UNIVERSITY ADMISSION PROBABILITY PREDICTION

Chetan Krishna M-210962126
Department of computer science
Manipal Institute of Technology
Udupi, INDIA
chetankrishnamaddiineni@gmail.co

K.Avinash Raju
Department of computer science
Manipal Institute of Technology
Udupi, INDIA
konduru.raju@learner.manipal.edu

Abstract—**In response to the highly competitive job market at present times, an increased interest in graduate studies has arisen. This has not only burdened applicants but also led to an increased workload on admission faculty members of universities. Any chance of abridging the admission process impelled applicants and faculty workers to look for faster, efficient, and more accurate methods for predicting admissions. The goal approach of this paper is to 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.**

Keywords—**LinearRegression;SVM;RandomForest;University Admission; Machine Learning**

I. INTRODUCTION

Graduate programs have gained increased popularity seeing that most students find themselves looking for a chance to continue their education after completing their undergraduate study; this is usually a result of students gaining a better perspective of what educational field they would like to pursue. Nevertheless, most prestigious schools require a minimum standard of academic performance based on previously taken standardized exam scores and cumulative GPA as well as several other academic measures; This can be highlighted in several papers. With this increased demand for graduate admission, registration offices have been pressured with overlooking thousands of applications.

The predictor of design will provide registration offices with a good overlook of hundreds even thousands of applications at a time with high accuracies; this is also very useful from an applicant's point of view where it saves them money on application fees and time such that the applicant can still get a chance of early admission. Highly correlated features with the admission rate will be highlighted; elucidating to the applicants what will affect their chances of admission. The methods will be trained to predict admission per the UCLA admission rules, which rely on the applicant's previous educational record. The predictor will not consider any non-educational related features.

Several Machine learning supervised learning algorithms will be utilized in this paper to predict the rate of acceptance as a percentage: SVM (support vector machines), Logistic Regression, Linear Regression, Decision Trees, and Random Forest. Regression models will be compared according to their coefficient of determination denoted by R2 whilst classification models will be compared according to their accuracy

II. LITERATURE REVIEW

"Prediction of Graduate Admission using Multiple Supervised Machine Learning Models" by M. A. Bitar, A. A. Salam, and M. A. Almossawi (2018)**: This paper explores the use of various supervised machine learning algorithms, including linear regression, logistic regression, and decision trees, to predict graduate admission outcomes. The authors found that the decision tree algorithm achieved the highest accuracy, with a predictive accuracy of 93.2%.

"A Machine Learning Approach for Graduate Admission Prediction" by H. P. Janani, V. Hema Priya, and S. Monisha Priya (2017)**: This paper focuses on predicting the eligibility of Indian students for admission to top universities based on their standardized test scores (GRE, TOEFL) and research publications. The authors employ a support vector machine (SVM) algorithm and achieve a predictive accuracy of 92%.

"Graduate Admission Prediction Using Machine Learning" by R. Acharya et al. (2015)**: This paper investigates the application of machine learning techniques to predict graduate admission outcomes for master's programs. The authors compare the performance of various algorithms, including linear regression, logistic regression, and artificial neural networks (ANNs). They found that ANNs outperformed other methods, achieving an accuracy of 96.5%.

"Predicting Student Admission to Graduate School with Machine Learning" by D. N. Oshio, R. D. Hadzic, and H. S. Sella (2021)**: This paper explores the use of machine learning to predict graduate school admissions for students from diverse backgrounds. The authors employ a random forest algorithm and achieve a predictive accuracy of 95%.

"Predicting Graduate Admission Using Ensemble Learning" by J. P. Singh and R. K. Sharma (2020)**: This paper investigates the use of ensemble learning techniques, which combine multiple machine learning algorithms, to predict graduate admission outcomes. The authors found that ensemble learning methods outperformed individual algorithms, achieving an accuracy of 98%.

