Machine Learning Framework for Prediction of Admission in Engineering College

Dr. B. Subba Reddy[1], B. Pallavi[2], B. Shruthi[2], T. Rohini[2]

[1]Professor & HOD, [2]UG Student, [1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (UGC-Autonomous), Maisammaguda, Secunderabad, Telangana, India

ABSTRACT

Utilizing machine learning (ML), enormous amounts of information can be re-evaluated and discover patterns that might not be immediately noticeable or recognizable to humans. ML strategies have increasingly been used to assess educational data such as student class performance. In the pursuit of the academic well-being of students, the utilization of neoteric technologies such as data mining, data management, and ML has increased. The idea of extracting undisclosed information from many raw databases is called data mining. Consequently, the exploration of knowledge acquisition relates to predictive ML models and subsequent decision-making. State-of-the-arts of data mining and ML have become more acceptable in predicting student examination evaluations such as grades, achievement, etc. Generally, conventional data mining for educational data analysis aimed at solving problems in an educational context can be described as educational data mining. Currently, intelligent computer-based methods such as artificial intelligence and data mining have been successfully applied to improve people's daily lives. A couple of million students participate in the bachelor's entrance examination at government-run universities each year in India. Nevertheless, only a few thousand are admitted after this competitive examination. In some cases, it was observed that many candidates struggled hard during this period. However, they could not get admission to a public university in India, resulting in an unforeseeable future. Numerous factors could be behind their unsuccessful admission to a public university, such as family circumstances, frustration, admission test anxiety, etc. However, Indian students need admission to a public university because private university education costs are too high for middle-income and low-income families. In contrast, the government primarily covers public university costs. Therefore, this project implements the prediction of college admission for engineering or college students using machine learning algorithm.

Keywords: Education, admission prediction, machine learning, supervised learning.

1. INTRODUCTION

Today, all higher education institutions, especially computer and engineering colleges, face challenges in the admissions process. Each university should strive for an admissions system based on valid and reliable admissions criteria that select candidates likely to succeed in its programs. In addition, each university should use the best possible techniques for predicting applicants' future academic performance before admitting them [1]. This would support university decision makers as they set efficient admissions criteria. However, most higher education institutions face challenges when they analyse their large educational databases to predict students' performance [2]. This is because they use only conventional statistical methods rather than new and efficient predictive techniques such as Educational Data Mining (EDM), which is the most popular technique to evaluate and predict students' performance. EDM is the process of extracting useful information and patterns from a huge educational database [3], which can then be used to predict students' performance. As a result of better information, student performance can be more effectively improved through more effective strategic programs.

Today, all higher education institutions face difficulties in the admission process. Every college ought to make a choice in its admission form which is dependent on legitimate and credible admissions procedures that select the student candidates prone to prevail in its programs. Furthermore, every college should use the most ideal [4] methods for foreseeing candidates' future academic performance before conceding them. This result would uphold college chiefs as they set effective admissions criteria. Recently, educational data mining (EDM), a subfield of datum mining, has appeared that has practical experience in educational datum that is the most common method to value and foresee students' execution. EDM is the way toward extricating helpful information and examples from an enormous educational database [5], which would then be able to be used to predict students' performance.

2. LITERATURE SURVEY

Mengash et. al [6] focuses on ways to support universities in admissions decision making using data mining techniques to predict applicants' academic performance at university. A data set of 2,039 students enrolled in a Computer Science and Information College of a Saudi public university from 2016 to 2019 was used to validate the proposed methodology. The results demonstrate that applicants' early university performance can be predicted before admission based on certain pre-admission criteria (high school grade average, Scholastic Achievement Admission Test score, and General Aptitude Test score). The results also show that Scholastic Achievement Admission Test score is the pre-admission criterion that most accurately predicts future student performance. Therefore, this score should be assigned more weight in admissions systems. We also found that the Artificial Neural Network technique has an accuracy rate above 79%, making it superior to other classification techniques considered (Decision Trees, Support Vector Machines, and Naïve Bayes). Acharya et. al [7] present a Machine Learning based method where we compare different regression algorithms, such as Linear Regression, Support Vector Regression, Decision Trees and Random Forest, given the profile of the student. We then compute error functions for the different models and compare their performance to select the best performing model. Results then indicate if the university of choice is an ambitious or a safe one.

