International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 12 (2020) pp. 056-065

Received: 18 May, 2019; Accepted: 3 March, 2020; Published: 31 March, 2020

Modified Random Forest based Graduates Earning of Higher Education Mining

Tahseen A. Wotaifi[1], Eman S. Al-Shamery[2]

[1] College of Information Technology, University of Babylon,
Hillah, Babil, Iraq
tahseen.ubabylon@gmail.com

[2] College of Information Technology, University of Babylon,
Hillah, Babil, Iraq
emanalshamery@itnet.uobabylon.edu.iq

***Abstract*: With the significant trend of students and families towards higher education and the great change in the labor market, great attention is paid to the issue of job opportunities and the earnings of graduates. However, according to the principle of contemporary education, the policies of educational institutions changed significantly in terms of preparing and qualifying the students to compete for employment. This study aims at 1) identifying the important factors (relevant features) affecting the earnings, and 2) designing a system to predict in the employment of alumni. The new major contributions presented in this work are: the identification of the most important factors by using the fuzzy logic technique in the filter methods for feature selection, and the suggested prediction model by controlling the bootstrap samples that are selected for building the forest in the random forest algorithm. The proposed system has been carried out in light of the higher education system in the United States (US) and has been implemented on the college scorecard dataset. This dataset contains nearly (8000) colleges and exactly (1825) features, so the mechanism of selecting the relevant factors and ignoring the irrelevant features is performed using four methods: Fuzzy-Selection Method (FSM), Mean Decrease Impurity (MDI), Drop-Feature Importance (DFI), and Wrapper-Forward Selection (WFS). According to these methods, it has been found that there is a reduction rate of selection for more than 98% of the factors. Therefore, the Modified Random Forest Regression (MRFR) model is used with two other models: Linear regression and Support vector regression for comparison to predict the earnings of graduates. The Mean Absolute Error (MAE) values for these models are (0.052), (0.068), and (0.068) respectively. The research findings are better in terms of reducing the number of factors and MAE in comparison to previous works.**

***Keywords*:** Earning Prediction of Graduates, Fuzzy-Selection Method, Education Data Mining, Random Forest, Linear Regression, Support Vector Regression.

I. Introduction

The educational sector, including students and graduates, represents the cultural face of the country. Countries that support the education sector and provide the requirements for graduates may notice improvement on an economic level [1]. The policies of educational institutions differ in the preparation and qualification of graduates in terms of experience and skills, which enable graduates or even students to compete in the labor market. It has been noted that there is a large gap in earnings between the graduates because of some of these institutions, as some have faculties with a relatively higher scientific credibility which eventually play an important role in employment [2]. The quality of higher education systems plays a vital role in presenting an appropriate context for educational institutions or colleges. Many countries provide an excellent environment for higher education such as academic freedom, excellent research tools, and others. Therefore, the systems of these countries occupy a high ranking according to Quacquarelli Symonds (QS) [3]. In this project, the higher education system in the United States (US) has been used as a base, as it represents one of the important educational systems in the world. In order to focus on other factors affecting the earnings of graduates, the US Department of Education launched a college scorecard dataset in September 2015. This dataset provides an insight for prospective students to select an appropriate university, because it contains detailed information about educational institutions as well as about former students themselves such as their demographics, race, family income, costs, financial aid, fees, and others [1]. Since not all of these factors affect the earnings of graduates, the important features need to be identified for which data mining techniques have been used.

Educational Data Mining (EDM) recently emerged because of the great interest of researchers in education and the analysis of educational databases [4]. This field is concerned with developing techniques to extract valuable information from the educational databases in order to analyze the student's orientation of the education as well as other aspects. In other words, EDM produces many methods that can help the educational system discover any information related to graduates and students such as the improvement of the learning experience of students, earning of alumni, students' failure rate, student performance, to eventually send alerts to the faculty [5].

This work has focused on the important features affecting the earnings of graduates according to the higher education

system in the United States. This is achieved by applying many data mining techniques on college scorecard dataset then extracting the most important factors and thus enabling prospective students to select the suitable educational institution before entering it.

Outline of the Paper:
Section (II) includes the related works. Section (III) shows the theoretical background. Section (IV) explains the methodology of research including dataset, data pre-processing, feature selection methods, and prediction models. Section (V) illustrates the results. Section (VI) covers the research conclusions. Finally, section (VII) contains the references.

II. Related Work

Being newly-released, there are not many works on college scorecard datasets. Unlike previous studies which revolve around the effect of the selection of university on the income of graduates, this study introduces a comprehensive analysis and then identifies the most important factors related to the earnings of graduates.

Agrawal et al., 2017, used a college scorecard dataset to identify the features which affect the earning of graduates. The study applied a variety of feature selection methods and predictions models for analyzing this dataset. This study has been characterized by many positive aspects, but neither were the findings at the required level nor was the number of important factors identified reduced by a small subset (170 features have been identified).

Miranda Strand and Tommy Truong, 2016, applied data mining techniques on a college scorecard dataset to predict the earnings of alumni. The study has not been exhaustive of all the features in this dataset as it was limited to data released from the US Treasury Department only, whereas in fact, many other features were found to affect the earnings.

