A Peer Reviewed
TECHNICAL JOURNAL Vol 2, No.1, October 2020
Nepal Engineers' Association, Gandaki Province
ISSN : 2676-1416 (Print)
Pp.: 89-98

PREDICTING ACADEMIC PERFORMANCE OF ENGINEERING STUDENTS USING ENSEMBLE METHOD

Tek Bist Bithari*, Sharan Thapa, and Hari K.C.
Department of Electronics and Computer Engineering, Pashchimanchal Campus, Institute of Engineering, Tribhuban University, Nepal
**E-mail: tekbist54@gmail.com*

Abstract

One of the common problems that most of the engineering institutions face in recent times is poor academic results. The statistics of the IOE semester result show that since 2009 A.D., the average pass percentage has been reducing in an average from 50% to 40% and moving towards decreasing scenarios. The study aims to predict an engineering student's academic performance based on their past educational records, demographic factors, family backgrounds, and other related factors. Firstly, a predictive model is built using the traditional classifiers Decision Tree, SVM, and Linear Regression, which had shown good performance in similar types of study. After that, we have implemented one of the popular ensemble Methods, voting, which is known for improving the individual classifier's performance. Voting classifier combines the predictions of base classifiers by averaging those predictions. The result revealed that the accuracy, precision, recall, and F1 score had been considerably improved by using ensemble voting than that of the individual classifiers. The data used in the study was collected directly from the hard copy personal files of each pass out student of Paschimanchal Engineering Campus, Pokhara between the years 2004 to 2015 AD.

Keywords

Academic performance, Engineering, Ensemble method, Voting, SVM

1. Introduction

One of the popular fields of interest in the recent times is Educational Data Mining (EDM). Although, in the context of Nepal the educational data are not properly organized, the data mining techniques have been used to extract useful knowledge from the available educational data. The research also contributes towards extracting hidden useful information from the educational data. The previous works similar to the study had mainly focused on using single model for prediction (Aman, Fazal et.al, 2019) but in this research the ensemble model has been used for the prediction. The ensemble method combines the result of multiple individual models that can improve the reliability and performance of the model. Ensemble techniques have been very popular for predictive modeling almost in every field at recent time.

The research contributes mainly to fulfill two objectives:

1) To compare and analyze the performance of the ensemble method in terms of various

performance measures like accuracy, recall, precision, and F1 score.

2) To predict the academic performance of engineering students of Nepal with various attributes.

In this research, we have used the data of engineering students to develop a predictive model that can classify a student's academic performance into one of the four categories (Excellent, Good, Medium, and Satisfactory). The real-world student data, including various educational, demographic, personal, and family attributes have been collected. The possible attributes influencing the academic result of the engineering student were identified by the extensive literature study (Kumari, Pooja, Praphula Kumar Jain, and Rajendra Pamula, 2019). The task was performed by using a supervised data mining technique i.e. Classification. The classification was first performed by using the three traditional classifiers Decision tree, Support Vector Machine (SVM), and Logistic Regression (LR). After that, ensemble voting was implemented by taking these three classifiers as base learners.

Ensemble Methods provide classification accuracy by aggregating the prediction of multiple classifiers. The ensemble method constructs a set of base classifiers from training data and performs classification by taking the vote on the predictions made by each classifier. In this model, for improving the classification accuracy, the voting algorithm was used.

2. Traditional Classifier

2.1 Decision Tree

A decision tree as the name suggest consist of a tree like structure where branch indicates the decision rule, leaf indicates the outcome and the internal node indicates the feature. The tree is partitioned in a recursive manner based on the attribute values. The decision tree can be easily understood by the humans as its visualization is like a flowchart diagram. The processing time for a decision tree depends upon the number of records and the number of attributes and hence it is faster than other machine learning algorithm like neural networks. The decision tree is capable of handling high dimensional data with better accuracy (Poojari. D, 2019).

2.2 SVM

One of the popular supervised machine learning algorithms that is mainly used for the classification task is Support Vector Machine (SVM). In the n-dimensional (where n is number of features) space the sample data is plotted where the coordinate represents the value of each feature. The main task is SVM is to find the optimum hyperplane that separates the multiple classes. The hyperplane can be single point, line, or a plane in the case of 1-dimension, 2-dimension, and 3-dimension respectively. The unseen data can be generalized and classified correclty if the optimum hyperplane is identified (Jinde, S. 2018).