Tsang et. al [8] proposes a novel machine learning methodology: entropy regularization with ensemble deep neural networks (ECNN), which simultaneously provides high predictive performance of hospitalization of patients with dementia whilst enabling an interpretable heuristic analysis of the model architecture, able to identify individual features of importance within a large feature domain space. Experimental results on health records containing 54,647 features were able to identify 10 event indicators within a patient timeline: a collection of diagnostic events, medication prescriptions and procedural events, the highest ranked being essential hypertension. The resulting subset was still able to provide a highly competitive hospitalization prediction (Accuracy: 0.759) as compared to the full feature domain (Accuracy: 0.755) or traditional feature selection techniques (Accuracy: 0.737), a significant reduction in feature size. The discovery and heuristic evidence of correlation provide evidence for further clinical study of said medical events as potential novel indicators. There also remains great potential for adaption of ECNN within other medical big data domains as a data mining tool for novel risk factor identification.

El-Bouri et. al [9] presents a deep learning method of predicting where in a hospital emergency patients will be admitted after being triaged in the Emergency Department (ED). Such a prediction will allow for the preparation of bed space in the hospital for timely care and admission of the patient as well as allocation of resource to the relevant departments, including during periods of increased demand arising from seasonal peaks in infections. Methods: The problem is posed as a multi-class classification into seven separate ward types. A novel deep learning training strategy was created that combines learning via curriculum and a multi-armed bandit to exploit this curriculum post-initial training. Results: We successfully predict the initial hospital admission location with area-under-receiver-operating-curve (AUROC) ranging between 0.60 to 0.78 for the individual wards and an overall maximum accuracy of 52% where chance corresponds to 14% for this seven-class setting. Our proposed network was able to interpret which features drove the predictions using a `network saliency' term added to the network loss function. Conclusion: We have proven that prediction of location of admission in hospital for emergency patients is possible using information from triage in ED. We have also shown that there are certain tell-tale tests which indicate what space of the hospital a patient will use. Significance: It is hoped that this predictor will be of value to healthcare institutions by allowing for the planning of resource and bed space ahead of the need for it. This in turn should speed up the provision of care for the patient and allow flow of patients out of the ED thereby improving patient flow and the quality of care for the remaining patients within the ED.

Hu et. al [10] proposed a stacking ensemble model with direct prediction strategy to predict the daily number of CVDs admissions using HAs data, air pollution data, and meteorological data. The sequential forward floating selection method with early stopping was applied for feature selection. Five machine learning models, including linear regression (LR), support vector regression (SVR), extreme gradient boosting (XGBoost), random forest (RF), and gradient boosting decision tree (GBDT), were utilized as base learners to construct the stacking model. We compared the performance of the proposed stacking model with the five base learners in three datasets. The experimental results indicated that our model performed best in three datasets under four evaluation criteria, including mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R 2). Particularly, in the CVDs dataset, the MAPE is 15.103 for LR, 11.862 for SVR, 10.571 for XGBoost, 10.378 for GBDT, 10.333 for RF, and 9.679 for the stacking model. Compared with the best base learner RF, the MAPE, RMSE, and MAE of the stacking model decreased by 6.3%, 7.4%, and 6.3%, respectively, and the R 2 improved by 1.7%. It is evident that the proposed stacking model can effectively forecast the daily number of hospitalizations for CVDs and provide decision support for hospital managers.

Fadil et. al [11] aims to apply and analyse the accuracy of the least square method to predict the number of prospective students. This method is very suitable to be used to predict the magnitude of variables in time series. The number of applicants entering each department is entered based on a period, namely Information Engineering, Information Systems, and Information Management. Existing data amounted to five time periods, consisting of five years of new student admission data. To calculate the error rate, mean absolute deviation, mean square error, and mean absolute percentage error is used. With prediction results for 2020 obtained, are 225 people for Informatics Engineering, 82 Information Systems, and 11 Informatics Management. Besides, the results of the measurement of forecasting suitability with these methods include Informatics Engineering as much as 0.79 % , Information Systems 0.8%, and Information Management as much as 3.4%. The results of this study prove that to predict the number of registrants on the admission of new students using the least square method.

