Mining of Completion Rate of Higher Education Based on Fuzzy Feature Selection Model and Machine Learning Techniques

Tahseen A. Wotaifi, Eman S. Al-Shamery

***Abstract*: *In the context of the great change in the labor market and the higher education sector, great attention is given to individuals with an academic degree or the so-called graduates class. However, each educational institution has a different approach towards students who wish to complete their university degree. This study aims at (1) identifying the most important factors that directly affect the completion, and (2) predicting the completion rates of students for university degrees according to the system of higher education in the United States. Unlike previous studies, this project contributes to the use of the fuzzy logic technique on three methods for feature selection, namely the Correlation Attribute Evaluation, Relief Attribute Evaluation, and Gain Ratio Method. Since these three methods give different weight to the same attribute, the fuzzy logic technique has been used to get one weight for the attribute. A great challenge faced throughout this study is the curse of dimensionality, because the college scorecard dataset launched by the US Department of Education contains approximately (8000) educational institutions and (1825) features. Applying the method used in this study to identify important features lead to their reduction to only (79). Accordingly, two models have been used to predict the completion rates of students for their university studies which are the Random Forest and the Support Vector Regression with a Mean Absolute Error (MAE) value of (0.068) and (0.097) respectively*.

***Keywords* : *Completion Prediction of Students, Fuzzy-Selection Method, Filter Method, Mining Higher Education, Random Forest, and Support Vector Regression*.**

I. INTRODUCTION

Students are the main stakeholders in educational institutions and their performance plays an important role in the economic and social aspect of the country through their experience and skills of study[1]. Before entering the institution, prospective students should know about the most important problems and obstacles that will be faced during the course of the study, as these factors may force them to dropout from the educational institution. In other words, prospective students should know the pros and cons of any educational institution before entering it [20].

The nature of the higher education systems plays a critical role in producing an appropriate context for educational institutions resulting in a great competition among these universities, for instance. In this project, the higher education system in the United States, being one of the world's most important educational systems. Apart from the fact that a great majority of the people in the US have an interest in education, many other elements made the educational system in the United States so important, including its large number of researchers and scientists, the freedom of expression, competent colleges and faculties, equality of chance, and non-discrimination among individuals in terms of religion, gender, and race [18][19]. The United States Department of Education has introduced excellent support that provides insight into students and their families by creating a college scorecard dataset in September 2015[2]. This dataset contains detailed information about educational institutions in the US themselves as well as former students who have entered these institutions and many other factors including costs, fees, financial aid, demographics of students and others [3][21]. The identification of the factors that help students complete their study or, on the contrary, force them to dropout, are considered as a recommendation system for prospective students [4].

Since higher education in the US is a good investment and students are always encouraged by their families as well as the university itself to complete the academic certificate [5], Educational Data Mining (EDM) has recently emerged [6]. Data mining techniques have been applied in many fields such as medicine, business, and markets and have achieved great successes. Recently, a clear indication of the direction of researchers towards the educational database has appeared. In other words, in order to analyze the orientation of students towards education and thus achieve high success rates for students and the educational institution itself, research interests in the use of data mining techniques in education have doubled. This emerging field is primarily concerned with students and the acknowledgement of the essential factors that affect their performance either positively or negatively through the analysis of educational databases [7].

Previous studies have merely focused on student assessment at a specific time such as the first year of the student after enrollment, whereas this study aims to analyze all educational institutions in the US comprehensively and identify the most important factors faced by students in

Revised Manuscript Received on September 25, 2019

Tahseen A. Wotaifi , College of Information Technology, University of Babylon, Hillah, Babil, Iraq Email: tahseen.ubabylon@gmail.com

Eman S. Al-Shamery, College of Information Technology, University of Babylon, Hillah, Babil, Iraq Email: emanalshamery@itnet.uobabylon.edu.iq

each educational institution. Through this analysis, prospective students and their families come to learn about educational institutions and know the advantages and disadvantages of each institution before entering it [8].

Outline of paper:

Section II contains the related work. Section III reviews the theoretical background. Section IV explains the methodology of research including data, preprocessing, and the developed fuzzy filter method. Section V illustrates the results. Section VI contains the conclusions. Finally, section VII reviews the references.