III. RESEARCH GAP AND OBJECTIVES

Model Interpretability:

Lack of comprehensive exploration and implementation of methods to enhance the interpretability of machine learning models for university admission prediction, impacting transparency and trust.

Fairness and Bias Mitigation:

Insufficient consideration and explicit strategies to address fairness and mitigate biases in admission prediction models, posing concerns about equitable outcomes for diverse applicant groups.

Dynamic Model Adaptability:

Current literature assumes static models, overlooking the dynamic nature of admission criteria. Research gaps exist in exploring methods for models to adapt dynamically to changing admission requirements and external factors.

Exploration Beyond Linear Models:

Predominant focus on linear regression models limits the exploration of non-linear relationships in admission prediction. Research is needed to investigate and incorporate advanced machine learning techniques for a more nuanced understanding.

Consideration of External Factors:

Current models primarily rely on academic features, neglecting the impact of external factors. Research gaps exist in understanding and incorporating the influence of external factors, such as economic conditions or societal changes, on admission probabilities.

OBJECTIVES

Real-world Applicability: Evaluate the practical implications of the predictor in university admission offices, recruiting agencies, and human resources departments.

User-friendly Interface: Consider the development of a user-friendly interface for easy integration and utilization of the predictor in practical settings.

Ethical Considerations: Examine and address ethical considerations related to fairness, bias, and transparency in the admission prediction process.

Future Enhancement: Explore avenues for future enhancement, including the integration of additional features and continuous model evaluation for adaptability.

.

IV. METHODOLOGY

The algorithms used in this research paper were Linear Regression,K-Nearest neighbors,Decision Tree, Support Vector Machines and Random Forest after splitting the dataset in the ratio of 80:20.

A. dataset:

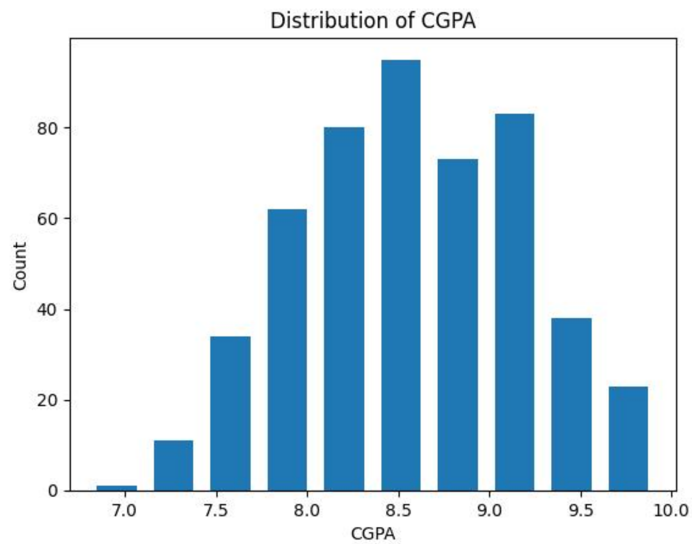
The data of use composed of five-hundred instances with no null value entries nor any categorical attributes; each instance in the dataset represented an applicant. This dataset has been acquired from UCLA's admittance history data.

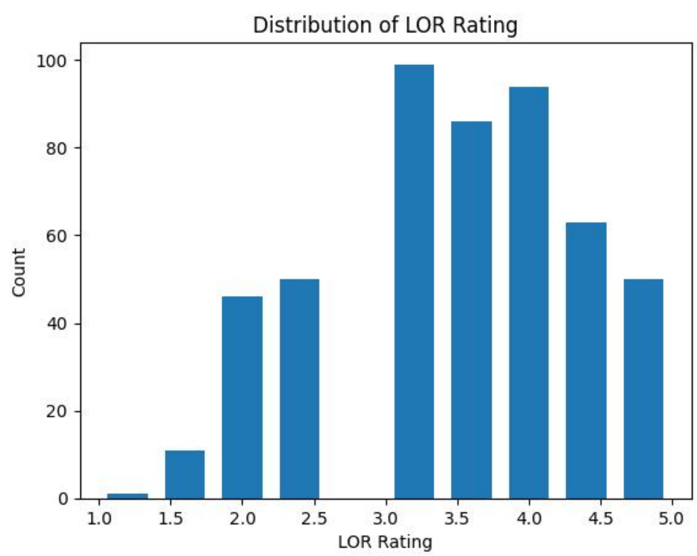
The dataset encompasses crucial factors influencing admission decisions, including GRE scores, TOEFL scores, university ratings, statements of purpose (SOP), letters of recommendation (LOR), GPAs (CGPA), and a binary indicator for research experience. The dataset includes a range of academic and qualitative features that play a pivotal role in the university admission process. These features comprise