Ewan Wright et al., 2017, analyzed the dataset using data mining techniques. The study has applied many feature selection methods to identify the features affecting the incomes of graduates. Although the study provided a scientific analysis of the college scorecard dataset, some categories of the college scorecard dataset have not been studied.

Nunley et al., 2016, used experimental data from a resume audit to estimate the effect of internship experience and particular college majors on employment opportunities. The study applied data mining techniques and found that these factors affect the earnings of graduates to a certain extent.

Elite Schools and Opting-In, 2018, presented a study on how attendance at elite colleges affects the future of graduates. Using College and Beyond data, this study focused on the statistical analysis of students' data with full-time and part-time. The study then expanded to include the impact of these colleges on male and female graduates and pointed out that the effect of college selectivity on earnings is significantly larger for females than for males.

III. Theoretical Background

A. Preprocessing

The main purpose of using data-preprocessing techniques is to prepare data, which tends to be of low quality. This is achieved through reducing the size of the dataset in order to obtain more efficient analysis, thus adapting the data to suit the analysis technique [6].

B. Feature Selection

Feature selection methods are important techniques to reduce the dimensionality of data by selecting the most important features from the features of the original dataset. These techniques are necessary because of the variety of attribute relevance; some are strongly relevant, whereas others are weakly relevant or irrelevant. Therefore, there are two important purposes of identifying the features which are strongly relevant to the target: to increase prediction accuracy and to reduce execution time [7].
Among the many techniques of feature selection, the most widely used are the filter technique, embedded technique, and wrappers technique [8].

1) Filter Methods

Filter methods select attributes based on a performance measure, regardless of the prediction model, and use feature ranking techniques as criteria for feature selection. After evaluating all features, the threshold is used to remove the features that are worth less than this threshold. There are many types of filter methods such as relief method, correlation method, information gain, and others. Each method has a different approach for evaluating the features [9].

2) Wrapper Methods

Wrapper methods search through the space of attributes subsets by using a prediction algorithm. Depending on the accuracy, the attribute can be either removed or retained in the features subset [10].

3) Embedded Methods

In contrast to the wrapper and filter approaches, in embedded method, the feature selection part and the learning part cannot be separated as the embedded method performs feature selection during the implementation of the prediction model [11].

C. Prediction Models

1) Random Forest

Random forest is an ensemble predictor technique that depends on the decision tree model. This technique can be used for both classification and regression. With regression, the result is returned based on the average output of all trees while with classification, the result is returned based on the votes of all trees [12]. There are many characteristics to this algorithm: this method works efficiently on huge databases and produces highly accurate predictors, handling thousands of input features without prune, and it is an efficient technique the estimation of missing values [13].

2) Support Vector Machine

The Support vector machine is a supervised learning method which has been used for regression and classification (with numerical data, this method is also called Support Vector Regression). This method uses the margin from the unlabeled example to the classification hyperplane as a metric for the importance of the example for learning [14]. This technique carries several characteristics: including the processing of linear predictions with high efficiency by the kernel function, avoiding the issue of dimensionality (because it works well with the high-dimensional dataset), and most importantly are the "support vectors" where the decision boundary are represented using a subset of the training examples [15].

3) Linear Regression

Linear regression is a supervised learning algorithm which forms a special case of regression analysis. The main idea of the linear regression model is to explain the relationship between a dependent variable (usually denoted by Y and it means target class) and one or more independent variables (the features) using a straight line [16]. In order to obtain the best results, the linear regression model attempts to make the vertical distance between the data points and the line as small as possible (fitting the line to the data points), in other words, this technique tries to minimize the sum of squares (least squares) [17].

IV. Research Methodology

A. College Scorecard Dataset

The US Department of Higher Education monitored its colleges from 1996 to 2015 and launched the dataset in September 2015 [16]. This dataset is designed to put the choice in the hands of students and families to compare how well colleges are preparing students and thus identify the appropriate educational institution [1]. The College scorecard dataset contains more than 8000 colleges and approximately 2000 factors including study costs, educational expenses, demographics of students, percentage of students (by gender, ethnicity, and color in each educational institution), financial aid (e.g. PALL grant and loans), family income (high, medium, and low), and many others. In order to understand this huge dataset, it has been divided into nine categories: financial aid, student, admission, costs, completion, repayment, school, academics, and earning. Each of these categories contains a number of factors as shown in Table 1 below [18].

Category	Number of factors	Description
Admission	25	Include information about the admissions rate and ACT /SAT scores
Cost	52	Include study fees and costs
financial aid	40	Include grants and loans offered for students.
Earning	73	information about the earnings of graduates.
Student	96	Information about the students such as demographics, family income, and others
Repayment	131	Includes repayment of student and default rate
School	170	Include information about colleges
Academics	228	The type of academic program in the Colleges
Completion	1013	Include US Treasury Department information and others

Table 1. Number of Factors in each Category and Description.

B. Data Preprocessing

Since the college scorecard dataset is very large and many of attribute values have been listed as "PrivacySuppressed" and "NULL", the preprocessing of data has been performed.

In this work, data pre-processing has been performed in three steps as follows:

1) *Data Cleaning*: deleting any attributes which conform to the following conditions:

(a) Features containing a single value in all instances (colleges).
(b) Attributes that are not useful in prediction (such as ID number).
(c) Attributes in which more than (50%) of their entries are missing values.

