Journal of Applied Statistics & Machine Learning
Vol. 3, Nos. 1-2, 2024, pp. 1-9, ISSN: 2583-2891
© Black Rose Publication
URL : www.blackroseindia.com/jasml

Comparative Analysis of Machine Learning Classifiers for Educational Institution Admissions

Jassimar Singh
Student BSC (Computer Science), Buffalo University NY USA

Received : 13 February 2024 • **Revised :** 11 March 2024;
Accepted : 20 March 2024 • **Published :** 29 December 2024

Abstract: There has been a comparative analysis done on fifteen classifiers (Models). When it came to predicting the students' admission status (binary), the Random Forest classifier (RF) was ranked No 1. Metrics including Accuracy (95.5%), Precision (97.2), Recalls (93.2) and F1-Score (95.2) were used to assess the models' performance on test/validated data. Accuracy (95.5%), Precision (95.3), Recalls (94,9) and F1 (95.0). The area under the ROC curve was also very high (0.99), suggesting that the RF model is the most effective and perfect model and that it can be used to predict the admission status for new applicants to educational Institutions.

Introduction

Artificial Intelligence (AI) emerges as a multidisciplinary domain in computer science, aiming to replicate human intelligence. This expansive field merges insights from various disciplines, including mathematics, cognitive science, neuroscience, and engineering, to create systems that can perform tasks typically requiring human intellect. AI systems are designed to mimic cognitive functions such as learning, reasoning, problem-solving, and understanding language. Just as our brains learn from experiences, AI learns from data, progressively improving its performance and accuracy over time through iterative processing. A key component of AI is Machine Learning (ML), a branch that empowers computers to learn from data and make decisions without being explicitly programmed for every task. Machine Learning is transforming industries by enabling systems to automatically learn and adapt from experience, much like humans do. At the heart of ML is the ability to recognize patterns and make data-driven decisions. This

TO CITE THIS ARTICLE

Singh, J. (2024). Comparative Analysis of Machine Learning Classifiers for Educational Institution Admissions, *Journal of Applied Statistics & Machine Learning,* 3(1-2), pp. 1-9.

capability is leveraged in numerous applications, from predicting stock market trends to recommending products on e-commerce sites. Machine Learning itself is an umbrella term encompassing several specialized fields, including Deep Learning, Reinforcement Learning, and Generative AI. Deep Learning is a subset of ML that employs neural networks with many layers to analyses vast amounts of data. These neural networks simulate the workings of the human brain, allowing the system to understand complex patterns and features within the data. This technology is particularly effective in tasks such as image and speech recognition, where it has achieved remarkable breakthroughs. Reinforcement Learning, another branch of ML, focuses on training algorithms through a system of rewards and penalties. In this paradigm, an agent learns to make decisions by performing certain actions and receiving feedback from the environment. This approach is widely used in areas like robotics and gaming, where an AI can learn optimal strategies by exploring different actions and outcomes. Generative AI, which includes models like ChatGPT, represents a cutting-edge advancement in AI. These models are designed to generate new content by learning from vast datasets. For instance, ChatGPT is trained on extensive text data, enabling it to produce human-like responses in natural language. Applications of Generative AI range from chatbots that handle customer inquiries to tools that assist in creative writing and content creation.

Within the realm of Machine Learning, two fundamental categories are Supervised Learning and Unsupervised Learning. Supervised Learning involves training models on labeled data, where each input is paired with the correct output. This method allows the model to learn the relationship between inputs and outputs, making it possible to predict outcomes for new, unseen data. It is commonly used for classification tasks, such as spam detection in emails, and regression tasks, like predicting house prices. Unsupervised Learning, in contrast, deals with unlabeled data. Here, the model tries to identify patterns and structures within the data without any prior knowledge of the correct outputs. This approach is often employed in clustering tasks, such as customer segmentation, where the goal is to group similar items together.

Our study focuses on Supervised Learning, specifically a binary classification problem where the target variable is the admission status of students, either “admitted” or “not admitted.” The primary objective is to predict the admission category of students based on a set of independent features. These features include GRE scores, TOEFL scores, university rating, statement of purpose, letter of recommendation, CGPA, research experience, and admission status.

To accomplish this, we apply 15 different Machine Learning models of classification. These models range from traditional algorithms like Logistic Regression and Decision Trees to more advanced techniques such as Support Vector Machines, Random Forests, and Gradient Boosting. Each model has its strengths and is suitable for different types of data and classification problems.

