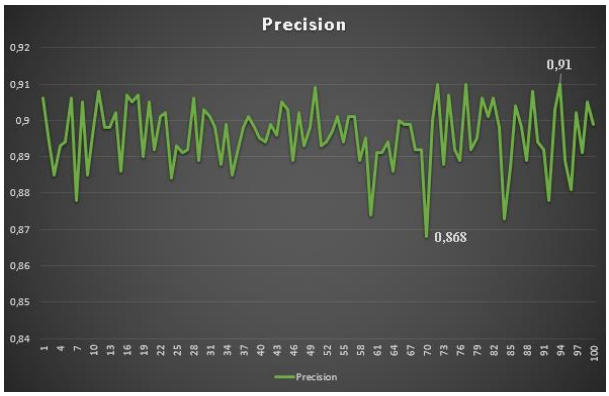

Figure 6.*Changes in precission value according to the number of Initial seeds.*

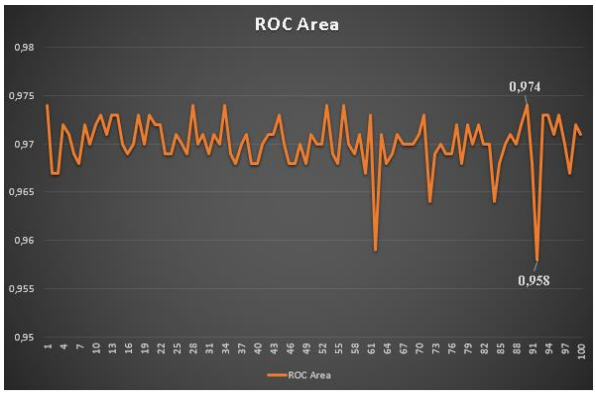

Figure 7*.Changes in Roc Area Valueaccording to the number of Initial seeds.*

The values in figure 5, figure 6 and figure 7 show that determining the initial seed value correctly can provide approximately 3% better classification in the "MoCap Hand Postures" dataset, which is the second dataset.

These results revealed the importance of determining the initial seed value correctly. Correct determination of this parameter, which is evaluated on 2 datasets and shows that classification can be made up to 5% better, is an important area that needs to be investigated. In these 2 data sets, initial seed values between 0 and 100 were evaluated. If the initial seed value is evaluated with a larger number of integers, there may be a greater improvement in the results. It is a promising new field of study, such as determining the number of seeds with the best success from all results according to each data set to further the study.

V. DISCUSSION AND CONCLUSION

In this study, 3 different data mining algorithms were applied to 2 randomly selected data sets in the UCI Machine learning pool and the results were obtained. At the same time, algorithms were evaluated with 4 different free data mining tools and the results obtained were compared. According to these results, the Random Forest Algorithm works best in all situations. However, it is also a subject of research that leads to different results in different tools.

There are parameters that cause different results in different mining software and can improve success results up to 5%. These parameters can be considered in 2 categories as the less effecting parameters and the more effecting parameters. The less effecting parameters (providing 1-2% change at most) are as follows; Best Partition feature (gini_index, information_ gain, gain_ ratio), sample type (mixed sampling, stratified sampling, linear sampling), sample reduction during decomposition (numerical to nominal, etc.), distance calculation techniques (Euclidean distance, Manhattan distance, Minkowski distance, etc.), transformation state (for Naive Bayes -> Laplace), maximum depth and number of trees (for random forest).

The more effective parameter is the initial seed value. Since the initial seed value can provide up to 5% improvement, this study was conducted on this value.

All other substances, except the number of Seed, can change the accuracy of the classification in the range of 1% to 3% when applied to this data set. On the other hand, the seed number can increase the accuracy of classification by approximately 5% in this data set. Therefore, this study focused on the effect of initial seed number and the results showed the importance of choosing the correct initial seed number.