2) *Processing missing values:* the missing values have been handled in two methods: the mode method and the mean method. Each method has been applied depending on the type of attribute value.

3) *Normalization*: In order to avoid having attributes with large values that control the results of the calculation, the min-max normalization method has been applied to normalize all attribute values into a range between (0) and (1). For more details, see Algorithm (1) below.

Algorithm 1. Data-Preprocessing

Input: Array of features (Dij) where i: number of instances and j number of features.
Output: Relevant features

```
// Data Cleaning
Begin
1 for i = 1 to n          // where n: number of features
2    for j = 1 to m       // where m: number of instances
3    if (all att. values are equal or missing  value>= 0.5)
4       remove the feature from (Dij)
5    end if
6  end for
7 end for
// Missing Values
8 for i = 1 to n
9    for j = 1 to m
10     if attribute value v is missing
11        if all attribute values in the feature are different
12           v = μ = mean
13        else
14           v = M = mode
```

15 end if
16 end for
17 end for
// Normalization
18 for i = 1 to n
19 set min and max to the initial value
20 for j = 1 to m
21 compute min and max in feature i
22 $\bar{v} = \frac{v - \min j}{\max j - \min j}$
23 DPij = $\bar{v}$
24 end for
25 end for
26 return DPij
end

C. Fuzzy-Selection Method (FSM).

According to the FSM, the fuzzy logic technique has been included in three filter methods (Relief Attribute Evaluation, Classifier Attribute Evaluation, and Correlation Attribute Evaluation) to identify relevant features. In general, the irrelevant features have been excluded in three stages:

1) Separately, all weak attributes by estimating the three filter methods are removed as follows: 1) Relief method evaluates the features with weights from (1) to (-1), so any feature weighing less than zero has been removed, 2) Correlation technique evaluates the feature with weight from (-1) to (1), so any feature with zero correlation has been removed.

2) It is natural that the factors chosen by all techniques are better than the factors neglected by one of the techniques, so the factors are selected, are the ones that have not been removed by any of the three techniques.

3) Since all features are evaluated by three filter methods and thus any feature has three different weights, the fuzzy logic technique has been applied in order to have one weight for each feature.

In this study, Membership Function (MF) has been adopted as triangular where Equation (1) below has been applied for calculating the fuzzy value of each crisp value (weight of feature).

$$MF = \frac{wi - a}{b - a} \quad (1)$$

Where: MF is a membership function, w_i representing input value (crisp value), *a* represents the lowest possible input, and *b* represents the highest possible input. Figure 1 below illustrates a triangular membership function:

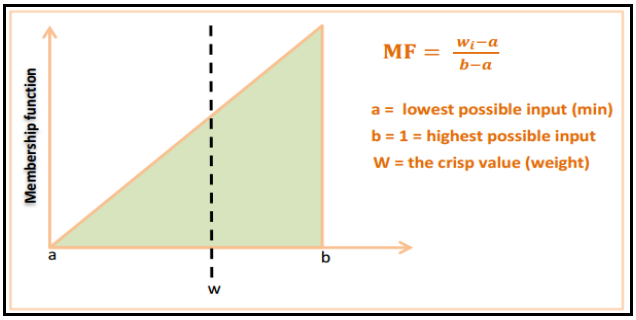


Figure 1. Triangular Membership Function.

For converting the fuzzy values to a crisp value and thus obtaining a single weight for each feature, the Center of Gravity (COG) method has been used. The equation of this method is explained below:

$$COG = \frac{\sum_{i=1}^{n} ?wi) \ * wi}{\sum_{i=1}^{n} ?wi)} \quad (2)$$

Where: COG represents the final weight of the feature, w_i represents a crisp value, and $\mu(w_i)$ represents an MF of the crisp value.

After applying the three stages above, any feature having a weight less than (0.4) has been ignored. This threshold has been chosen so that the irrelevant factors are significantly reduced with an acceptable error rate. After this step, the number of remaining features represents the most important factors affecting the earning of graduates. The number of the remaining variables are (120) only after FSM. Algorithm 2 shown below summarizes FSM:

Algorithm 2. FSM

Input: The output of the algorithm (1).
Output: significant features
// Relief Method
Begin
1 for f = 1 to n //where n represents the number of features
2 compute weight w of feature F according to relief method
3 if w < Θ // Θ = 0
4 delete f from the array
5 else
6 RF(f) = F
7 end if
8 end for
// Correlation Method
9 for f = 1 to n //where n represents the number of features
10 compute weight w of feature F according to correlation.
11 if w = Θ // Θ = 0
12 delete f from the array
13 else
14 CO(f) = F
15 end if
16 end for
// Classifier Method
17 for f = 1 to n //where n represents the number of features
18 compute weight w of feature F according to classifier tech.

19 CL(f) = F
20 end for

// Voting

21 for f = 1 to n
22 if (F ∉ RF(f) or F ∉ CO(f) or F ∉ CL(f) then
23 delete feature F from the array
24 end if
25 end for

// Fuzzy Logic

26 for i = 1 to n //where n is number of features
27 for f = 1 to m //where m is weight in three tech.
// compute membership function
28 $MF = \frac{wi - a}{b - a}$ //where a: the lowest possible input and b = 1 = the highest possible input.
29 end for

// compute the center of gravity

30 $COG = \frac{\sum_{i=1}^{n} (?wi) * wi}{\sum_{i=1}^{n} (?wi)}$
31 if COG < Θ // Θ < 0.4
32 delete feature F from an array
33 else
34 SF = COG
35 end for
36 return SF
End

Figure 2 below illustrates the process of selecting the most important features (factors) affecting the earnings of graduates according to the FSM.