By exploring a diverse set of models, we aim to identify the most effective approach for predicting student admissions. This comprehensive analysis not only enhances our understanding of the factors influencing admission decisions but also provides valuable insights into the performance of various Machine Learning techniques. Ultimately, the goal is to develop a robust predictive model that can assist educational institutions in making data-driven admission decisions, ensuring a fair and efficient selection process for prospective students.

Methods and Material

Machine learning Models generally involve the following steps :

1. **Important Libraries :** Downloads necessary Libraries such as Pandas, NumPy , Matplotlib , Seaborn sklearn, dataprep and pycaret.
2. **Data Collection and preprocessing** : The data for the study was down loaded from Kaggle dataset1 . Data set comprises of 400 examples of independent variables/features namely Gre_Score, Toffel Score ,University Rating ,Statement of Purpose ,Letter of Recommendation, CGPA, Research Experience and Chance of admission . Sample size of examples was enhanced to 800 by using bootstrap method. The target binary variable i.e., Admission Status was to admit the student if chance of Admission is >= 0.80 otherwise not Admit . Preprocessing of data was carried out using the dataprep library2. It was observed that data is free from outliers (Figure 1).
3. **Data Splitting** : we split the dataset into training (80 %) and testing (20 %) sets. The testing set was used to evaluate the performance of the Models.
4. **Model Selection**: Since the target variable is binary, we selected 15 classifiers algorithms by using pycaret library and chose the random forest classifier as its performance was significantly higher among all the 15 models (Table 1) considered. The Random Forest algorithm was introduced by Leo Breiman8 in 2001 as an extension of bagged decision trees. It has since become a fundamental algorithm in

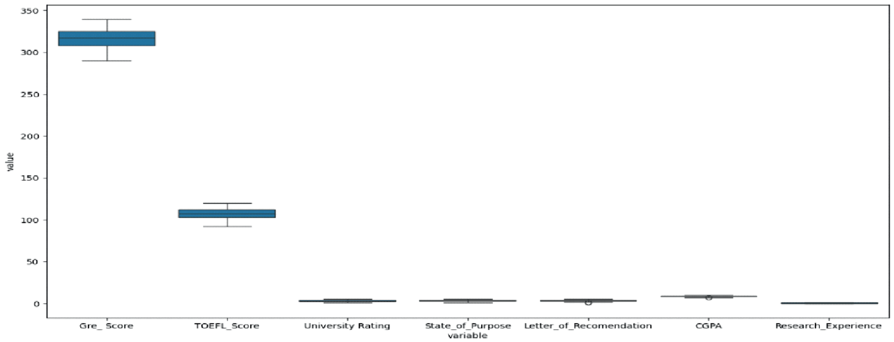


Figure 1: Boxplot of data by the seven categorical features

machine learning. Random Forest's feature importance scores helped the users to understand the relative importance of features in predictive modelling[9] . Another invariant of RF is Extra Tree[10] classifier has been explored to further enhance performance and reduce overfitting. RF model was found to be not overfitting as confirmed by slight variation between training and testing score and out of bag error is only 3.14 %. Extra Tree ranked 2nd as for the performance is concerned .

5. **Model Training** : 15 machine learning models for classification were trained on training data set on the chosen features
6. **Model Hyperparameter tuning** : RF was not tuned as it performs very well with default hyperparameters .
7. **Model Evaluation** : Evaluate the model's performance on testing data using appropriate evaluation metrics such as accuracy, precision, recall , F1-score and ROC-AUC
8. **Model Comparison** : Compared the performance of the 15 models to determine which one provide better classification results .
9. **Deployment of Model** : Better-performing model i.e. RF model was saved using joblibs library for futures use . However, it can be deployed either on as a web application, API, or integrate it into your university's admission system to make real-time predictions for new applicants.

Results and Discussion

Python Tool package such as Pandas, NumPy, Matplotlib, SkLearn, seaborn, dataprep and Pycaret libraries has been used to apply different data mining algorithms with their parameters and attempted to select best algorithm so we can use that to make prediction .The FIG 2 shows that there are the data is slightly imbalance as the number of students in admit category and not admit are 460 and 340 respectively. SMOTE method was used to take care of Imbalance data.

