

Analysis of Girls Vocational High School Students' Academic Failure Causes with Data Mining Techniques

Omer Deperlioglu¹ and Fatma Sibel Birtil²

¹*Afyon Kocatepe University, Vocational School of Afyon, Ali Cetinkaya Campus, Izmir Yolu 8 Km, 03200, Afyonkarahisar, Turkey*

²*Ministry of Education, Vocational School of Ali Cetinkaya, Department of Computer Programming, Afyonkarahisar, Turkey
Telephone: +90505 402 01 06, E-mail: fssert@gmail.com*

KEYWORDS Academic Success. Clustering Method. Data Mining. Reasons of Academic Failure. Vocational Secondary Education of Girls

ABSTRACT In this paper, factors that affect the failure of students at the Vocational High School for Girls were studied and a survey was performed to determine these factors. Since the collected data formed a large database, a data mining technique was used to more easily establish the correlations and rules. Based on content of the data and the subject of the study, the clustering method was selected as an evaluation tool. The collected data were separated into three clusters. Each cluster was investigated to determine the factors that play a role together in students' failure, as well as the relations between them. Finally, the findings of the applied survey, taking into account the work aimed at reducing the causes of failure, are expressed.

INTRODUCTION

Every individual in the society strives to be successful, but everyone cannot be equally successful. In one's life, an individual is designated successful or unsuccessful in the school environment, too. There are many different factors which affect student's failure. In the school environment, the concept of success takes the concept of academic achievement. Academic achievement is the average of the grades that student receive in their courses. If this average is low, it is referred to as academic failure. Academic failure is not only related to student effort and the courses themselves. There are many factors that could cause students to fail, such as cultural and socio-economic characteristics and the relationship between children, parents, and teachers (Schneider and Lee 1990; Gomes-Neto and Hanushek 1994; El-Hassan 1998; Birtil 2011).

Attaining academic success is more difficult nowadays compared to previous years because

students, especially those going through puberty, are more affected by their families, home and school environments, teachers' attitudes, and similar factors, resulting in academic failure (Yi et al. 2015; Urbanaviciute et al. 2015; Volodina et al. 2015; Heers et al. 2015). Many studies have been conducted to examine the problems of vocational education and some recommendations have been presented. But this is viewed from these studies for legislation and functioning that Vocational Education and Training system should be considered as a whole (Lin et al. 2015; Misbah et al. 2015; Rudenko 2015).

Data mining methods represent a valid approach for the extraction of precious information from existing Database. Educational Data Mining (EDM) deals with developing methods for discovering knowledge from data that come from educational domain (Bhawana and Bharti 2014).

This paper consists of eight chapters. As mentioned in the introduction, the concepts related to the research are provided. In the Method section, the issue of the data mining is presented. Data mining application areas, the studies on this issue, and the steps to be followed in the creation of data mining models are discussed. The clustering method of data mining used in this study is described in detail. In the Results section, the techniques developed and used in the implementation phase of a survey study for the Girls Vocational High School students are

Address for correspondence:

Omer Deperlioglu
Afyon Kocatepe University,
Vocational School of Afyon,
Ali Cetinkaya Campus,
Izmir Yolu 8 Km,
03200, Afyonkarahisar, Turkey
Telephone: +90532 397 79 09
E-mail: deperlioglu@gmail.com,
deperlioglu@aku.edu.tr

presented. In the Discussion and the Conclusion sections, the findings of the applied survey, taking into account the work aimed at reducing the causes of failure, are expressed. Then, the Recommendations and Limitations sections are given.

MATERIAL AND METHODS

Education can be defined as the process of changing or creating desired behaviors. To create a behavioral change in a person is not an easy thing. Considering the factors affecting education, the most important thing is to determine which factors affect students more. Thus, the quality of education can be improved by reducing or completely eliminating these factors. However, the data used in these assessment, are very excessive, because family, school and environmental factors prompt a lot of questions. To analyze these questions from obtained results by the students and assess results of the database will not be a very simple process. One of the methods that can be used for this application is data mining which is a method of creating a lot of meaningful data.

Data mining is a data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational research environment. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational databases for better understanding, improved educational

performance and assessment of the student learning process (Chan et al. 2008; Chandra 2010). Educational data mining is the application of data mining techniques to educational data, and so, its objective is to analyze these types of data in order to resolve educational research issues (Barnes 2009).