2.3 Logistic Regression

Logistic Regression is a popular supervised machine learning algorithm that uses the sigmoid function (an S-shaped curve that takes any real value and plot and map it between a value 0 and 1). It is mainly used for binary classification task but it can also perform multiple classification as well (R-BLOGGERS, 2019).

3. Ensemble Voting

The main concept of ensemble voting is to combine the predictions of multiple classifiers. This is a popular technique that may be used

to enhance the performance of model than that of a single model. This technique can be used in the case of both classification and regression. In the case of classification, the predictions are performed for each model then it is summed and the label with majority vote is predicted. Voting can be performed in two ways i.e., soft voting and hard voting. Hard voting involves summing the predictions for each individual model and predicting the label with majority vote. Soft voting involves summing the predicted probability for each label and the class label with largest probability is predicted (Brownlee.J, 2020).

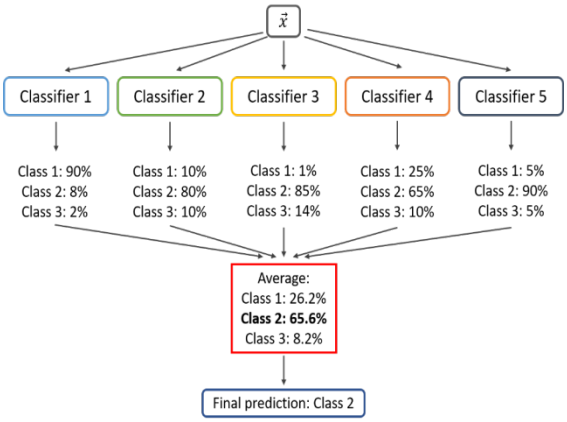

Figure1: Ensemble voting with the soft voting process

(Source: https://mc.ai/ensemble-learning-techniques-votingclassifier/)

4. Related Works

Prediction of student's academic performance is a hot area of research for long. The research community has widely focused on identifying the most predictive features influencing student's academic performance at different levels of their studies. Aman, Rauf, and Ali proposed and intelligent predictive model that accurately recommends a student with an appropriate choice of their study at master's level. A real-world dataset of 1,021 records obtained from the University of Peshawar, consisting of eight academic features, seven socio-economic features, and one demographic feature, was used. Three predictive classes, including the 1st Division, 2nd Division and Fail were used. Logistic Model Trees (LMT) were used for the creation of a predictive model, and its performance was compared with the Random forest and J48 model. The highest 83.15% accuracy was achieved when all the three features were taken into consideration(Aman, Fazal et al.,2019).

Praphula, Kumari, and Pamula, proposed using ensemble Methods in the student academic performance prediction. The result showed significant improvement in all the evaluation measures when ensemble Methods were used. The ensemble voting outperformed the bagging and boosting Methods(Kumari, Pooja, Praphula Kumar Jain, and Rajendra Pamula, 2019).

Mukesh and Sahal, reviewed the research articles related to Student's Academic Performance (SAP) from 2012 to 2017 in order to identify the most used data mining algorithm in the study. It was found that the mostl used classification was among Neural Network(NN), Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbours (KNN), Logistic Regression (LR), etc. where DT has the best performance than others (Kumar, Mukesh, and Yass Khudheir Salal,2019)

In Grade prediction of student academic performance with multiple classification models by Zhang, Liu, and Xue, they performed an exhaustive comparative study on the datasets of students' information provided by the university of electronic science and technology. They compared the performance of supervised

machine learning algorithms such as Naive Bayes, Decision Tree, Multilayer Perceptron and Support Vector Machine. In this research, a variety of models for grade prediction of student academic performance were compared and analyzed based on the grades in students' achievements and the school's behavior information. The results showed that the multilayer perceptron classification model (with an accuracy of 65.90% on the training set and 64.02% on the test set) was the most effective way. It provides a basis for the scientific decision-making of the teaching management department (Zhang, Xu et al,2019).

Alsalman, Yasmeen Shaher et al. have built a classification model to predict academic students' performance in Jordanian universities. While working on predicting the performance, several attributes have been tested, but actually, some of them were effective on students' performance prediction. The failure time was the most powerful attribute then the internet dependency in studying, came in the second level. The work status and the marital status did not showed a direct influence on the prediction, while attributes like the scholarship and Guardian had a degree of effectiveness on prediction. For the teachers and university administration, this model can help them in predicting the students' academic performance to take appropriate actions to enhance this performance (Alsalman, Yasmeen Shaher et al, 2019).