II. RELATED WORK

The predictability of students' completion of their studies has got many researchers' concerns and contributions and has taken extensive strides in the use of data mining techniques, yet few have used comprehensive educational databases such as college scorecard dataset. Most studies have been performed based on the individual characteristics of graduates, such as family and personal background, individual major, and others [1][2]. Therefore, one of the priorities of this study is to focus on educational institutions including small and less elite ones as well as the characteristics of individual students.

[Ali Daud a, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D. Lytras, Farhat Abbas, Jalal S. Alowibdi., (2017)], applied a classification model to predict whether or not a student will be able to complete his degree or not and found that the family's expenditure, income, and assents, as well as the students' personal information are the most influential factors regarding their goal. In scientific terms, the mechanism of work is interesting but, as mentioned earlier, the shortcomings of most research in this field are due to their dependence on the personal factors of students.

[Peter Shea, Temi Bidjerano., (2014)], introduced important models about students who participated in online and distance education. Through this study, they have concluded that these students not better academically prepared and were, in fact, possibly somewhat less academically prepared and/or less likely to graduate than students who actually attended the educational program.

[Abeer Badr El-Din Ahmed, Ibrahim Sayed Elaraby., (2014)], used one of the classification techniques (decision tree) to predict the final grade for students (excellent, very good, good, acceptable, and fail). Although this study provided a good model for determining student success or student failure, it would have been better to address failure frequency rather than to determine failure itself.

[Zahrah Alharbi, James Cornford, Liam Dolder, and Beatriz De La Iglesia., (2016)], attempted to use data mining techniques to discover student performance problems and then identify students at risk of poor performance. This study has contributed to the building of a recommendation system for students but its disadvantage is the lack of data used.

[Tolga Demirhan and Ilker Hacioglu., (2017)]. In this study, the dataset has been collected through the voluntary participation of students at Trakya University, Tunca Vocational School (Distance Education) in a questionnaire. This dataset has been analyzed and data mining techniques have then been applied to identify important factors and predict the success or failure of students.

[Davis Jenkins., (2011)] provided a method or system of recommendation to the community colleges to measure the completion rates of students for their program, and a set of suggestions has been submitted to the administrations of these colleges. The first shortcoming is that this paper is restricted to only a certain group of students, namely the students of community colleges. On the other hand, judging most students who wish to enter these colleges and considering them students who do not have goals is not appropriate.

[Al-Barrak, Mashael et al.,2016], used educational data mining to examine a dataset at educational institutions or universities. This study has applied educational data mining methods to predict students' final GPA (poor, average, good, very good, and excellent) depending on the grades in earlier courses. A number of classification rules have been created by using the J48 decision tree model.

III. THEORETICAL BACKGROUND

1- Higher Education Dataset.

Rapid advances in information technology have resulted in a significant increase in educational databases. These huge educational databases contain a wide range of valuable information about former students as well as educational institutions[7]. In order to extract this valuable information, a new research field has emerged for this purpose, called EDM which provides a detailed analysis of student databases. In other words, it extracts hidden patterns in student databases in order to understand students' attitudes toward education [8].

2- Feature Selection.

Feature selection is the process of selecting the most important set (the best subset) of attributes from original attributes[9]. Because some features are weakly relevant, redundant or irrelevant, the use of the feature selection method plays an active role in this field by identifying the factors that are strongly relevant to the target and thus improving the accuracy of the model[10]. The features which are strongly relevant to the target are identified through three broad-use strategies: the filter method, embedded method and wrapper method. Although all these methods are used to reduce the data dimensionality, the mechanism is different for each method. Filter techniques will be clarified as they have been applied to this project.

2.1 Filter methods

These methods select features based on a performance measure regardless of the prediction algorithm. That is, they are always used before the prediction model [11]. There are many types of filter methods such as correlation attribute evaluation, relief attribute evaluation, information gain, gain ratio, and others [12].

- Gain ratio method: this is an information gain method modified by the gain ratio method to reduce the bias on high-branch features, so this method focuses on the size and number of branches when choosing the feature. This is achieved through the following

equations:

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

$$GainRATIO_{split} = \frac{GAIN_{split}}{SplitINFO}$$

- Correlation method: the correlation method gives weight to any attribute between (1) and (-1) according to the correlation between this feature and the target (it evaluates each feature by its individual predictive ability). This is achieved through the following equation:

$$cor(x,y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

- Relief method: This method takes the instance in the account when evaluating the attribute. In the classification, this technique randomly selects the two nearest neighbors (nearest miss and hit) and then evaluate the attribute while in regression. This method depends on the principle of probability.