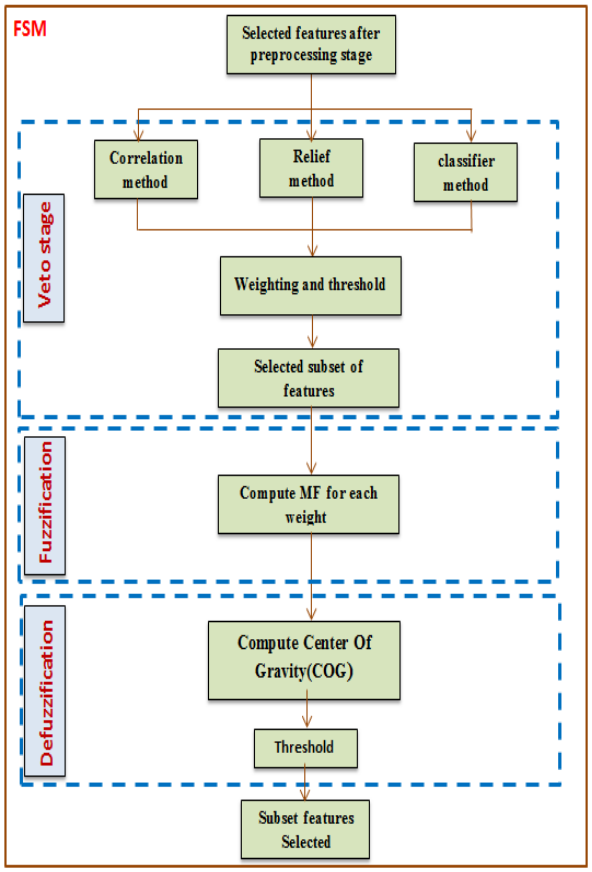


Figure 2. FSM for Earning.

D. Embedded Method

The embedded methods are always implicitly implemented with prediction techniques. Depending on the random forest algorithm, the remaining attributes after the FSM are evaluated again. These features have been evaluated based on the decrease in impurity when the attribute is selected as the split node in any tree of the forest. The mean of all contributions across all trees for a given feature is taken as importance or weight to that feature, and this is called the Mean Decrease of Impurity (MDI).

After that, any feature with a significance degree (weight) less than (0.04) is removed. The number of features that remained after this process were only (56). Algorithm 3 below explains the MDI method:

Algorithm 3. MDI

Input: The output of the algorithm (2).

Output: significant features

Begin

1 for i = 1 to ntree

// compute impurity in each tree

2 for f = 1 to n //n is a number of the feature in a single tree

3 compute impurity IM before splitting

4 select f as a split node

5 compute impurity to the left and right child (LC and RC)

// compute the importance of features IMPF

6 IMPFij = IM – (LC+RC)
7 end for
8 end for
// compute mean decrease impurity (MDI)
9 for i = 1 to n
10 MDI = (sum of IM in all trees)/(number of trees)
11 end for
12 return MDI
End

Figure 3 shows the process of selecting the important features that affect the earning of graduates according to the MDI method.

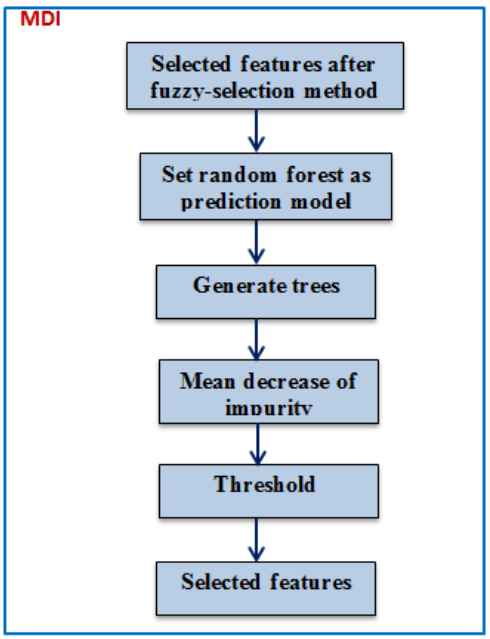


Figure 3. MDI Method.

E. Wrapper Method

The random forest algorithm is applied with the wrapper methods to select the best subset of remaining attributes after the MDI method. Two different approaches to the wrapper methods have been applied for this selection: Drop-Feature Importance (DFI) and Wrapper-Forward Selection (WFS).

1) DFI: According to this method, the attribute has been evaluated by dropping the attribute from the dataset and then re-evaluating it. The importance of each feature is the difference between the error with this feature and the error after dropping it. Then, any feature having a weight less than (0.1) has been removed. The remaining variables or features are only (35) features.

2) WFS: In this method, the forward selection is applied to select the best subset of features. According to WFS method, each feature is tested separately where it starts with an empty set and then evaluates the features based on the forward selection. Finally, the best subset of features is selected. After the WFS method, the remaining variables are only (20) features.