In today's technology, the data mining method is a conversion method which converts large amounts of data into information, saving time. Using data mining techniques reduces costs, increases revenues, improves productivity, uncovers new opportunities, makes new discoveries, and identifies scams.

Data Mining Process

Data mining is a method that can be used in many fields. This method is composed of a combination of stages which must be completed results. Data mining is a process that consisting of five phases as shown in Figure 1.

1. *Defining the Research Problem*: The first phase of the process, in the purpose of data mining operation, is fully defined.
2. *Data Preparation Phase*: The phase of the preparation of the data in its collection which consists of assembly and cleaning, and selecting and converting steps.
3. *Modeling Phase*: Data are collected and passed to the modeling stage after making sure they can be used. At this stage, the selected model is the most important issue to be applied.

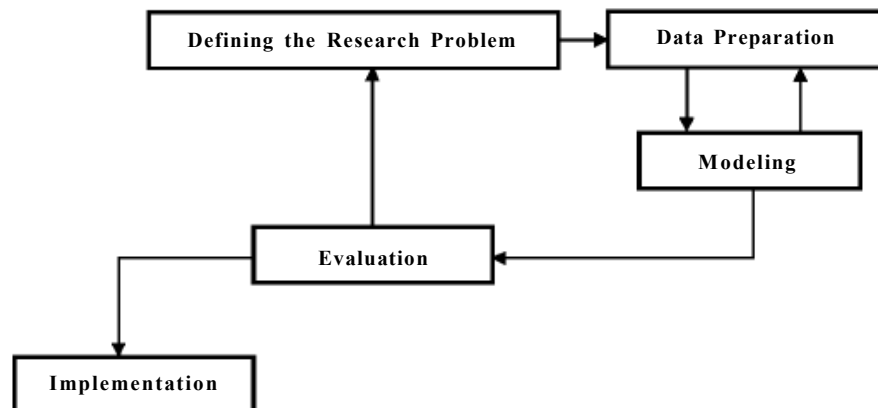


Fig. 1. Data mining process

Source: Author

4. *Evaluation Phase:* The results of the modelling stage are obtained. At this stage, obtained data, is assessed.
5. *Implementation Phases:* In this phase, the model is applied and the results are obtained. But the process phase has not been finalized yet because the model may need to be rearranged against changes.

Data Mining Models

In data mining, two main models used are descriptive and predictive. Descriptive models are the identification of patterns from the data. The predictive models moving from the known results are aimed at estimating the value of the results for unknown data sets (Han et al. 2006). There are commonly used statistical models in the descriptive and predictive methods: classification, regression, clustering, association rules, and artificial neural networks. The classification and regression models are the predictive models, whereas clustering, association rules, and neural networks are related to the descriptive models (Han et al. 2001).

Cluster Analysis

Cluster analysis is a separation process used to group data relevant. The purpose of clustering objects within the group, as distinct from objects in other groups to create themselves to be similar to the one another (Zaizane et al. 2002). Cluster analysis is a solution process consisting of several steps. Clustering a set of n objects into k groups is usually moved by the aim of identifying internally homogenous groups, according to a specific set of variables. In order to accomplish this objective, the most common starting point is computing a matrix called a dissimilarity matrix which contains information about the dissimilarity of the observed units. Thus, the data matrix is formed. Then the distance matrices are obtained with a suitable similarity measure. A suitable clustering technique is selected and applied. The objects will be allocated to the cluster by applying the technique. The final stage of analysis is the interpretation of the significance of the clustering results (Byran 1994).

FINDINGS

Data were collected through “the Survey of Causes for Failure” applied to students. A sample of “the Survey of Causes for Failure” was

obtained from RAM (the Guidance Research Center of the Ministry of Education). The survey consists of three parts. The first part contains questions to learn about the students. In the second part, students have questions about their perceived reasons for failure that consists of 37 questions. In the third section, there are 34 questions related to teacher attitudes, classroom and exams. The prepared survey was transferred to the web using Google Docs, that an application of The Google Inc.