There are several methods of assigning the initial seed number. Two of the most common methods are Random and Partial. The seed number determines the starting order of the data set. If the same algorithm and the same data set are used and the result is desired to be the same, the seed number should be selected the same. In clustering too, changing the number of cores allows the algorithm to start with different random cluster centers. In this case, the random number generator must be of good quality to get good results. The core number is also used in the field of computer security. Therefore, it is very important to choose the correct seed number. In the next study, it is planned to work on the program that enables the selection of the most suitable seed value for the data set to be mined.

VI. REFERENCES

[1] M. S. Durmuş, "Veri kümeleme algoritmalarının performansları üzerine karşılaştırmalı bir çalışma," Yüksek Lisans tezi, Elektrik-Elektronik Mühendisliği, Pamukkale Üniversitesi, Denizli, Türkiye, 2005.

[2] Y. Farhang, "Face extraction from ımage based on K-Means Clustering Algorithms," *International Journal Of Advanced Computer Science And Applications*, vol. 8, no. 9, pp. 96-107, 2017.

[3] H. Kaya and K. Köymen, "Veri madenciliği kavramı ve uygulama alanları," *Doğu Anadolu Bölgesi Araştırmaları Dergisi,* vol. 6, no. 2, pp. 159-164, 2008.

[4] Q. Chen, Y. Wan, X. Zhang, Y. Lei, J. Zobel and K. Verspoor, "Comparative analysis of sequence clustering methods for deduplication of biological databases," *Journal of Data and Information Quality (JDIQ),* vol. 9, no. 3, pp. 1-27, 2018.

[5] M. A. Alan, "Veri madenciliği ve lisansüstü öğrenci verileri üzerine bir uygulama,"*Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, vol. 33, pp. 165-174, 2012.

[6] S. Özşen and R. Ceylan, "Comparison of AIS and fuzzy c-means clustering methods on the classification of breast cancer and diabetes datasets," *Turkish Journal of Electrical Engineering & Computer Sciences,* vol. 22, no. 5, pp. 1241-1254, 2014.

[7] G. Kayakutlu, I. Duzdar, E. Mercier-Laurent and B. Sennaroglu, "Intelligent association rules for innovative SME collaboration," *International Federation for Information Processing (IFIP) International Workshop on Artificial Intelligence for Knowledge Management*, Springer, Cham, 2014, pp.150-164.

[8] A. M. Moawad, A. M. Gadallah and M. H. Kholief, "Fuzzy ontology based approach for flexible association rules mining," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, pp. 328-337, 2017.

[9] T. Pala, I. Yücedağ and H. Biberoğlu, "Association rule for classification of breast cancer patients," *Sigma*, vol. 8, no. 2, pp. 155-160, 2017.

[10] R. A. Shah and S. Asghar, "Privacy preserving in association rules using a genetic algorithm," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 22, no. 2, pp. 434-450, 2014.

[11] I. C. Yeh and C. H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473-2480, 2009.

[12] A. Dhall, G. Sharma, R. Bhatt, and G. M. Khan, "Adaptive digital makeup," in*International Symposium on Visual Computing*, Berlin, Heidelberg, 2009, pp. 728-736.

[13] E. J. Lauría, A. D. March, "Combining bayesian text classification and shrinkage to automate healthcare coding: A Data Quality Analysis," *Journal of Data and Information Quality (JDIQ),* vol. 2, no. 3, pp. 13, 2011.

[14] K. Rangra and K. L. Bansal, "Comparative study of data mining tools," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 6, pp. 216-223, 2014.

[15] F. Bulut and I. O. Bucak, "An urgent precaution system to detect students at risk of substance abuse through classification algorithms," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 22, no. 3, pp. 690-707, 2014.

[16] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi and E. M. Al-Shawakfa, "A comparison study between data mining tools over some classification methods," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 2, pp. 18-26, 2011.

[17] A. Tekerek, "Veri madenciliği süreçleri ve açık kaynak kodlu veri madenciliği araçları," *Akademik Bilişim*, vol.11, pp. 2-4, 2011.

[18] M. Dener, M. Dörterler and A. Orman, "Açık kaynak kodlu veri madenciliği programları: WEKA'da örnek uygulama," *Akademik Bilişim*, vol. 9, pp. 11-13, 2009.