Soobramoney, Ranjin and Alveen Singh, compared all the prior works done on identifying students at risk that were done using different individual machine learning algorithms such as SVM, Artiifcial Neural Network (ANN), Random Forest, J48, and so on. After reviewing this entire prior works, they seem largely unsure of anyone specific machine learning algorithm that performs best in every context in order to identify students that are at risk. It is necessary to consider several factors in order to get successful performances from machine learning algorithms for the prediction of students at-risk. Consequently, they concluded that instead of a single machine learning algorithms, the Ensemble Methods like Bagging and Boosting could lead to greater accuracy(Soobramoney, Ranjin, and Alveen Singh,2019).

Vasileva, Ekaterina E., Daniil S. Kurushin, and Sergey S. Vlasov used a multilayer perceptron as a type of neural network to to predict the total grade point average (GPA) at the end of the university based on data on the academic performance of full-time students in the 2nd year of study. Though the data was very limited in this study on the test sub-sample, the error of the best NN was one and a half percent, which is an excellent result for such a small sample(Vasileva, Ekaterina E., Daniil S. Kurushin, and Sergey S. Vlasov ,2019).

5. Methodology

The framework that has been used in building a predictive model is shown in the Figure 2.

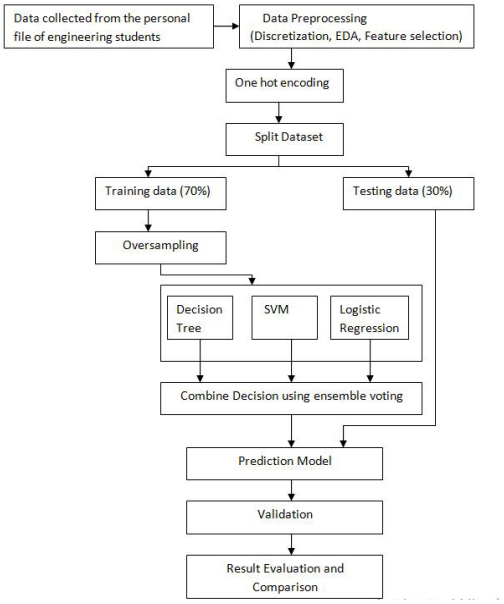

Figure 2: Detail Methodology Steps

5.1 Data Collection

The engineering students' data with required attributes were not available directly on the softcopy, so we collected the data from the hardcopy personal file of each pass-out student from Paschimanchal Engineering Campus, Pokhara, between the years 2004 to 2015 AD. Each student's personal file was provided by the exam department of the respective campus, which includes all the forms filled during the admission, a copy of past academic transcripts, a character certificate, mark sheets of every semester, BE transcripts, and many others.

Table 1: Attribute Description

Attributes	Descriptions
Gender	Male, Female
Scholarship Type	Type of scholarship student has got.
Age	Age of the student at the time of BE admission.
Entrance Rank	Rank in the IOE entrance exam.
Interest	Priority given to the admitted program during the time of BE admission.
Plus two percentage	Aggregate percentage of class 11 and 12.
Plus two location	Location of the plus two college.
Gap	Gap between plus2 and BE admission.
Main subject	Main subject chosen by the student during plus two.
Class ten percentage	Percentage obtained by the student in SLC.
School Location	Location of the school.
School Type	Type of the school.
Ethnicity	Ethnic group of the student.
Batch	Year in which student is admitted.
Program	Program in which student is admitted.
Father Job	Job in which father is engaged.
Class (target)	Excellent (if BE%>=80 or Failure times=0) Good (if BE%>=75 or Failure times<=5) Medium (if BE%>=70 or Failure times<=10) Satisfactory (others)

5.2 Data pre-processing

The real-world dataset is prone to miss values, noisy instances, outliers, and so on. Therefore it becomes necessary to clean the dataset which can lead to optimum performance. After cleaning the dataset we got 2445 cleaned records for further processing.