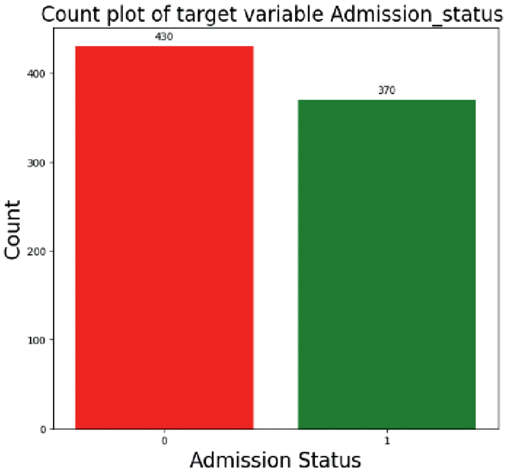


Figure 2

A confusion matrix is a 2X 2 Table 1 that is used to evaluate the performance of a machine learning model . It shows the number of correct and incorrect predictions made by the model when compared to the actual outcomes.

Classification	*Predicted Positive*	*Predicted Negative*	*Total*
Actual Positive	True Positive (TP)	False Positive (FP)	**TP+FP**
Actual Negative	False Negative (FN)	True Negative (TN)	**FN+TN**
Total	TP+FN	FP +TN	TP+TN+ FP+FN

Accuracy, Precision, Recall and F1 -Score : The performance of the ML models was evaluated using fours metrics namely Accuracy, Precession, Recall and F1 Score on training as well on test data.

Accuracy : is the measure of how often the model is correct in its predictions. But it may not be the best metric for imbalanced datasets, where the number of positive and negative samples are significantly different.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision : Precision is the measure of the proportion of true positives among all the samples that the model predicted as positive.

Precision = (T P)/ Predicted Positive (TP+FN)

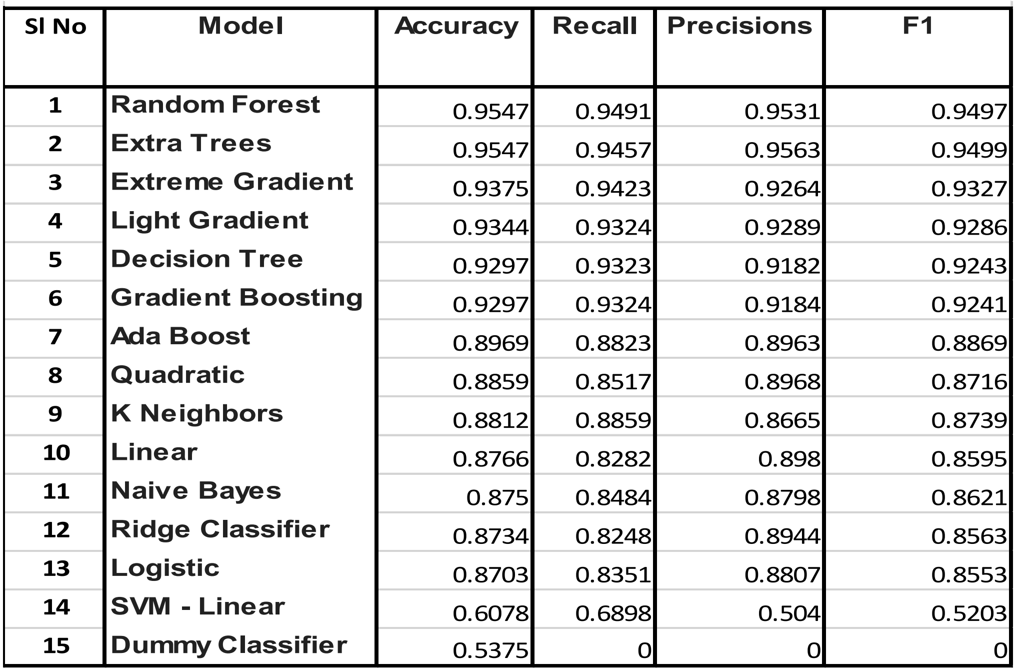
Recall : is the measure of the proportion of true positives among all the actual positive samples. Recall measures the model's ability to correctly identify all positive samples.

Recall =TP/actual Positive (TP+FP)

The F1-score : is the harmonic mean of precession and recall

Accuracy, Precision, Recall and F1-Score for 15 classifiers on training data are shown table 2 . It is evident from table 2 that that Random Forest model performed best as compared to all others models as Accuracy (95.5%), Precision (95.3), Recalls (94,9) and F1(95.0) respectively is relatively high than any other models.