[19] E. Atagün and İ. D. Argun,"A comparison of data mining tools and classification algorithms: Content producers on the video sharing platform,"in *The International Conference on Artificial Intelligence and Applied Mathematics in Engineering,* 2019, pp. 526-538.

[20] Ş. E. Şeker, *İş Zekası ve Veri Madenciliği*, 1. baskı, İstanbul, Türkiye: Cinius Yayınları, 2013.

[21] WEKA. (2020, April 1). *Weka* [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/.

[22] M. Kaya and S. A. Özel, "Açık kaynak kodlu veri madenciliği yazılımlarının karşılaştırılması," *14. Akademik Bilişim Konferansı*, Mersin, Türkiye, 2014, pp.5-7.

[23] M. Turanlı, U. H. Özden and S. Türedi. (2020, October 21). *Avrupa Birliği'ne aday ve üye ülkelerin ekonomik benzerliklerinin kümeleme analiziyle incelenmesi.* [Online]. Available: http://acikerisim.ticaret.edu.tr/xmlui/handle/11467/891#sthash.tFw7f06H.dpbs.

[24] A. Tiwari and A. K. Sekhar, "Workflow based framework for life science informatics," *Computational Biology and Chemistry*, vol.31, no.5-6, pp.305-319, 2007.

[25] KNIME.(2020, April 1).*KNIME.* [Online]. Available: http://www.knime.org/.

[26] ORANGE.(2020, April 1).*ORANGE.* [Online]. Available: http://orange.biolab.si/.

[27] RAPIDMINER. (2020,April 1). *RAPIDMINER.* [Online].Available: http://www.rapidminer.com/.

[28] S. Kırışoğlu and A. Yakupoğlu , "Veri madenciliği ile üniversite bilişim teknik servis hizmetleri analizi", *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, vol. 8, no. 1, pp. 326-333, 2020.

[29] Wikipedia. (2020, April 1). *NaiveBayesClassifier.* [Online].Available: http://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[30] Wikipedia. (2020, April 1). *RandomForest.* [Online].Available: https://en.wikipedia.org/wiki/Randomforest.

[31] Towards Data Science. (2020, April 1).*RandomForestAlgorithm.* [Online].Available: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[32] J. R. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221-234, 1987.

[33] L. Mason, P. L. Bartlett and J. Baxter, "Direct optimization of margins improves generalization in combined classifiers," in*Proceedings of the 1998 Conference On Advances İn Neural İnformation Processing Systems*, 1999, pp. 288-294.

[34] D. Dua and E. Karra Taniskidou.(2017, December 30).*UCI Machine Learning Repository* [Online]. Available: http://archive.ics.uci.edu/ml.

[35] A. Gardner, R. R. Selmic, J. Kanno and C. A. Duncan. (2016, November 22). *UCI Machine Learning Repository* [Online]. Available: http://archive.ics.uci.edu/ml.

[36] E. Atagün and I. Duzdar Argun, "Performance analysis of data mining software with parametric changes," *International Journal of Forensic Software Engineering,* vol. 1, no. 2-3, pp. 115-143, 2020.

[37] M. A. Mahdi, S. E. Abdelrahman and R. Bahgat, "A high-performing similarity measure for categorical dataset with SF-tree clustering algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 496-509, 2018.

[38] A. Stetco, X. J. Zeng and J. Keane, "Fuzzy C-means++: Fuzzy C-means with effective seeding initialization," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7541-7548, 2015.

[39] M. A. Rahman and M. Z. Islam, "Application of a density based clustering technique on biomedical datasets," *Applied Soft Computing*, vol. 73, pp. 623-634, 2018.

[40] P. Eliasson and N. Rosen. (2020, December 21). *Efficient K-means clustering and the importance of seeding*. [Online]. Available:https://www.semanticscholar.org/paper/Efficient-K-means-clustering-and-the-importanceof-Eliasson Ros %C3%A9n/460178da1a4a7403a0a2db4cf52962ea7b4de29b.