5.2.1 Discretization: Discretization was used to discretize the numeric attribute into nominal ones based on the class information. We used a discretization mechanism to transform the student performance from numerical values to nominal values, which represents the class labels of the classification problem. We divide the dataset into four nominal intervals (Excellent, Good, Medium, and Satisfactory) based on the BE percentage and number of failure times during the study.

5.2.2 Exploratory Data Analysis: The data used in the study was real-world data, so there were many missing values, outliers, noisy instances, and so on. Therefore various Methods like box plot, scatter plot were used to remove irrelevant data.

5.2.3 Feature Selection: The feature selection process selects an appropriate subset of features that can efficiently describe data,and remove irrelevant data. In the above dataset, the batch attribute does not have any relation with the performance of the student, so it was removed.

After the entire pre-processing task, the categorical data were converted into the numeric data. For this one of the popular techniques, one hot encoding was used. One hot encoding was used to convert the nominal categorical data into the numeric ones. The categorical, ordinal data like entrance rank, interested faculty, the class were converted into numeric by using the replace method, which preserves the order of the data.

The entire data was then spitted into training and testing data. The training data was imbalanced with most of the instances falling to 'Good' class and very few falling to 'Excellent' class. This could hamper the learning of the model. Therefore one of the resampling techniques SMOTE was used to balance the majority and minority classes of training data. Synthetic Minority Over-sampling Technique (SMOTE) is an Over sampling technique that generates synthetic samples from the minority class (Xiaojun Wu and Sufang Meng,2016). It works by creating synthetic observations based upon existing minority observations.

The balanced training data was used for training each of the traditional classifiers i.e., Decision tree, SVM, and Logistic Regression. The performance measure of each model was evaluated using the unseen test data. In order to improve the performance ensemble voting as used by taking these three classifiers as base learners. Voting classifiers combine the predictions of base classifiers (Decision tree, SVM, and Logistic Regression) by averaging those predictions.

6. Result Analysis

6.1 Tools used

Python has been used as the programming language for coding purposes. Python is a widely used, high-level, general-purpose programming language. The Python language's key features are its code readability and its syntax that allows its user to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The ipython jupyter notebook text has been used as a text editor.

6.2 Evaluation measures

For the imbalanced dataset, accuracy is not an appropriate metric (Acharya, Anal, and Devadatta Sinha, 2014) for classification

performance evaluation because in the imbalanced dataset, the accuracy of the learner will be high even when the classifier classifies all the majority samples correctly and misclassifies all the minority samples since the number of majority samples is much more than the number of minority samples, Under such circumstances, accuracy cannot reflect reliable predictions for the minority class. Therefore, metrics beyond accuracy, such as precision, recall, and F1 score have been used to evaluate the performance of the model.

Precision is the ratio of a number of correct positive result and the number of positive results predicted by the classifier, and is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

A recall is the ratio of the number of correct positive results and the number of *all* relevant samples and is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy of any prediction model can be given as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

F1-score is the harmonic mean of precision and recall, and is given by:

$$\text{F1-score} = \frac{2*precision*recall}{precision+recall}$$

6.3 Evaluation using cross-validation

The training set was cross-validated, and the mean training accuracy and the validation accuracy were evaluated as shown in table 2. The result indicates that the SVM got the highest testing accuracy among the individual classifiers. The result shows that the difference between training and testing accuracy has been very less in the case of ensemble voting. Thus, overfitting has been removed by applying ensemble voting.

Table 2: Cross-validation result with 10 fold

Accuracy Score	DT	SVM	LR	Voting
Training Set	81.73	82.08	78.6	84.03
Validation set	76.0	77.1	73.1	79.73
Testing set	74.2	78.3	72.0	82.01

6.4 Evaluation Results

The learning curve for each individual classifier and ensemble voting has shown that the mean training and validation accuracy was closer in the case of ensemble voting. This indicates that the predictive model in the case of ensemble voting has been trained well over other the individual classifiers and shows no sign of overfitting