Algorithm 4 below illustrates both the DFI method and WFS method:

Algorithm 4. Wrapper Methods

Input: The output of the algorithm (3).
Output: significant features

// DFI

Begin
1 set random forest as a prediction model
2 compute MAE before dropping the feature MAE_B

$MAE\_B = \frac{1}{n}\sum_{i=1}^{n} |Pi - Ai|$ // where Pi is predicted value and Ai is actual value

3 for i = 1 to n
4 drop feature i
5 compute MAE after dropping the feature MAE_A
// Compute the significant feature (SD)
6 SD = MAE_A – MAE_B
7 if SD > Θ // Θ = 0.1
8 SDi = SD
9 end if
10 return this feature to the dataset
11 end for

// WFS

12 set random forest as the prediction model
13 set forward selection as a search strategy
14 start with the empty set
15 for f =1 to n // where n is number of features
16 eval[f] = cross-validation(RF, f, 10, Random)
17 end for
18 WFS = the best result in array eval[f]
19 merge WFS with other features
20 re-evaluation until all attributes are merged and evaluated
22 return WFS
End

Figures 4 and 5 below show the process of selecting the important features affecting the earnings of graduates according to the DFI and WFS methods.

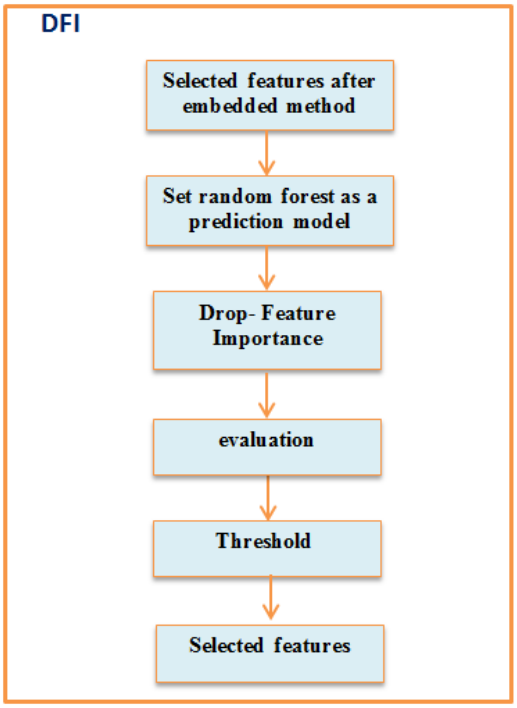


Figure 4. DFI Method.

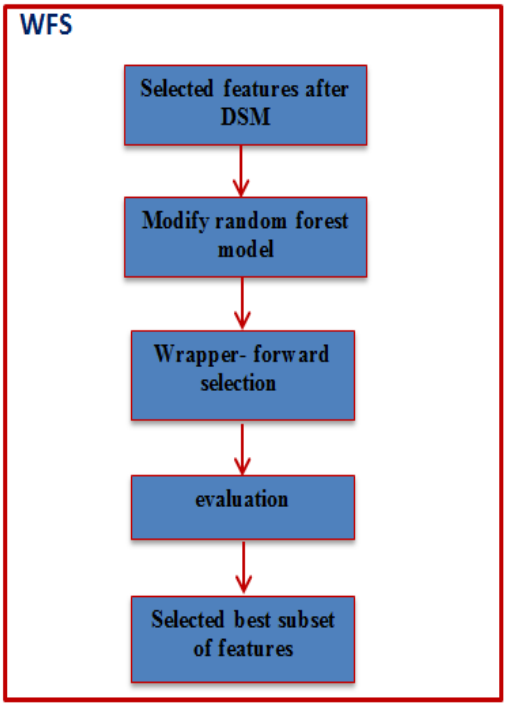


Figure 5. WFS Method.

Generally, the main objective of using each of FSM, MDI, DFI, and WFS is to exclude irrelevant features and identify the most important ones which affect graduates' earnings. Each method contributed to reducing a certain number of insignificant features.

In order to validate this subset of features which have been identified as having the most impact on the earning of

graduates, the prediction models have been used. The Random forest algorithm has been improved to predict the earning. The results of this method are compared with two important techniques: the linear regression and the support vector regression.

V. Experimental Results and Discussion

A. Modified Random Forest Regression (MRFR) Model

Reducing the features to the lowest possible with maintaining the least prediction error is the main purpose of this paper. Therefore, with the methods used to select important factors on earnings, the random forest algorithm has been improved. The random forest algorithm selects the features in each tree randomly. This algorithm is improved through the identification of the best features in all trees of the forest. After the WFS method, 20 features are identified as the best subset of all features in the dataset. These features are evaluated in three methods: FSM, MDI, and DFI. So, in order to have one weight for each attribute, the average has been calculated. The top five features with higher weights have been identified in all trees of the forest. Table 2 shows the best features according to the three methods above.