Reliability Analysis

First, the survey’s reliability coefficient was calculated using the program SPSS 11.5 for Windows. Conducted to determine the internal consistency reliability of the scale after the first analysis, the total value of Cronbach’s alpha was calculated as $\alpha=0.8855$ for the entire test. Based on the negative impact on the alpha value of a total of 6 items (a1, a6, a7, a22, a23, b1) were removed from the survey and analysis was performed again. The total value of Cronbach’s alpha for the entire test after the second analysis was calculated as $\alpha = 0.8912$. A total of 3 items (a13, a18, a19) which negatively impacted the alpha value were removed from the survey and analysis was performed again. After the third analysis, the total value of Cronbach’s alpha for the entire test was calculated as $\alpha = 0.8918$. Two items in total had a negative impact on the alpha value (a17, a25) and were removed from the scale and an internal consistency analysis was repeated on the remaining components. Following the fourth analysis, the total Cronbach’s alpha value for the entire test was calculated as $\alpha = 0.8920$. On the fourth analysis no components were found to negatively impact the alpha value so internal consistency was reached.

As a result of the four-stage internal consistency analysis, the 71-point scale was reduced to 60 items, and it was concluded that the internal consistency of (the Cronbach’s Alpha) could not be upgraded from $\alpha=0.8920$. In this type of survey, an alpha reliability coefficient which is above 0.7 is deemed sufficient. The Survey of Causes for Failure results showed that alpha coefficient was 0.8920; thus it can be said it is a reliable questionnaire for this survey.

Data Cleaning, Selection and Conversion

Questionnaire data were obtained using Google Docs as an Excel spreadsheet. The data fields

irrelevant for the analysis were removed that is, name and surname, class, age, gender fields. The ages of all students were the same and all students in the school were girls. In addition, the full name and class field had no importance in this application. All the data entered by students were examined and many blank lines were found where students declined to respond. In this case, this student's data was removed from the registry database. The survey was completed by a total of 555 students, and about 542 of these pieces of data were used in the research students as seen Figure 2.

Modeling Phase

SPSS Clementine Client 10.1 software was used in the modeling phase of the study. SPSS Clementine is an integrated visual modeling tool developed for data mining applications.

After adjustment-related data was passed to the modeling stage, the modeling process was completed by a clustering algorithm. The process of grouping similar objects is called clustering.

Students' survey responses as to the causes of the failure of the survey in this study were divided into clusters. Thus, the supplied answers on the questionnaire will be detected which is desired and which is associated with each factor. Generally, two algorithms were used in the clustering process of the Clementine environment. These are K-means and K-medoids algorithms. The most common method used in the clustering method is a K-means method. K-means algorithm implementation is easy and less complexity compared to other clustering methods

(Han and Kamber 2001). The data was divided into clusters using the K-Means method as seen in Figure 2. Data on the results of the K-Means clustering process modeling phase was divided into three clusters. Other modeling done by changing the number of clusters was determined to be the logical consequence of three sets. Out of 542 total records, 289 were located in the first set, 88 in the second set, and 165 in the third set as seen in Figure 3.

RESULTS

Before detailed examination of the data sets, it was determined that all students marked the questions, which "I agree" with above 50 percent. The answer, which is "I agree" for given questions about causes of failure of all the students who participated in the survey were shown in Table 1.

Analyzing the responses of the students to item 4, the students themselves related reasons that include:

"My attention very quickly falling apart", which applied to 78.78 percent of student. "I have basic lack of information" which pointed to 74.35 percent of the students, and "I do not want to study the courses that I will fail", which was marked by 69.37 percent of the students. Then, other items are "I get too nervous during the exam", "I deal with extracurricular subjects", "I do not give myself a kind of study", respectively.

In part b of the survey, reasons for failure related to exams, classroom environment, and teacher attitudes examined the item "Exams are done in such a way that in the succession " was

Table 1: The questions that over the 50% of all students marked "I Agree" answer in the survey