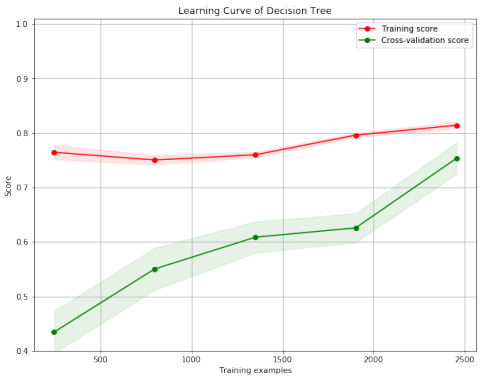

Figure 3: The learning curve of DT

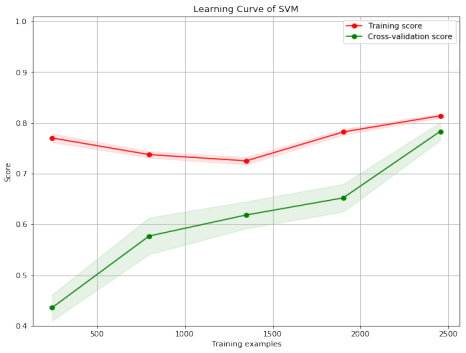

Figure 4: The learning curve of SVM

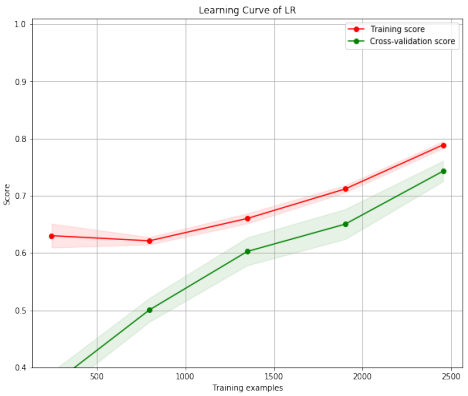

Figure 5: The learning curve of LR

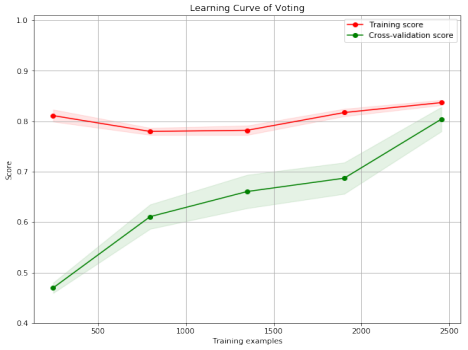

Figure 6: The learning curve of Ensemble Voting

Table 3 shows that SVM has demonstrated better performance in the case of traditional classifiers. The accuracy, precision, recall, and F1 score for SVM were better than other classifiers. Ensemble voting has shown the best accuracy of 82 %, which was a significant improvement in comparison to the individual classifiers. Similarly, recall, precision, and F1 score were also better in ensemble voting, as shown in the Figure 7.

Table 3: Classification results using ensemble voting

Classifier	DT	SVM	LR	Voting
Accuracy	74	78	72	82
Recall	75	80	73	83
Precision	75	78	72	82
F1-score	75	78	72	82

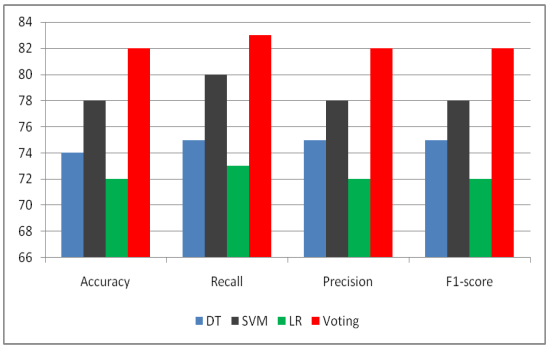

Figure 7: Comparison analysis using traditional classifiers and ensemble voting

Validation is an important phase in building the predictive model; it analyses, how realistic the predictive model is. The model was validated by using 30% unseen student records which were previously split into test data. Figure 8 shows the confusion matrix for ensemble voting. Among the 79 excellent students the predictive model predicts 57 correctly. Out of 321 good students, 266 students were correctly classified as good. Similarly, out of 191 medium students 156 students were correctly classified as medium and out of 143 satisfactory students 123 students were correctly classified as satisfactory students. The result shows the reliability of the proposed model.