3- Data Mining Techniques

3.1 Random Forest.

Random forest is a collective learning method which builds a large number of decision trees. It is highly suitable for large data and can be used as classification or regression models, which implies that these models predict continuous or discrete values[13]. Random forest model is characterized by (1) the way to build of a group of individual trees, (2) the procedure which is used to generate and modify these individual trees, and (3) the way the predictions of each individual tree are combined to produce more unique and consistent predictions [14].

3.2 Support Vector Regression.

In regression problems there is a Support Vector Machine (SVM) named Support Vector Regression (SVR). It is a supervised learning method that is characterized by the usage of kernels where it can handle non-linear prediction with high efficiency through a nonlinear kernel function, the absence of local minima, the sparseness of the solution and the capacity control obtained by acting on the margin, or on a number of support vectors [15]. One of the most important strengths is that it is used to create classification models or methods of regression as well as achieve important results with large datasets [16].

IV. RESEARCH METHODOLOGY

1- Dataset (college Scorecard Dataset).

In September 2015, the United States Department of Higher Education contributed a significant step in serving a large segment of people, especially educational ones by creating college scorecard dataset[2]. The United States Department of Higher Education monitored its educational institutions from 1996 to 2015 and then released this dataset[13]. The College Scorecard dataset is very huge where it includes many factors for educational institutions, as well as their former students such as the demographics of students in each college or educational institution, cost of study, educational expenses, number of students (by gender, ethnicity, and color in each educational institution), financial aid (e.g. PALL grant and loans), the SAT and ACT scores in each institution, and many others. Since this dataset is very huge, it has been divided into nine categories which are the categories of financial aid, student, costs, admission, repayment, school, completion, academics, and the earnings. Each of these categories contains a number of features about educational institutions in the United States and their former students. Many of the details of this dataset are given in Table (1) below.

Table1: number of features & description

Category	No. of features	Description
Cost	52	Includes study costs and fees
earning	70	This category includes information on family income and the earnings of graduates
financial aid	40	Includes loans and grants offered for students
Completion	1013	Includes US Treasury Department information as well as students completion rates within four, six and eight years
Academics	228	Information about the type of academic program available in the educational institution
Admission	25	Includes admissions rates and SAT/ACT scores
Repayment	131	Includes students' repayment and default rates
School	170	Important information about universities
Student	96	Includes demographics of the student's body

2- Data-preprocessing.

There are two main reasons for using data preprocessing: 1) to reduce the dimensionality of data in order to achieve efficient analysis, and 2) for adaptation of data to best fit more with the analysis method[17]. In order to better understand the dataset which has been used in this study as well as processing the missing values, data-preprocessing has been performed.

Generally, four steps have been performed in data preprocessing: data cleaning. Handling missing values, data normalization, and numeric to nominal.

2.1 Data cleaning step:

According to this step, the features have been removed

which conform to the following conditions

- attributes that contain a single value in all universities (instances).
- features that are of no use in prediction (e.g. ID number).
- attributes that have more than 50% of their entries as "NULL" or "PrivacySuppressed" values.

2.2 Handling missing values step:

Since the dataset used in this study is very large and many values in the dataset have been listed as "NULL" and "PrivacySuppressed", the preprocessing of data have been performed (NULL means that the value is missing), whereas (PrivacySuppressed refers to the fact that most of these values are found in small educational institutions that have fewer than 30 students). One of the easiest ways to deal with these values is to remove any feature with PrivacySuppressed entries or NULL entries. However, this method is not desirable with college scorecard dataset, so the missing values have been processed using mode and mean method. Each method has been applied based on the type of feature.

2.3 Normalization step:

This step is performed in order to avoid features with large values that control the results as well as to normalize all feature values into a range between 0 and 1 using the min-max normalization method.

For more details, see Algorithm (1) which summarizes what has been mentioned above.

Algorithm 1: Data preprocessing

// input and output

Input: Array of attribute (D_{ij}) where i: number of instances and j number of attributes.