Features	FSM	MDI	DFI	Mean
AVGFACSAL	**0.284**	**0.158**	**0.079**	**0.615**
MALE_RPY_7YR_RT	**0.303**	**0.181**	**0.130**	**0.597**
DEP_INC_AVG	**0.280**	**0.124**	**0.092**	**0.625**
PELL_EVER	**0.334**	**0.173**	**0.230**	**0.598**
MD_FAMINC	**0.296**	**0.139**	**0.196**	**0.554**
HI_INC_WDRAW_RT	0.196	0.122	0.040	0.427
DEP_WDRAW_ RT	0.203	0.121	0.041	0.448
NOPELL_ENRL_RT	0.210	0.112	0.060	0.457
NOPELL_WDRAW_RT	0.192	0.111	0.051	0.415
NOPELL_COMP_RT	0.210	0.091	0.061	0.477
SAT_AVG_ALL	0.251	0.132	0.073	0.547
DEP_INC_PCT_LO	0.276	0.150	0.125	0.554
PAR_ED_PCT_1STGEN	0.199	0.123	0.042	0.432
INC_PCT_H2	0.252	0.142	0.098	0.515
DEP_INC_PCT_M1	0.195	0.115	0.067	0.404
PAR_ED_PCT_PS	0.207	0.143	0.047	0.432
SATMT25	0.247	0.112	0.055	0.573
CDR3	0.267	0.153	0.10	0.548
FIRST_GEN	0.207	0.142	0.040	0.44
TUITIONFEE_OUT	0.219	0.125	0.049	0.483

Table 2. Average of Three Methods.

Table 3 below shows a brief description of the top five features:

Feature	Description
AVGFACSAL	Average salaries of the faculty.
MALE_RPY_RT	The repayment rate of male students.
DEP_INC_AVG	Average income of dependent students.
PELL_EVER	Students who do not need PALL_GRANT.
MD_FAMINC	Average family income.

Table 3. The top Five Features.

Figures 6, 7, 8 and 9 below show the significant degree (weights) for the 20 features that have been classified as most important and are listed in Table 2 according to the FSM, MDI, DFI and Mean methods:

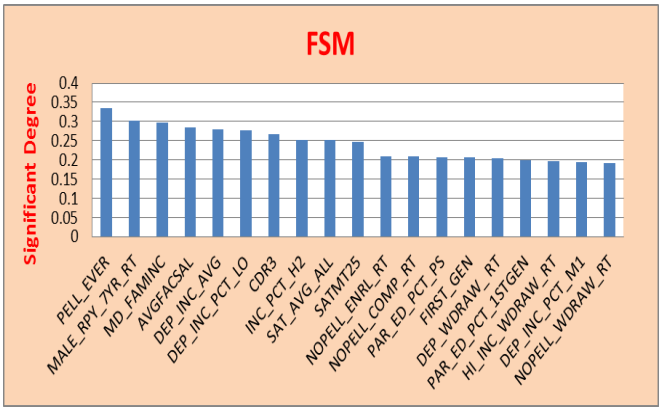


Figure 6. Significant Degree of each Feature (FSM).

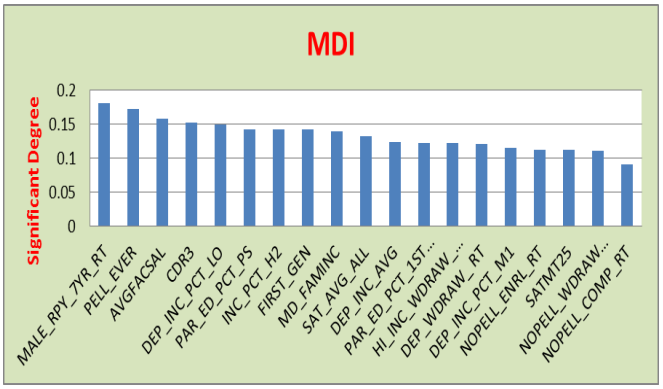


Figure 7. Significant Degree of each Feature (MDI).

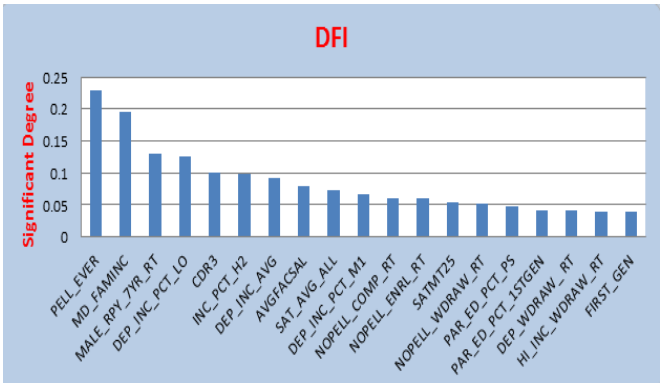


Figure 8. Significant Degree of each Feature (DFI).

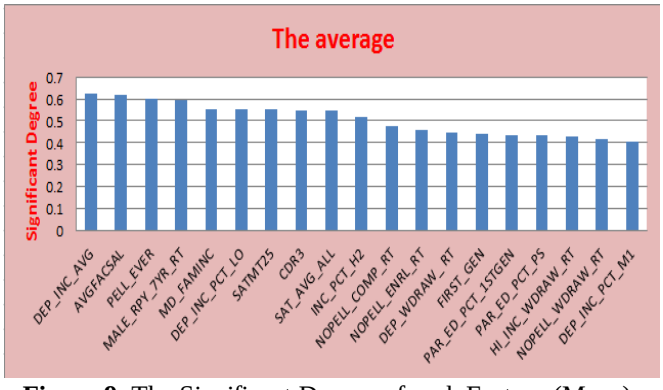


Figure 9. The Significant Degree of each Feature (Mean).