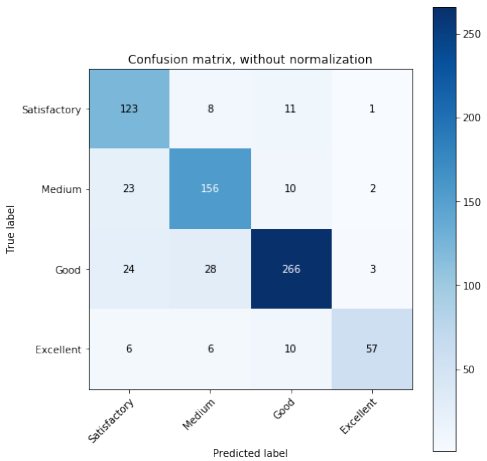

Figure 8: Confusion matrix of ensemble voting

7. Discussion

There has been a lot of prior studies done on Student Performance Prediction(SAP) all over the world for students of various levels. The prior works suggest that only the past academic results of the student are sufficient predictor of student academic performance in future. There are lot of factors that influence the peformance of the students in examination. The study includes possible factors(demographic factor, family factor, personal factor) that can have influence in the acadmenic performance. The literature suggest that some specific machine learning cannot provide best result in every context as given by authors (Soobramoney, Ranjin and Alvin singh,2019) in the case of Student Performance Prediction(SAP). The use of the ensemble method was able to improve the result than that of the single model. Hence the study add a brick towards improving the accuracy of the model.

8. Conclusion

In this study the predictive model has been built on the data of engineering students which includes various attributes. The multi-class classification has been done to predict the students into four categories (Excellent, Good, Medium, and Satisfactory). The classification has been done by three individual traditional classifiers first and then the voting was done in the second phase. The result obtained shows significant improvement in the performance when the ensemble method was implemented. It also removes the slight overfitting which was seen in the case of some individual classifier. In the future, discovering various other attributes that influence an engineering student's academic performance can further improve the result.

References

Aman, Fazal et al., 2019. "A Predictive Model For Predicting Students Academic Performance." *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*

Zhang, Xu et al,2018. "Grade Prediction Of Student Academic Performance With Multiple Classification Models." *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*

Alsalman, Yasmeen Shaher et al.,2019 "Using Decision Tree And Artificial Neural Network To Predict Students Academic Performance." *2019 10th International Conference on Information and Communication Systems (ICICS)*

Kumari, Pooja, Praphula Kumar Jain, and Rajendra Pamula.,2018. "An efficient use of ensemble Methods to predict students academic performance." *2018 4th International Conference on Recent Advances in Information Technology (RAIT).*

Kumar, Mukesh, and Yass Khudheir Salal.,2019. "Systematic Review of Predicting Student's Performance in Academics." *Int. J. of Engineering and Advanced Technology* 8.3 pg:54-61

Soobramoney, Ranjin, and Alveen Singh.,2019. "Identifying Students At-Risk With An Ensemble Of Machine Learning Algorithms." *2019 Conference on Information Communications Technology and Society (ICTAS)* (2019):

Vasileva, Ekaterina E., Daniil S. Kurushin, and Sergey S. Vlasov.,2019. "Early Prediction Of The Grade Point Average Of University Students Diploma:

Neural Network Approach." *2019 XXII International Conference on Soft Computing and Measurements (SCM))*

Poojari, D., 2019. *machine-learning-basics-descision-tree-from-scratch-part-i-4251bfa1b45c*. Retrieved July 24, 2020, from towardsdatascience.com: https://towardsdatascience.com/machine-learning-basics-descision-tree-from-scratch-part-i-4251bfa1b45c

Jinde, S., 2018. *support-vector-machines-svm-b2b433419d73*. Retrieved July 25, 2020, from medium.com: https://medium.com/coinmonks/support-vector-machines-svm-b2b433419d73

R-BLOGGERS.,2019. *logistic-regression-in-r-a-classification-technique-to-predict-credit-card-default*. Retrieved July 25, 2020, from r-bloggers: https://www.r-bloggers.com/logistic-regression-in-r-a-classification-technique-to-predict-credit-card-default/

Brownlee, J. ,2020. *voting-ensembles-with-python*. Retrieved JUly 25, 2020, from machinelearningmastery: https://machinelearningmastery.com/voting-ensembles-with-python/

Xiaojun Wu and Sufang Meng,2016."E-commerce customer churn prediction based on improved SMOTE and AdaBoost," *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, Kunming, 2016, pp. 1-5, doi: 10.1109/ICSSSM.2016.7538581.

Acharya, Anal, and Devadatta Sinha.,2014. "Early prediction of students performance using machine learning techniques." *International Journal of Computer Applications*