Output: Relevant features

// data cleaning step

Begin

1 for i = 1 to n do

2 for j = 1 to m do

3 if ((all feature values are equal) or(missing value>= 0.5))

4 remove the feature from D_{ij}

5 end if

6 end for

7 end for

// handling missing values

8 for i = 1 to n do

9 for j = 1 to m do

10 if attribute value v is missing

11 if all attribute values in the feature are different

12 v = mean

13 else

14 v = mode

15 end if

16 end for

17 end for

// data normalization

18 for i = 1 to n do

19 set max and min to zero

20 for j = 1 to m do

21 calculate min and max in feature j

22 $\hat{v} = \frac{v - min_j}{max_j - min_j}$

23 $DP_{ij} = \hat{v}$

24 end for

25 end for

26 return DP_{ij}

End

3- Fuzzy-selection method.

The variable response in this study (target) is the completion rates for students who completed within 150 percent of the expected time. In order to understand a college scorecard dataset that contains a large number of features, three techniques of feature selection have been used: Correlation Attribute Evaluation, Relief Attribute Evaluation, and Gain Ratio Method. The key role of these methods is to reduce the dimensions of the data and to identify the features that are relevant to the target. Since each of these methods evaluates the same attribute with a different weight, the fuzzy logic technique has been used to obtain one weight for each attribute.

Overall, the irrelevant features have been reduced in three stages:

- Separately, all weak features estimated by the three techniques mentioned earlier have been removed and as follows: 1) according to the Relief Attribute Evaluation method, which gives weight to each feature between (1) and (-1), all attributes weighing less than zero have been deleted, 2) according to the Correlation Attribute Evaluation technique, any feature with zero correlation with the target has been deleted, 3) according to Gain Ratio Method, any feature with weight equal to zero is deleted.
- It is natural that the features chosen by the three techniques are better than the attributes neglected by one of the methods, so the features that have been selected are those that are not neglected by any of the three methods.
- Finally, the fuzzy logic has been applied in order to have one weight for each attribute, as each of the three filter methods evaluates the same attribute with a different weight or rank. After performing the fuzzy logic technique, any feature having a weight less than a threshold has been removed.

Figure 1 below illustrates the process of selecting the most important features (factors) affecting the completion rates of students according to the fuzzy-selection method.

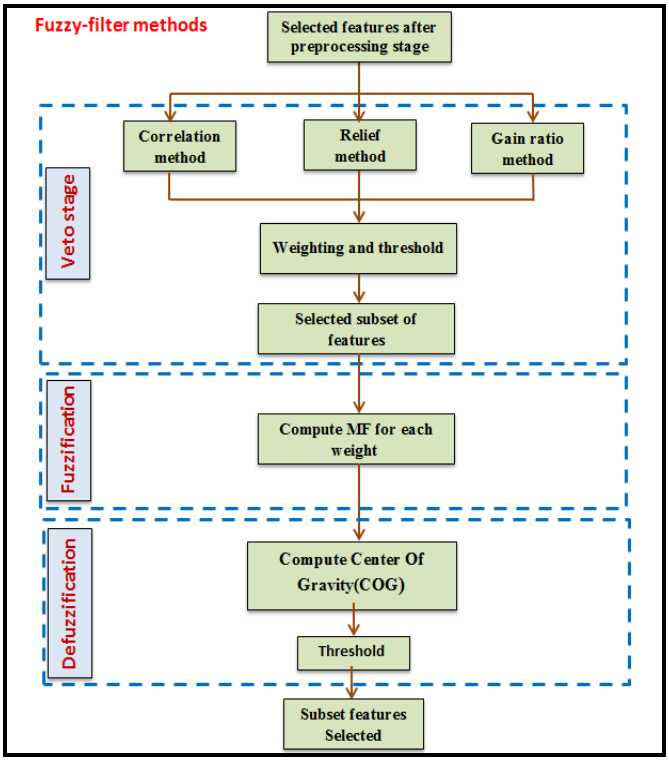


Figure 1: FSM for completion

In Fuzzification, the Membership Function (MF) for the three weights obtained by the filter methods above has been computed through the application of the trapezoidal shape.

Figure 2 below explains the process of calculating the membership function for each of crisp values.