Questions	Value %
1 Exams are done in such a way that in the succession. (b24)	84.32
2 Some courses and subjects are very heavy, such as I'm having trouble understanding. (b16)	79.52
3 My attention very quickly falling apart. (a4)	78.78
4 Difficult courses in the curricula of overlap. (b10)	76.38
5 I have basic lack of information. (a14)	74.35
6 I cannot succeed in the courses I was not interested. (b15)	70.11
7 I do not want to study the courses that I will fail. (a12)	69.37
8 I get too nervous during the exam. (b18)	65.87
9 Most of the time I cannot take notes that I expected. (b33)	62.18
10 I deal with extracurricular subjects. (a16)	60.15
11 There's no one at home to help my lessons. (a32)	58.3
12 I do not give myself a kind of study. (a3)	56.09
13 I cannot be successful in the course owing to dislike of teachers (b27)	53.32
14 The teacher offers lessons in a boring and monotonous way. (b28)	51.66
15 Difficult questions being asked in the exam. (b19)	51.29

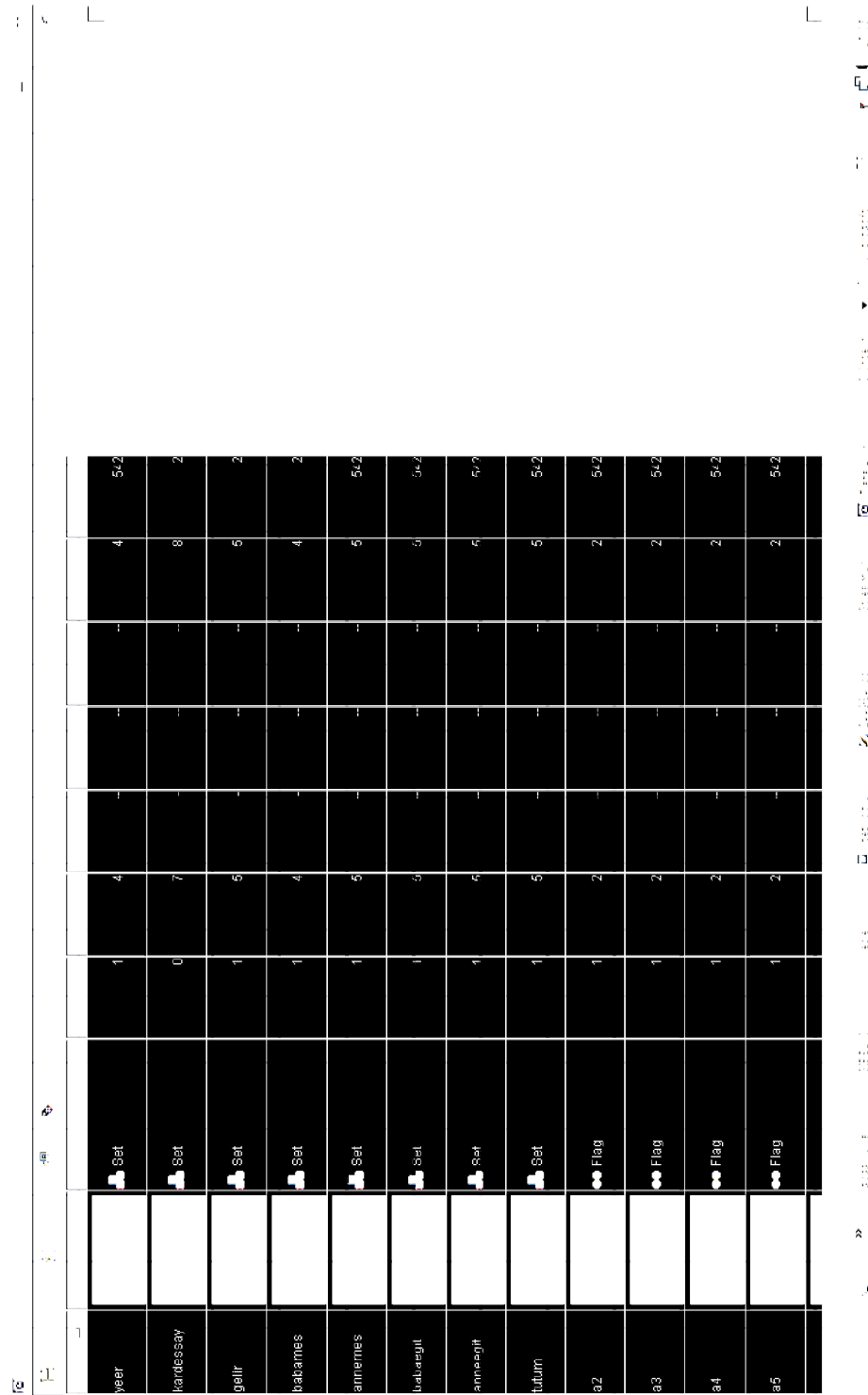


Fig. 2. The data clustering with K-Means method
Source: Author

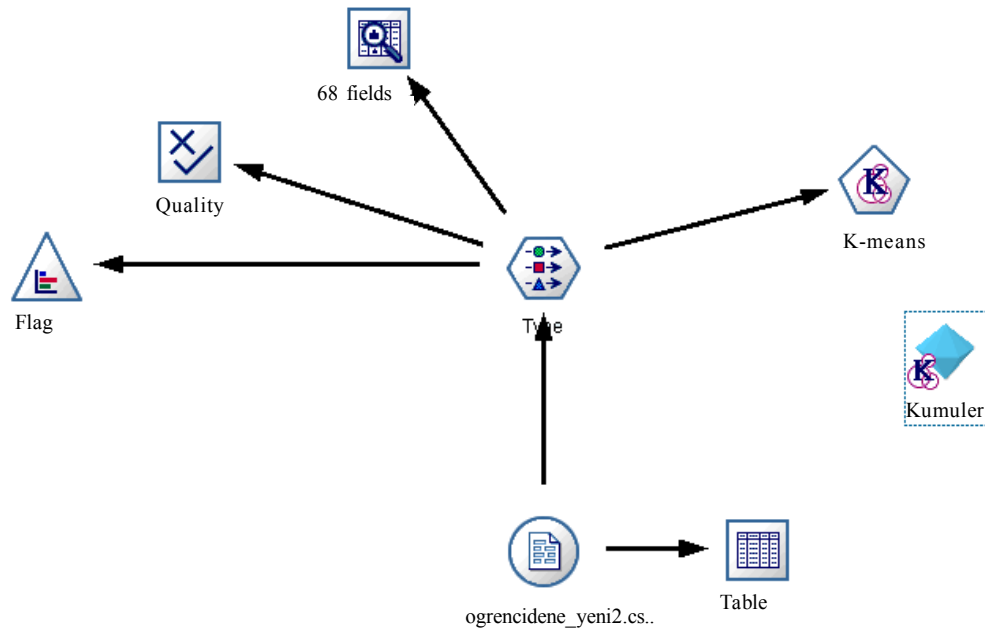


Fig. 3. Three sets in this study

Source: Author

marked by the 84.32 percent of the students. Second, “Some courses and subjects are very heavy, such as I’m having trouble understanding “in the item have pointed to 79.52 percent of the students. Thirdly, “Difficult courses in the curricula of overlap” item is marked by 76.38 percent of the students. Then, other items are “I don’t like this school”, “I get too nervous during the exam”, “Most of the time I cannot take notes that I expected”, “I cannot be successful in the course owing to dislike of teachers”, “The teacher offers lessons in a boring and monotonous way “ and “Difficult questions are being asked in the exam”, respectively.