Bitar et. al [12] implement and compare several supervised predictive analysis methods on a labeled dataset based on real applications from the prestigious university of UCLA; Regression, classification, and Ensemble methods are all the supervised methods that are to be employed for prediction. The dataset relies profoundly on the academic performance of the applicants during their undergrad years.

The coefficient of determination, as well as precision and accuracy, are the measures used to compare the different models. All predictive methods proved to show accurate results, however; certain methods proved to be more promising than others were. Predictions were obtained within short time frames, which in turn will cut down the time in the admission process.

3. PROPOSED SYSTEM

3.1 Dataset description

GRE Score - Out of 340

TOEFL Score - Out of 120

University Rating - Between 1 to 5 (5 being the best)

SOP - Between 1 to 5 (5 being the best)

LOR - Between 1 to 5 (5 being the best)

CGPA - Out of 10

Research - 1 if student has research experience, else 0

Chance of Admit - Probability of getting accepted into graduate program

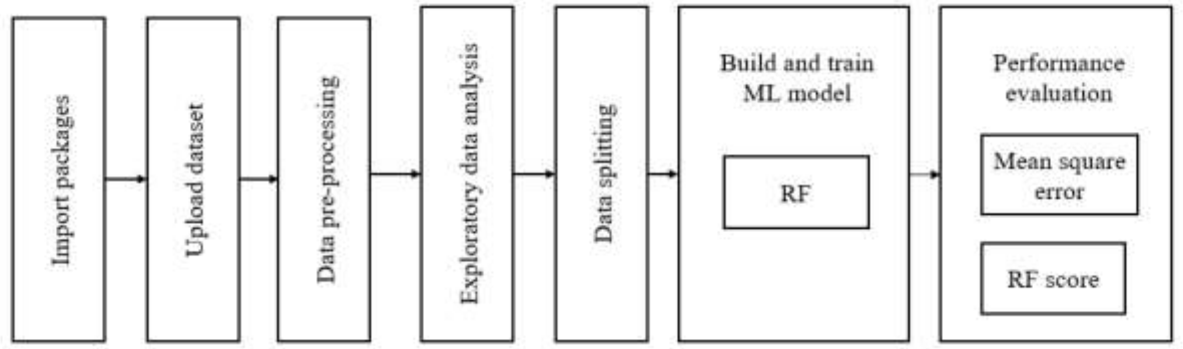


Fig. 1: Block diagram of proposed system.

3.2 Data Preprocessing in Machine learning

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

3.2.1 Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So, we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

3.3 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

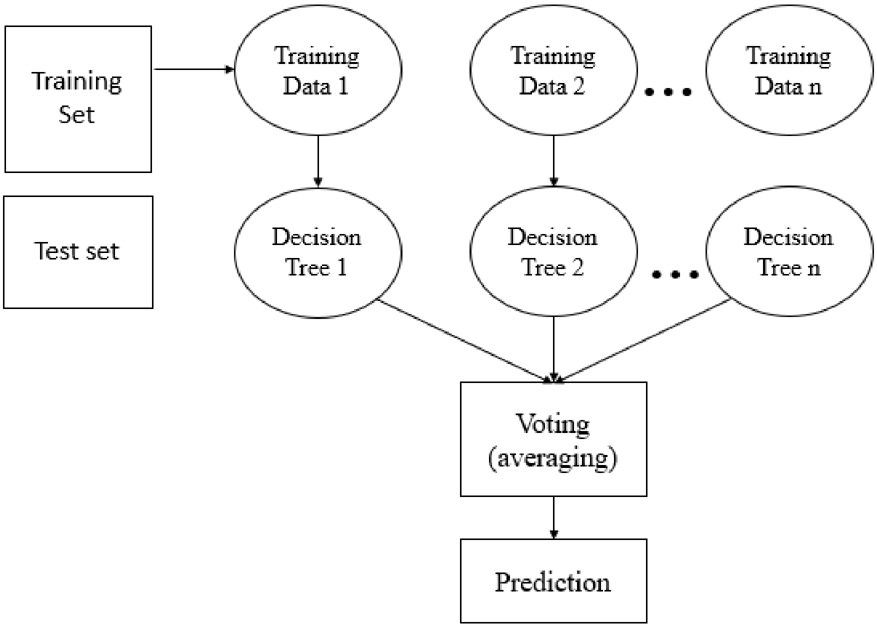


Fig. 2: Random Forest algorithm.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

3.3.1 Important Features of Random Forest

- **Diversity-** Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality-** Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization-**Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split-** In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability-** Stability arises because the result is based on majority voting/ averaging.