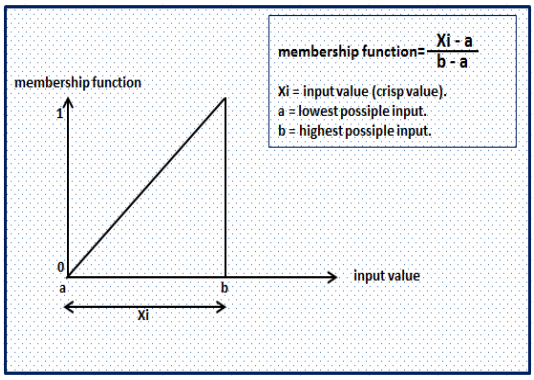


Figure 2: fuzzy logic to identify significant features

One weight has been obtained for each feature using defuzzification. This is achieved using the Center Of Gravity method (COG) according to the following equation.

$$\mu o = \frac{\sum_{i=1}^{n} \mu(Xi) * Xi}{\sum_{i=1}^{n} \mu(Xi)}$$

Where:
μ(Xi): is the membership function of the crisp value and Xi: is the crisp value (weight of feature).

The goal of this study is to reduce the number of factors to as least as possible while maintaining high accuracy. For this purpose, any feature having a weight less than (0.5) has been neglected. After this step, the number of remaining features represents the most important factors affecting students' completion. These factors will be clarified in the subsequent sections and it is hoped that they will be taken into account by prospective students. Algorithm (2) illustrates the algorithmic descriptive:

Algorithm 2: Fuzzy-selection methods

// input and output
Input: The output of data preprocessing algorithm..
Output: the significant features
// RRelief method
Begin
1 for f = 1 to n //where: n is the number of attributes
2 calculate weight (w) of feature (F) according to RRelief method
3 if w <Θ // Θ = 0
4 delete (f) from the dataset
5 else
6 RF[f] = F
7 end if
8 end for
// correlation method
9 for f = 1 to n //where: n is the number of attributes
10 calculate weight (w) of feature (F) according to correlation method.
11 if w=Θ // Θ = 0
12 remove f from the dataset
13 else
14 CO[f] = F
15 end if
16 end for
// Gain ratio method
17 for f = 1 to n //where: n is the number of features
18 calculate weight (w) of feature (F) according to gain ratio method.
18 if w = Θ // Θ=0
19 remove f from the dataset
20 else
21 GR(f) = F
22 end if
23 end for
// veto
24 for f = 1 to n
25 if (F ∉ RF[]) or F ∉ CO[f] or F ∉ GR[f] then
26 remove attribute F from the dataset
27 end if
28 end for
// Fuzzy logic
29 for i = 1 to n // where: n is number of features
30 for f = 1 to m do // where: m is weight in three methods.
// compute MF
31 $MF = \frac{w_f - a}{b - a}$
32 end for
// calculate the center of gravity
33 $COG = \frac{\sum_{i=1}^{n} \mu(w_i) * w_i}{\sum_{i=1}^{n} \mu(w_i)}$
34 if COG < Θ // Θ < 0.5

```
35      remove attribute F from dataset
36   else
37   SF = COG
38  end for
39 return SF
End
```

After closely studying the characteristics of prediction techniques and finding the most appropriate methods for our data, two different approaches have been used to know whether the factors chosen were the most likely to affect students' completion rates. In the first technique, this study sought to represent the data in form of a group of trees through random forest technique, while the second technique is support vector regression.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The main objective of this study is to identify the most important factors affecting the completion of students from a large database and eventually to reduce these factors to the lowest possible with high accuracy as possible as. In order to better understand this dataset, a number of steps have been taken in this study to identify features that are strongly relevant to the target and to exclude weak relevant features.

Table3: Error across the two models

Models	*MAE*	*RMSE*
Random forest	*0.068*	*0.097*
SV Regression	*0.072*	*0.10*

The first reduction has been performed through data pre-processing, and the number of features remaining after this process is (912) features. The second reduction has been done by fuzzy-selection method for feature selection. This has been accomplished by using the fuzzy logic technique on filter methods to provide one weight for each feature. Then any feature with a weight less than 0.5 has been discarded. After this process, only (79) features remained. Table 2 shows the top 10 factors yielded after data preprocessing and the fuzzy-selection method.