When modeling process results obtained in three sets were analyzed, the students in the first set, selected 9 out of 60 causes of failure. The second cluster of students stated that Article 41 makes them fail. Students in the third cluster selected 26 items affecting their failure. Noteworthy, was that the students in the first cluster, although the reasons are not faced but a lot of these reasons push them to fail. The students in the third set say that the amount of material makes them fail and the students in the second set, of more than half of the questions in the survey argued that it failed them.

In the first set, there were 53.3 percent of the students. The questions that the students in this set had answered as “I agree” were answered the same way the students in the other two sets. This cluster consists of students who have answered “I agree” for 9 items. These students were able to better evaluate the causes of their failure and they were able to select 9 items from 60 items. Students in this set are assumed to be forced to adapt to multiple topics within the same day because students have stated that the courses and exams are stacked back to back which negatively affects their performance.

There were 16.3 percent (88 students) of the group surveyed in the second set. Students in the second set, have answered 41 “I agree” from 60 items for the cause of their failure. Analyzing the responses of students in the second set, an item that was not found in the other two groups of students was found to have marked often. These students have stated that there were problems with themselves. Students in this set have said that they had a problem, but they could not tell anyone their problem. In addition, they felt isolated and stated that they find themselves flawed. It is observed that these students had

communication problems. It is believed they cannot share their problems with their families and their responses to questions about their families were investigated because they feel isolated. These students believe that their efforts are not appreciated enough. They said that their families always compare them with those more successful. They also complain that their family forces them to the study continuously. In addition, students in this set gave the answer "I agree" to 24 of the 34 items related to school.