3.3.2 Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm.

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.3.4 Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

- **Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

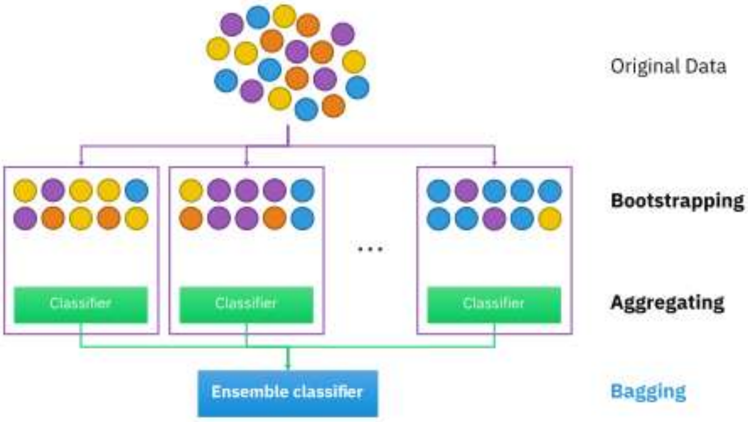


Fig. 3: RF Classifier analysis.

- **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

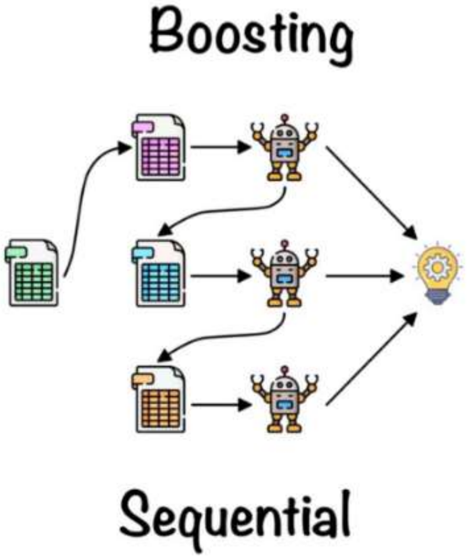


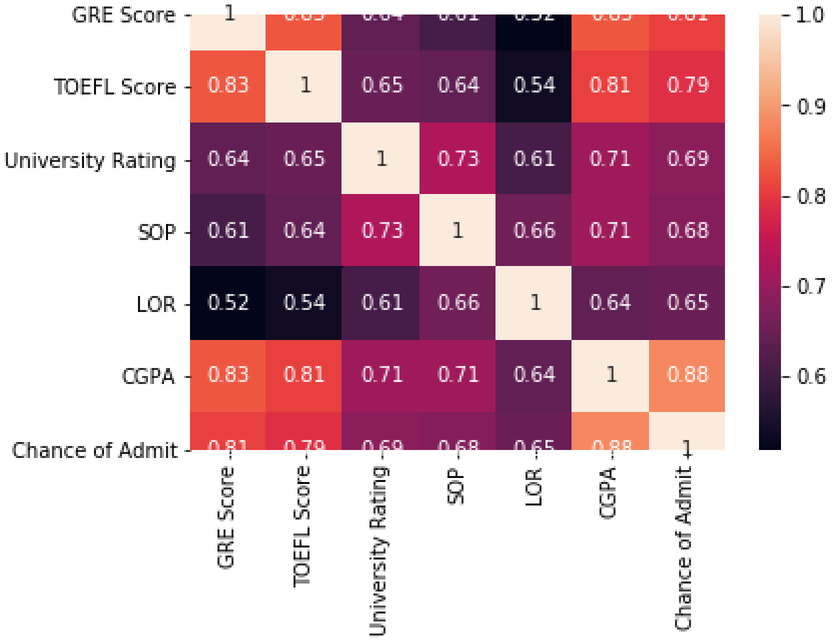
Fig. 4: Boosting RF Classifier.