With reference to the figures above, it has been found that the top five features shown in Table 2 are important in estimating the three methods (FSM, MDI, and DFI).

Both linear regression and support vector regression have been used in the comparison with the MRFR model. The performance of the models is evaluated through 10-fold cross-validation with Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as error metrics.

In this study, the irrelevant features are excluded and the factors affecting the earnings of graduates have been identified through four stages or four methods: Fuzzy-Selection Method, Mean Decrease of Impurity, Drop-Feature Importance, and Wrapper-Forward Selection.

After applying the Fuzzy-selection Method, the number of remaining features are only (120). The results of the three models are illustrated in Table 4.

Models	MAE	RMSE
Random forest model	0.055	0.077
Linear regression	0.060	0.081
SV regression	0.060	0.082

Table 4. The Results after The FSM.

The number of the remaining features are only (56) by using the Mean Decrease of Impurity method. The results of the three models are explained in Table 5.

Models	MAE	RMSE
Random forest	0.054	0.076
Linear regression	0.062	0.084
SV regression	0.062	0.085

Table 5. The Results after The Embedded Method.

The number of remaining features are only (35) after applying the Drop-Feature Importance. The results of the three models are shown in Table 6.

Models	MAE	RMSE
Random forest	0.052	0.073
Linear regression	0.063	0.085
SV regression	0.063	0.086

Table 6. The Results after The DFI Method.

The number of remaining features are only (20) by applying Wrapper-forward selection. The results of the three models are illustrated in Table 7.

Models	MAE	RMSE
MRF model	0.052	0.073
Linear regression	0.068	0.091
SV regression	0.068	0.091

Table 7. The Results after The WFS Method.

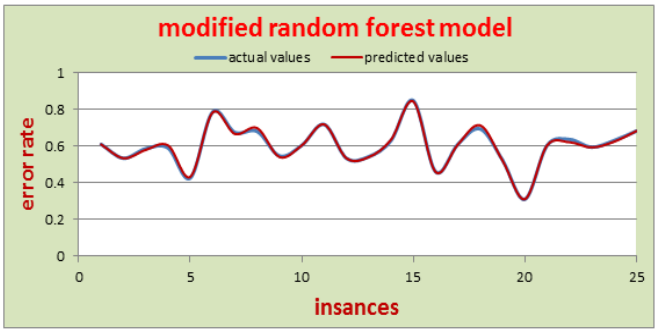
Although the factors are reduced to a large number through the WFS method, the results are similar to the previous stage as shown in Tables 6 and 7 above. In comparison with previous studies, for example, the first paper in the related

(Agrawal, Ganesan, & Wyngarden, 2017. "Prediction of Post-Collegiate Earnings and Debt" from Stanford University), our findings are better in terms of error rates as well as the number of remaining factors (relevant features) which are much lower, as indicated in Table 8 below:

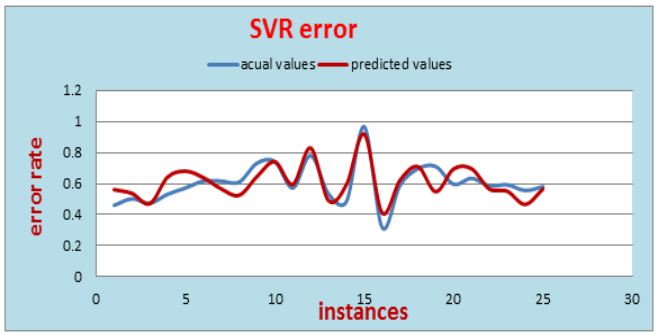
The current study (Modified Random Forest based Graduates Earning of Higher Education Mining). Number of remained factors are 20.		**The previous study (Prediction of Post-Collegiate Earnings and Debt). Number of remained factors are 170.**	
Models	MAE	models	MPE
MRFR model	0.052	Neural network	11.93
SV regression	0.068	Weighted Lin. Reg.	9.60
Linear regression	0.068	Linear regression	13.42
		KNN	14.11
		SVM	23.80

Table 8. The Comparison With The Previous Studies.

Figure 10, 11, and 12 show the difference between predicted and actual values for the three models.



Figures 10. Error for Earning With MRFR Model.



Figures 11. Error for Earning With SVR Model.

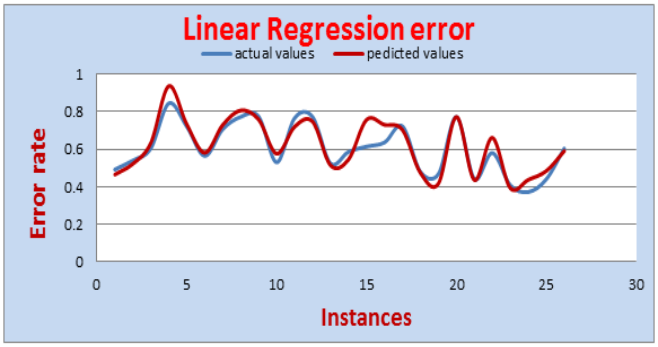


Figure 12. Error for Earning With Linear Regression Model.