There are 30.4 percent (165 students) of the group surveyed in the second set. Students in the second set have given 26 "I agree" answer from 60 items for the cause of the failure. Students in the third cluster are a group of students with problems associated with the school. These students have commonly marked the item of "The teacher offers lessons in a boring and monotonous way" and they don't like these teachers. Also they have commonly marked the items of "I cannot be successful in the course owing to dislike of teachers." and "I do not want to study the courses that I will fail".

DISCUSSION

This study aims to determine the factors that cause students to fail. In this way, efforts can be made to eliminate the factors that lead to failure. For the evaluation of the survey clustering method was selected among possible data mining methods. Data mining is more than a simple transformation of data using technology developed from databases, statistics, and machine-learning. Instead, data mining involves integration, rather than a simple transformation, of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis (Ma et al. 2000; Kotsiantis et al. 2004; Khan 2005; Han and Kamber 2006; Hijazi and Naqvi 2006).

Our results found that attaining academic success is more difficult nowadays compared to previous years. This is because students, especially those going through puberty, are more affected by their families, home and school environments, teachers' attitudes, and similar factors, resulting in academic failure as other studies in literature (Yi et al. 2015; Urbanaviciute et al. 2015; Volodina et al. 2015; Heers et al. 2015).

CONCLUSION

In this paper, the factors that affect the failure of students were studied and a survey was performed to determine these factors. Since the collected data formed a large database, a data mining technique was used to more easily establish the correlations. In evaluation of the survey, clustering method was selected, among possible data mining methods by considering the content of the data and the subject of the study. The collected data is separated into three clusters; each cluster is investigated to determine the factors that play a role together in students' failure, as well as the relations between them. It is determined that the students in the first cluster could not adapt to more than one issue in the same day. In the second cluster, students had stated that there were some problems that were relevant to them. Students in this cluster have had some problems that they cannot discuss it with anyone. Students within the third cluster were a group of students who had problems in general with the school.

RECOMMENDATIONS

Some seminars about paying attention, defeating the nervousness at the exam, and overcoming the feeling of failure can be given to the students in the first set. If these seminars included all students, they would be more useful because all the students in all sets selected these 9 items. To be scattered difficult courses in a lesson program can be discussed with the school administration. Also, the school administration cannot put difficult courses in a row in the lesson program and reduce the number of examination to be held in one day. This will provide relief for students in the first set.

Because the students in the second set have personal problems with themselves, it would be appropriate to conduct individual interviews with students. Also, it is seen that they had continuously talked about their family as a reason of their failure. So their families should be contacted and spoken too. If the communication problem is solved for the students who thought they had communication problems with their families, it can help solve other problems. In this set, the students' failure can be tied to several reasons implies that they are quite confused.

Students in the third set see teachers as the most important reason for the failure of students. Therefore, various seminars about good com-

munication with students, preparation and evaluation of exam questions and creating lectures with new techniques can be given to teachers. In this way, that student achievement levels would increase. Also taking into consideration the fact that the students stated that they did not like school; in order to attract the interest of students, if the school organized sporting or cultural activities and excursions. Changes may be made in the school environment that would attract the attention of students. It would help to get student inputs, before undertaking these activities, as it will make them feel important.

LIMITATIONS

This study was only conducted for the girls' vocational school in Afyonkarahisar in Turkey. The same work can be made in other regions of Turkey and even in other countries in Europe and similarities, and differences can be examined. In addition, the researchers have done work for the general students' academic failure. Studies for each cluster based on the results obtained in this study can be made separately. Thus, the results of this study can be examined in detail.

In this study, standards measurement survey that was prepared by the ministry of Turkish national education was used. By expanding the survey, different properties of the students such as cultural, physical and biological can be added. Thus, these features may also be evaluated.