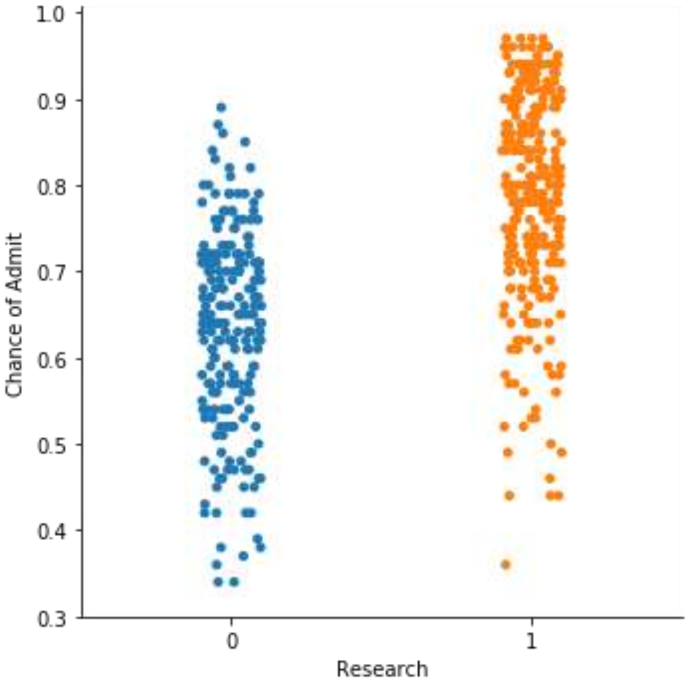
Advantages of proposed system

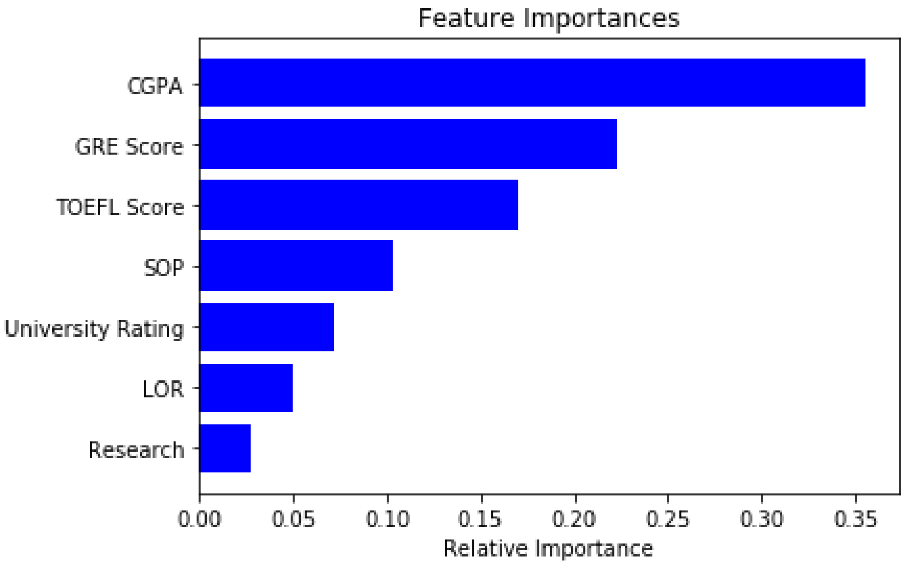
- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