Table 2: Comparative Performance of the Machine learning Classifications Models on Training Data



Sl No	Model	Accuracy	Recall	Precisions	F1
1	Random Forest	0.9547	0.9491	0.9531	0.9497
2	Extra Trees	0.9547	0.9457	0.9563	0.9499
3	Extreme Gradient	0.9375	0.9423	0.9264	0.9327
4	Light Gradient	0.9344	0.9324	0.9289	0.9286
5	Decision Tree	0.9297	0.9323	0.9182	0.9243
6	Gradient Boosting	0.9297	0.9324	0.9184	0.9241
7	Ada Boost	0.8969	0.8823	0.8963	0.8869
8	Quadratic	0.8859	0.8517	0.8968	0.8716
9	K Neighbors	0.8812	0.8859	0.8665	0.8739
10	Linear	0.8766	0.8282	0.898	0.8595
11	Naive Bayes	0.875	0.8484	0.8798	0.8621
12	Ridge Classifier	0.8734	0.8248	0.8944	0.8563
13	Logistic	0.8703	0.8351	0.8807	0.8553
14	SVM - Linear	0.6078	0.6898	0.504	0.5203
15	Dummy Classifier	0.5375	0	0	0

The performance of the Random Forest classifier was also evaluated on test data(table 3 and confusion matrix)

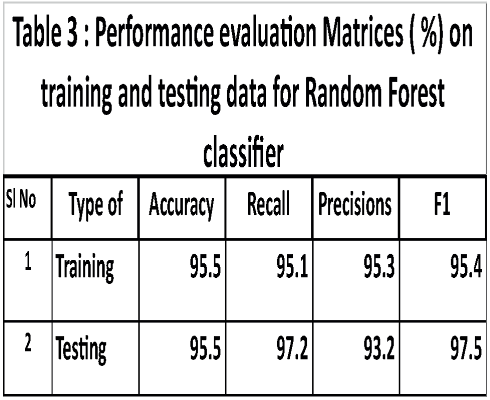


Table 3 : Performance evaluation Matrices (%) on training and testing data for Random Forest classifier

Sl No	Type of	Accuracy	Recall	Precisions	F1
1	Training	95.5	95.1	95.3	95.4
2	Testing	95.5	97.2	93.2	97.5

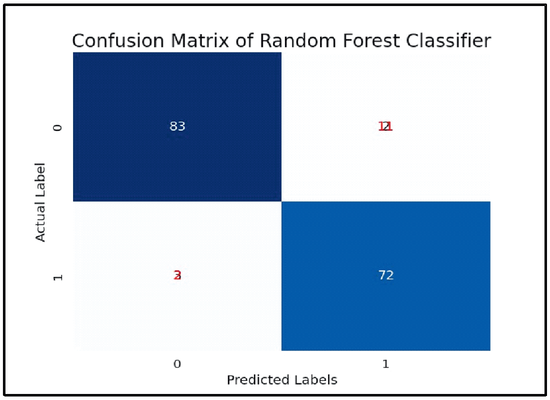
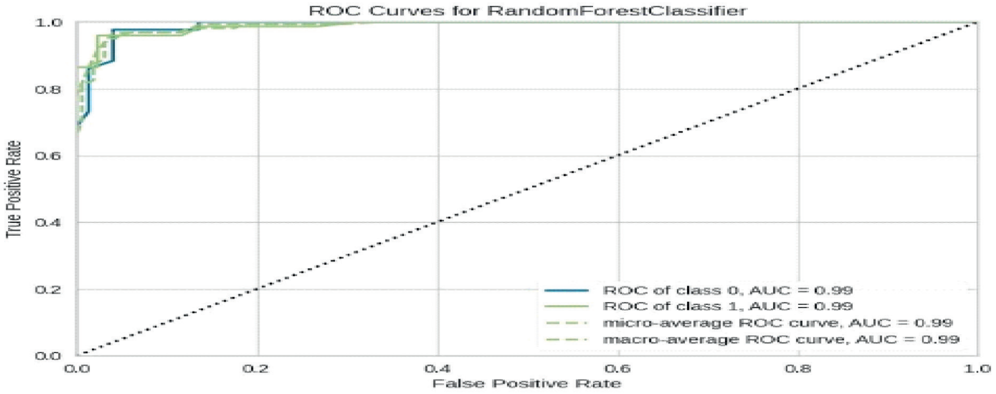


Table 3 and Confusion shown above clearly indicates that four metrics has high rate and very close to each other on test as well as on training data indicating there is no overfitting.