Table 2: the top 10 factors

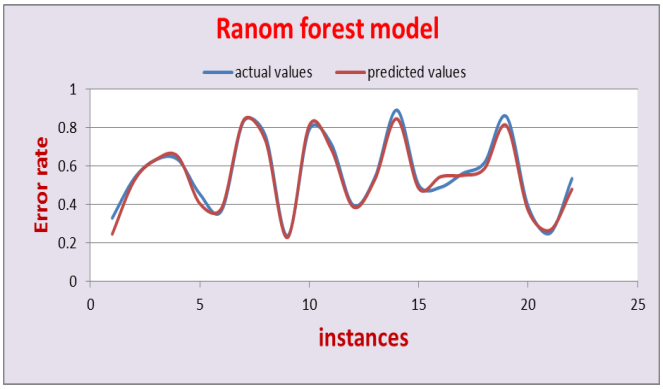
Category	Weight	Description
IND_INC_PCT_M1	*0.767*	*% independent students who their income between ($30,001-$48,000)*
IND_RPY_3YR_RT	*0.755*	*% independent students who are not in default on their federal loans for three years*
INC_PCT_M2	*0.718*	*% independent students who their families income between $48,001-$75,000*
APPL_SCH_PCT_GE3	*0.718*	*% students who send their FAFSA report to at least three institutions*
ACTMTMID	*0.711*	*ACT midpoint for mathematics*
HI_INC_RPY_7YR_RT	*0.710*	*% students who are not in default on their federal loans and*
APPL_SCH_PCT_GE2	*0.710*	*% students who send their FAFSA report to at least two institutions*
CDR3	*0.708*	*Rates of default on repayment their families income greater than $75,000*
APPL_SCH_PCT_GE4	*0.700*	*% students who send their FAFSA report to at least four institutions*
ADM_RATE_ALL	*0.698*	*Admission rates in the educational institution*

Given the advantages of data mining algorithms, it has been found that both of the random forests techniques that succeed in dealing with huge data as well as the support vector regression technique are the closest techniques to our data.

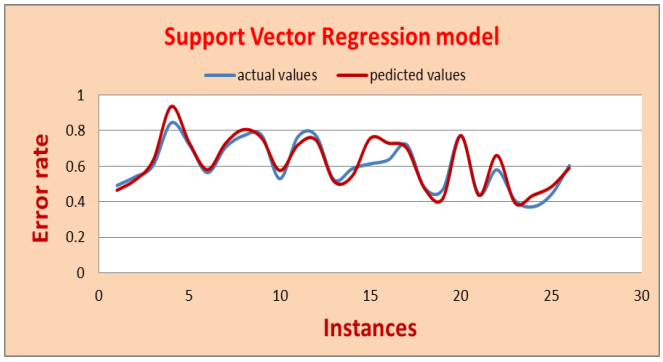
The 10-fold cross-validation scheme has been implemented (i.e. 9-fold for the training set, 1-fold for the testing set, 10 rounds in total). Depending on both the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), it has been found that the Random Forest Algorithm was slightly better than the Support Vector Regression algorithm. Table 3 below shows the amount of error for both models.

Compared to previous studies that used a simple dataset and did not have many features, this study has used a huge dataset (college scorecard dataset) containing thousands of features, and these features have been greatly reduced with a small error ratio.

The difference between the predicted and the actual values of the Random Forest model and Support Vector Regression model is shown visually in Figures (2) and (3).



Figures 3: Error for completion with random forest.



Figures 4: Error for completion with SVR

Additionally, this project presents the methods of academic prediction on a huge dataset recently launched by the US Department of Education, called College Scorecard dataset. In order to facilitate the understanding of these large data sets which contain hundreds or thousands of features to prospective students and their families, this study aims to identify the most important factors that affect the completion rates of students who are expected to complete their degree within six years. For this purpose, three attribute selection methods have been applied, including Relief Attribute Evaluation, Correlation Attribute Evaluation, and Gain Ratio Method. Then the fuzzy logic technique has been used. The aim of using fuzzy logic is that some features have been evaluated as weakly relevant according to one of the three techniques described above, whereas the same features have been evaluated as relevant according to other techniques. Consequently, the attribute is neglected if it is evaluated by one of these techniques as being weakly relevant to the target.

After looking at the most important factors or the top five attributes that have been evaluated through the previous procedures, it has been found that both "the independent students from different categories", "the students who send their reports to more than three institutions for obtaining financial aid", and "ACT midpoint degree of students in mathematics" are the features that have a greater correlation with student completion rates or the target.

This study reinforced the results achieved by using two important data mining techniques which are the Random Forest Technique and Support Vector Regression Technique. The results were at the level of work that was performed in this study, and both MAE and RMSE have been used to determine the error ratio in the prediction model.

Overall, the results showed that the random forest technique was slightly better than the Support Vector Regression. The reason for this may be because the dataset is too huge, and the random forest technique performed a type of attribute selection as well an important task of pruning the nodes.