Generally, this work presents the techniques of academic prediction on a large dataset recently released by the US Department of Education, namely the college scorecard dataset. In order to understand this large dataset that contains thousands of factors about educational institutions and former students, this study has focused largely on feature selection methods to indicate the factors that are suitable for the prediction task. Four methods of feature selection have been applied respectively on this dataset: FSM, MDI, DFI, and WFS. Practically, the most important method is the FSM as it presented a relatively higher reduction rate when implemented. Using this method, the features are evaluated in more than one stage and, eventually the number of features have been reduced to only 120. The remaining features are reduced again by the MDI method to half the initial number (56 features only) with better results in the prediction model. After the MDI method, the DFI method is applied. Approximately 21 factors have been eliminated using the DFI method with only 35 features remaining. With this reduction, the results of the prediction model are better. Finally, the Wrapper-Forward Selection method is used where 20 features have been identified as the best subset of the original features.

This study focused on the random forest algorithm, after that, the performance of this algorithm has been enhanced. In order to obtain better results, the top five factors are identified in all trees of the forest. By looking at Table (2), it has been found that each of PELL_EVER (students who do not need PALL_GRANT), MD_FAMINC (average family income), AVGFACSAL (Average salaries of the faculty), DEP_INC_AVG (average income of dependent students), and MALE_RPY_7YR_RT (repayment rate of male students) have the highest weights. Therefore, these features are identified in all trees of the forest.

VI. Conclusion

This paper aims at identifying the most important features affecting the earning of graduates as well as designing a system for predicting the median earnings of alumni. This is achieved by using a large dataset recently released by the US Department of Education, namely the college scorecard dataset. The issue of dimensionality is a great challenge faced in this study because this dataset contains thousands of features which are not all relevant to the target class. In order to better understand this dataset, four feature selection methods have been used: Fuzzy-Selection Method, Mean Decrease of Impurity, Drop-Feature Importance, and Wrapper-Forward Selection. The main contribution is found in the Fuzzy-Selection Method, as the features have been evaluated by three filter methods (Correlation method, Relief method, and classifier method) before the fuzzy logic model is proposed to select the best features. Using this method, the features were reduced to only 120, and it has been found that the saving rate of selection was for more than 95% for features. These features have been evaluated again using an embedded method (Mean Decrease of Impurity) with random forest technique. Through this method, the features contributing significantly to the reduction of impurity are given greater importance and thus the features have been reduced to only 56 ones.

The remaining features have been evaluated using Drop-Feature Importance method by the contribution of each feature in minimizing the predicting error. Using this method, the features have been reduced to only 35. Eventually, Wrapper-Forward Selection method has reduced these features to only 20. After applying the four methods mentioned above, it appears that there is a saving rate of selection for more than 98% of the features. In the prediction stage, the random forest algorithm has been improved by controlling in the selection of the bootstrap sample where the top five features have been identified in all trees of the forest. Two prediction models are also used: linear regression and support vector regression, and the results showed that these features are agreed upon by all models but the random forest model has yielded in slightly better results than the other two models. Finally, this study has contributed to providing understandable and concise factors and it is hoped that this work complements with the rest of the research in this field by offering more detailed insights about the earning of graduates.

References

[1] 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.

[2] 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).

[3] Thakur, M. (2007). The impact of ranking systems on higher education and its stakeholders. Journal of Institutional Research, 13(1), 83–96.

[4] 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.

[5] Jacob, J., Jha, K., Kotak, P., & Puthran, S. (2015). Educational Data Mining techniques and their applications. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 1344–1348).

[6] 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.

[7] 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).

[8] Miao, J., & Niu, L. (2016). A survey on feature selection. *Procedia Computer Science*, *91*, 919–926.

[9] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.

[10] Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. In Icml (Vol. 1, pp. 74–81).

[11] Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In Feature extraction (pp. 137–165). Springer.

[12] 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.

[13] Pretorius, A., Bierman, S., & Steel, S. J. (2016). A meta-analysis of research in random forests for classification. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)* (pp. 1–6).

[14] G*eorge, G., & Raj, V. C. (2011). Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. ArXiv Preprint ArXiv:1109.1062*.

[15] Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*, *12*(1), 1–7.

[16] Sunthornjittanon, Supichaya, "Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand" (2015). University Honors Theses. Paper 131. doi: 10.15760/honors.137.

[17] Hurwitz, M., & Smith, J. (2018). Student responsiveness to earnings data in the College Scorecard. *Economic Inquiry*, *56*(2), 1220–1243.

[18] Luo, B., Zhang, Q., & Mohanty, S. D. (2018). Data-Driven Exploration of Factors Affecting Federal Student Loan Repayment. ArXiv Preprint ArXiv:1805.01586.

Author Biographies

Tahseen Ali AL-Wotaifi is an assist lecturer at the faculty of Information Technology, University of Babylon / Hillah, Babil, Iraq. His research interests include Data Mining and Machine Learning.

Eman Al-Shamery received the BSc and MSc degrees in Computer Science from the University of Babylon, Iraq, in 1998 and 2001, respectively. After completing her MSc, she worked as an assistant lecturer at the Department of Computer Science, the University of Babylon. In 2013, she received her PhD in Computer Science from the University of Babylon. Currently, she holds a professor position at Software Department, University of Babylon. Her current research interests include artificial intelligence, bioinformatics, machine learning, neural networks, deep learning and data mining.