ROC of random Forest Classifier: The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance, plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The ROC-AUC is especially useful for evaluating models on imbalanced datasets, providing a single metric to compare different classifiers' effectiveness across all classification thresholds.



- The area under the curve (AUC) for a Receiver Operating Characteristic (ROC) curve is a measure of a classifier's performance. For ROC class 1(admitted) and class not admitted (0), the area under the curve was found to be 99 Percent. It implies that an AUC of 0.99 indicates that the classifier has excellent performance.

It means that there is a 99% chance that the model will correctly distinguish between a randomly chosen positive instance (class 1, admitted) and a randomly chosen negative instance (class 0, not admitted).

- High Discriminatory Power: This high AUC value signifies that the classifier is very effective at distinguishing between the two classes. It suggests that the model makes a very few errors in classification.

Near Perfect Classification: An AUC close to 1 indicates near-perfect classification. In our study the AUC is to close to 1, the better the model is at predicting 1s as 1s and 0s as 0s. An AUC of 0.99 means that the model is almost perfect in its predictions.

Thus, we find that performance of the Random Forest classifier not only achieved rank one among 15 classifiers models considered but also found to be perfect model for prediction.

Among the Seven Features considered CGPA (30%) is found to be (Figure 3) most important feature followed by Gre- Score (22%) to classify the students.

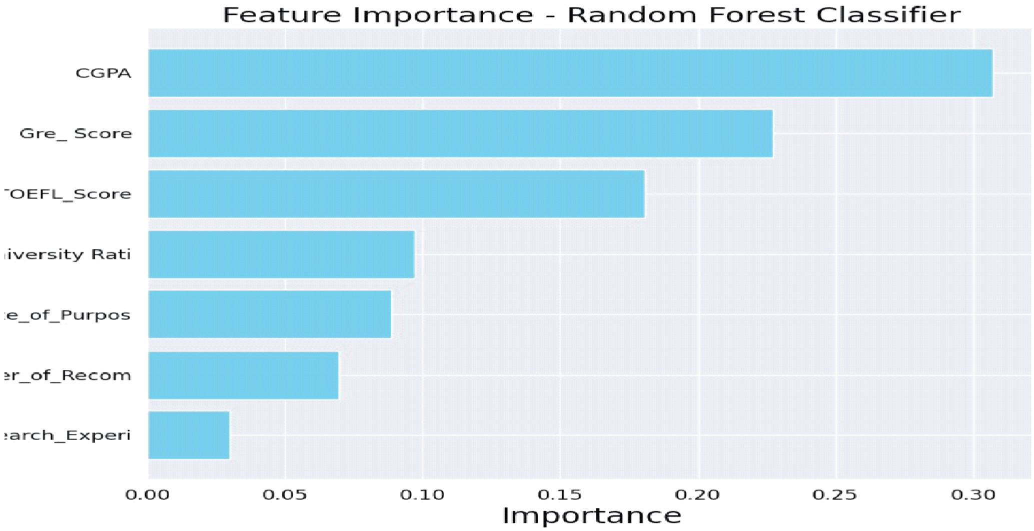


Figure 3

Features importance of the RF model is shown in figure 3 which indicates that CGPA score was most and research experience is least important features among the seven features considered .

The RF model was deployed of the new data and performing very well to predict the admission status of new students .

References

Kaggle data set data-for-admission-in-the-university https://www.kaggle.com/datasets/ashaydattatraykhare/data-for-admission-in-the-university

dataprep is a preprocessing library(https://pypi.org/project/dataprep/)

sklearn documentation for classification (https://scikit-learn.org/stable/supervised_learning.html)

NumPy documentation (https://numpy.org/doc/1.26/)

Matplotlib for plotting the graph (https://matplotlib.org/stable/index.html)

seaborn: statistical data visualization (https://seaborn.pydata.org/)

pycaret documentation for classification (https://pycaret.readthedocs.io/en/stable/api/classification.html)

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Liaw, A., & Wiener, M. (2002). Classification and regression by Random Forest. R News, 2(3), 18-22.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.