REFERENCES

- Barnes T, Desmarais M, Romero C, Ventura S 2009. Educational Data Mining 2009. *2nd International Conference on Educational Data Mining, Proceedings*. Cordoba, Spain, 1-3 July.
- Birtil FS 2011. *Analysis of Girls Vocational High School Students' Academic Failure Causes With Data Mining Techniques*. Master's Thesis. Afyonkarahisar: Afyon Kocatepe University.
- Bhawana A, Bharti G 2014. Review on data mining techniques used for educational system. *International Journal of Emerging Technology and Advanced Engineering*, 4(11): 325-329.
- Byran FJ 1994. *Mainly, Multivariate Statistical Methods*. 2nd Edition. London: Chapman – Hall.
- Chan AYK, Chow KO, Cheung KS 2007-2008. Online course refinement through association rule mining. *Journal of Educational Technology Systems*, 36(4): 433-444.
- Chandra E, Nandhini K 2010. Knowledge mining from student data. *European Journal of Scientific Research*, 47(1): 156-163.
- El-Hassan K 1998. Relation of academic history and demographic variables to grade retention in Lebanon. *The Journal of Educational Research*, 91(5): 279-288.
- Gomes-Neto JB, Hanushek EA 1994. Causes and consequences of grade repetition: Evidence from Brazil. *Economic Development and Cultural Change*, 43(1): 117-148.
- Han J, Kamber M 2006. *Data Mining Concepts and Techniques*. Waltham, USA: Morgan Kauffmann Publishers Inc.
- Han J, Kamber M, Tung AK 2001. Spatial clustering methods in data mining: A survey. In: H Miller, J Han (Eds.): *Geographic Data Mining and Knowledge Discovery*. London: Taylor and Francis, 201-231.
- Heers M, Klaveren CV, Groot W et al. 2014. The impact of community schools on student dropout in pre-vocational education. *Economics of Education Review*, 41: 105-119.
- Hijazi ST, Naqvi RS 2006. Factors affecting student's performance: A case of private colleges. *Bangladesh e-Journal of Sociology*, 3(1): 1-10.
- Khan ZN 2005. Scholastic achievement of higher secondary students in science stream. *Journal of Social Sciences*, 1(2): 84-87.
- Kotsiantis S, Pierrakeas C, Pintelas P 2004. Prediction of student's performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5): 411-426.
- Lin MH, Hu J, Tseng ML, Chiu AS et al. 2016. Sustainable development in technological and vocational higher education: Balanced scorecard measures with uncertainty. *Journal of Cleaner Production*, doi: 10.1016/j.jclepro. 2015.12.054.
- Ma Y, Liu B, Wong CK et al. 2000. Targeting the Right Students Using Data Mining. *Proceedings of KDD, International Conference on Knowledge discovery and Data Mining*, Boston, USA, August 20-23, pp. 457-464.
- Misbah Z, Gulikers J, Maulana R et al. 2015. Teacher interpersonal behaviour and student motivation in competence-based vocational education: Evidence from Indonesia. *Teaching and Teacher Education*, 50: 79-89.
- Rudenkoa D, Morosovab E 2015. Prospects for the development of further vocational education in the Tyumen Region of Russia. *Procedia - Social and Behavioral Sciences*, 214: 693-699.
- Schneider B, Lee Y 1990. A model for academic success: The school and home environment of East Asian Students. *Anthropology and Education Quarterly*, 21(4): 358-377.
- Urbanaviciute L, Pociute B, Kairys A et al. 2016. Perceived career barriers and vocational outcomes among university undergraduates: Exploring mediation and moderation effects. *Journal of Vocational Behavior*, 92: 12-21.
- Volodina A, Nagy G, Koller O 2015. Success in the first phase of the vocational career: The role of cognitive and scholastic abilities, personality factors, and vocational interests. *Journal of Vocational Behavior*, 91: 11-22.
- Yi H, Zhang L, Yao Y et al. 2015. Scott Rozelle Exploring the dropout rates and causes of dropout in upper-secondary technical and vocational education and training (TVET) schools in China. *International Journal of Educational Development*, 42: 115-123.
- Zaiane OR, Foss A, Lee CH et al. 2002. Data Clustering Analysis: Scalability, Constraints and Validation. *Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Taipei, Taiwan. May 2002, pp. 28-39.

Paper received for publication on April 2015

Paper accepted for publication on February 2016