4. RESULTS AND DISCUSSION

4.1 Correlation matrix







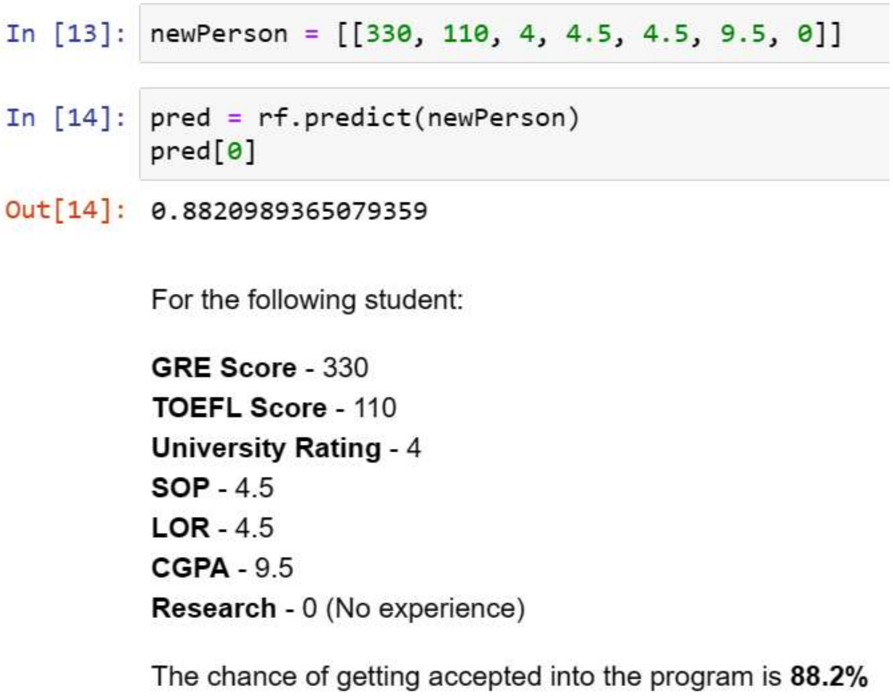
Train MSE: 0.035577157557505 21

Test MSE: 0.05607613271535168

Train R^2: 0.9354862330956815

Test R^2: 0.8486805018919078

4.2 Prediction on test data

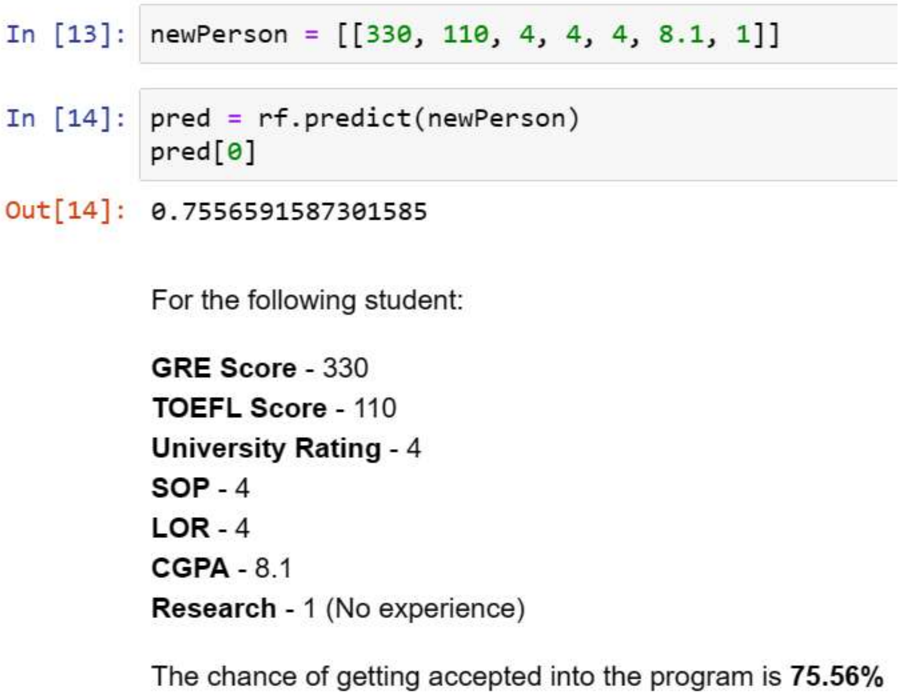


```
In [13]: newPerson = [[330, 110, 4, 4.5, 4.5, 9.5, 0]]

In [14]: pred = rf.predict(newPerson)
         pred[0]

Out[14]: 0.8820989365079359
```

For the following student:

GRE Score - 330
TOEFL Score - 110
University Rating - 4
SOP - 4.5
LOR - 4.5
CGPA - 9.5
Research - 0 (No experience)