VI. CONCLUSION

This study is concerned with providing models for an academic prediction of the possibilities of students to pass their studies and obtain an academic certificate. This is achieved by using a huge dataset recently launched by the US Department of Education called College Scorecard dataset. It is difficult to understand the dataset by prospective students and their families because it contains hundreds or thousands of features. Therefore, the main work for this project is to determine the most important factors affecting the completion rate of students who are expected to complete their degree within six years. A different approach has been presented in the process of determining the factors that affect the completion of students, as each feature has been evaluated by three filter methods for feature selection which are the Correlation Attribute Evaluation, Relief Attribute Evaluation, and Gain Ratio Method. These three techniques provide different weights to the same feature, so the fuzzy logic technique has been applied in order to have one weight for each feature. After this procedure, the features with the highest weights have been adopted. As such, this project has identified the most significant factors affecting the completion rates of students, and it has been found that the saving rate of selection was for more than 92% for features. Through the method used to identify significant features, and by looking at features that have higher weights, it has been found that independent students from different categories, and students who send their reports to more than three universities to receive financial aid, as well as students who excel in ACT midpoint in mathematics had a stronger correlation with the target. Overall, through the fuzzy logic technique used in this study, it has been noticed that some categories have been excluded completely (e.g. an academic category) while other categories appeared strongly (e.g. students, cost, financial aid, and repayment category). The results showed that these factors are agreed upon by both the random forest technique and support vector regression technique because these features have been reduced as much as possible while maintaining high accuracy in comparison to previous researches. Finally, this study has contributed to providing concise and understandable factors and it is therefore hoped that this work complements with the rest of the research in this field by offering a more detailed insight about the students' completion of their studies.

REFERENCES

1. Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 415–421).
2. Agrawal, M., Ganesan, P., & Wyngarden, K. (2017). Prediction of Post-Collegiate Earnings and Debt. CS.
3. Wotaifi, T. A., & Al-Shamery, E. S. (2018). FUZZY-FILTER FEATURE SELECTION FOR ENVISIONING THE EARNINGS OF HIGHER EDUCATION GRADUATES. Compusoft, 7(12), 2969–2975.
4. Alharbi, Z., Cornford, J., Dolder, L., & De La Iglesia, B. (2016). Using data mining techniques to predict students at risk of poor performance. In 2016 SAI Computing Conference (SAI) (pp. 523–531).
5. Wright, E., Hao, Q., Rasheed, K., & Liu, Y. (2018). Feature Selection of Post-graduation Income of College Students in the United States. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 38–45).
6. Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. Journal of Educational and Behavioral Statistics, 42(1), 85–106.
7. Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. ArXiv Preprint ArXiv:1201.3417.

8. Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. International Journal for the Scholarship of Teaching and Learning, 4(2), 17.
9. Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. In 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1200–1205).
10. Hall, M. A. (1999). Feature selection for discrete and numeric class machine learning.
11. Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of Database Systems, 532–538.
12. Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. In Machine Learning: Proceedings of the Fourteenth International Conference (ICML97) (Vol. 5, pp. 296–304).
13. Luo, B., Zhang, Q., & Mohanty, S. D. (2018). Data-Driven Exploration of Factors Affecting Federal Student Loan Repayment. Retrieved from http://arxiv.org/abs/1805.01586
14. Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493–507.
15. Kumar, M., & Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. In Indian institute of capital markets 9th capital markets conference paper.
16. Li, Y., Bontcheva, K., & Cunningham, H. (2009). Adapting SVM for data sparseness and imbalance: a case study in information extraction. Natural Language Engineering, 15(2), 241–271.
17. Alasadi, S. A., & Bhaya, W. S. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences, 12(16), 4102–4107.
18. Thakur, M. (2007). The impact of ranking systems on higher education and its stakeholders. Journal of Institutional Research, 13(1), 83–96.
19. Eckel, P. D., & King, J. E. (2004). An overview of higher education in the United States: Diversity, access and the role of the marketplace. American Council on Education.
20. Buchmann, C., & DiPrete, T. A. (2006). The growing female advantage in college completion: The role of family background and academic achievement. American Sociological Review, 71(4), 515–541.
21. Wotaifi, T. A., & Al-Shamery, E. S. (2018). FUZZY-FILTER FEATURE SELECTION FOR ENVISIONING THE EARNINGS OF HIGHER EDUCATION GRADUATES. Compusoft, 7(12), 2969–2975.