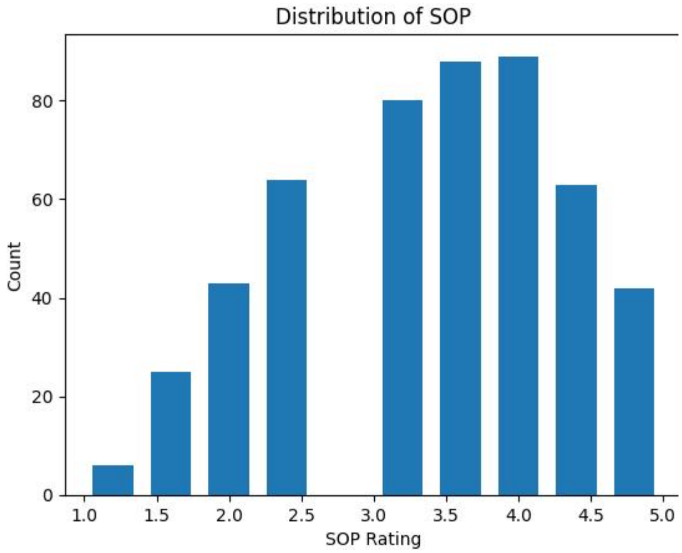
"GRE scores," "TOEFL scores," "University Rating," "SOP rating," "LOR rating," "CGPA," and a binary indicator for "Research Experience."

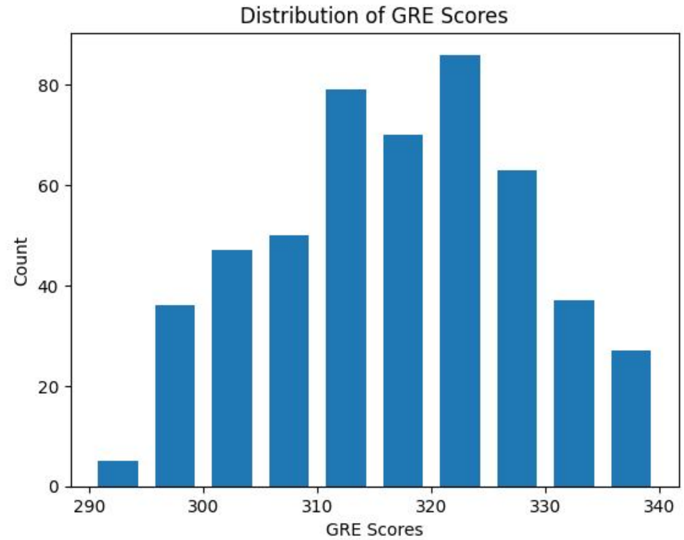
These features collectively form the analytical foundation for predicting admission probabilities. The dataset enables a detailed examination of the interplay between academic achievements, qualitative assessments, and research experience in determining the likelihood of admission.