The chance of getting accepted into the program is **88.2%**



```
In [13]: newPerson = [[330, 110, 4, 4, 4, 8.1, 1]]

In [14]: pred = rf.predict(newPerson)
         pred[0]

Out[14]: 0.7556591587301585
```

For the following student:

GRE Score - 330
TOEFL Score - 110
University Rating - 4
SOP - 4
LOR - 4
CGPA - 8.1
Research - 1 (No experience)

The chance of getting accepted into the program is **75.56%**

5. CONCLUSION

Academic performance is the primary concern for most colleges in most countries. There are extensive quantities of data generated in learning systems. This data holds hidden knowledge that could be used to heighten the students' academic success. In this research, a suggested model of student achievement prediction was constructed totally on ensemble methods. The predictive model by classifiers random forest method (bagging and boosting) deal with raising these classifiers' benefits. The retrieved results expose that there is an enhancement in these models over the conventional classifiers. Then, the proposed method combines two different classifiers with one of the bagging or boosting process. This method gave better results than previous methods that contribute to the growth of the accomplishment of students and educational systems. We will assemble information from numerous understudies of different instructive organizations and use some great data mining techniques to deliver a substantial yield. This project empowers instructional frameworks, foundations, understudies, and instructors to fortify their performance.

REFERENCES

[1] L. H. Son and H. Fujita, "Neural-fuzzy with representative sets for prediction of student performance," Applied Intelligence, vol. 49, no. 1, pp. 172–187, 2019.

[2] S. Bharara, S. Sabitha, and Bansal, "Application of learning analytics using clustering data mining for students' disposition analysis," Education and Information Technologies, vol. 23, no. 2, pp. 957–984, 2018.

[3] C. M. D. Bondoc and T. G. Malawit, "Classifying relevant video tutorials for the school's learning management system using support vector machine algorithm," Global Journal of Engineering and Technology Advances, vol. 2, no. 3, pp. 1–9, 2020.

[4] K. Coussement, M. Phan, A. De Caigny, D. F. Benoit, and A. Raes, "Predicting student dropout in subscription-based online learning environments: the beneficial impact of the logit leaf model," Decision Support Systems, vol. 135, 2020.

[5] Dhankhar, K. Solanki, and A. Rathee, "'Predicting student's performance by using classification methods," Journal of advanced trends in computer science and engineering, vol. 8, no. 4, pp. 1532–1536, 2019.

[6] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," in IEEE Access, vol. 8, pp. 55462-55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[7] M. S. Acharya, A. Armaan and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862140.

[8] G. Tsang, S. -M. Zhou and X. Xie, "Modeling Large Sparse Data for Feature Selection: Hospital Admission Predictions of the Dementia Patients Using Primary Care Electronic Health Records," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 9, pp. 1-13, 2021, Art no. 3000113, doi: 10.1109/JTEHM.2020.3040236.

[9] R. El-Bouri, D. W. Eyre, P. Watkinson, T. Zhu and D. A. Clifton, "Hospital Admission Location Prediction via Deep Interpretable Networks for the Year-Round Improvement of Emergency Patient Care," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 1, pp. 289-300, Jan. 2021, doi: 10.1109/JBHI.2020.2990309.

[10] Z. Hu, H. Qiu, Z. Su, M. Shen and Z. Chen, "A Stacking Ensemble Model to Predict Daily Number of Hospital Admissions for Cardiovascular Diseases," in IEEE Access, vol. 8, pp. 138719-138729, 2020, doi: 10.1109/ACCESS.2020.3012143.

[11] I. Fadil, M. Agreindra Helmiawan and D. Indra Junaedi, "Analysis of Data Prediction Generate by Admission Student Application using Least Square Method," 2020 8th International Conference on Cyber and IT Service Management (CITSM), 2020, pp. 1-5, doi: 10.1109/CITSM50537.2020.9268840.

[12] Z. Bitar and A. Al-Mousa, "Prediction of Graduate Admission using Multiple Supervised Machine Learning Models," 2020 SoutheastCon, 2020, pp. 1-6, doi: 10.1109/SoutheastCon44009.2020.9249747.