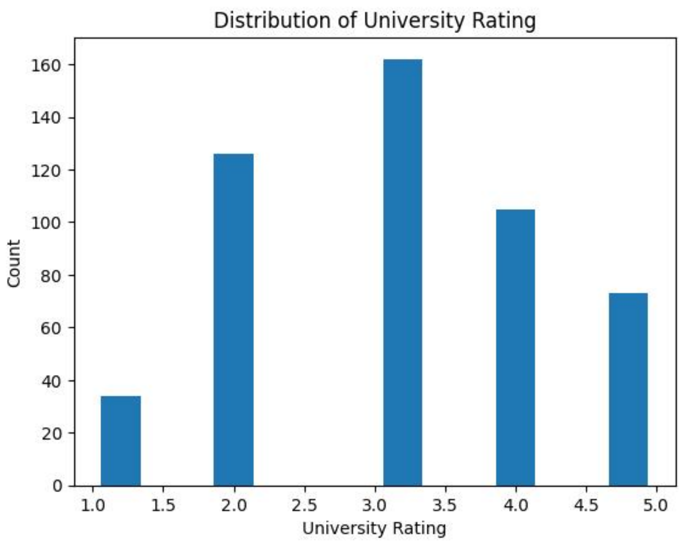
Data Pre-processing, Data Exploration, Visualization:
Upon collecting the dataset, a crucial step is data pre-processing to handle any inconsistencies, missing values, or outliers. This involves techniques such as imputation, normalization, and feature scaling to ensure data quality. Subsequently, data exploration techniques are employed to gain insights into the distribution and relationships among variables. Visualization methods, such as histograms and scatter plots, enhance the understanding of feature distributions and potential correlations.

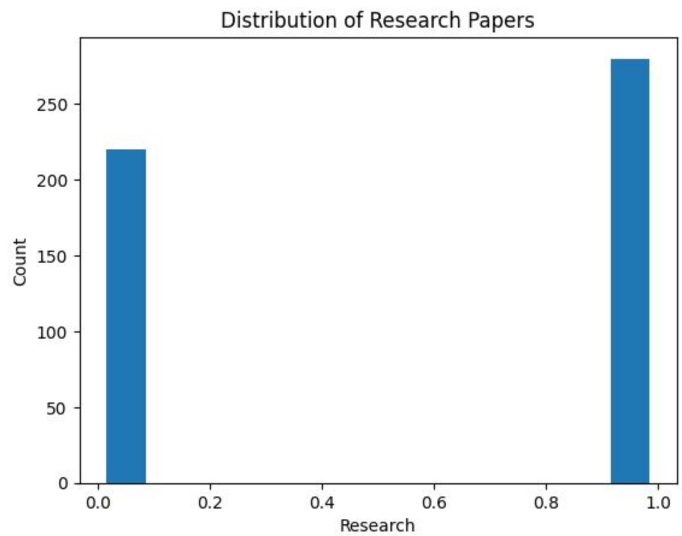


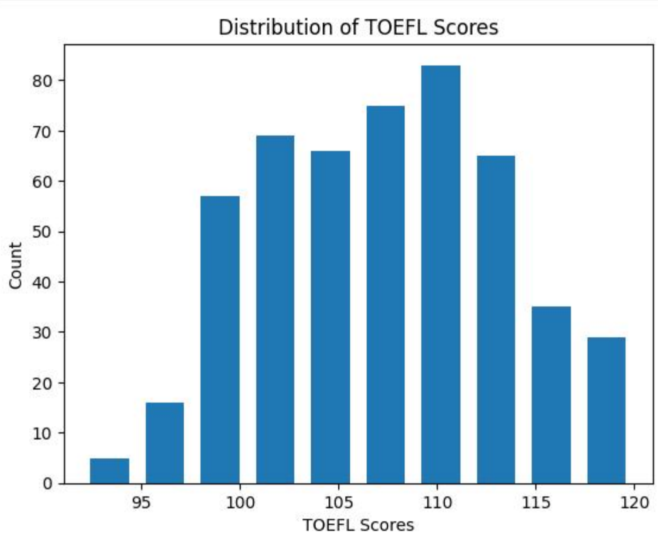






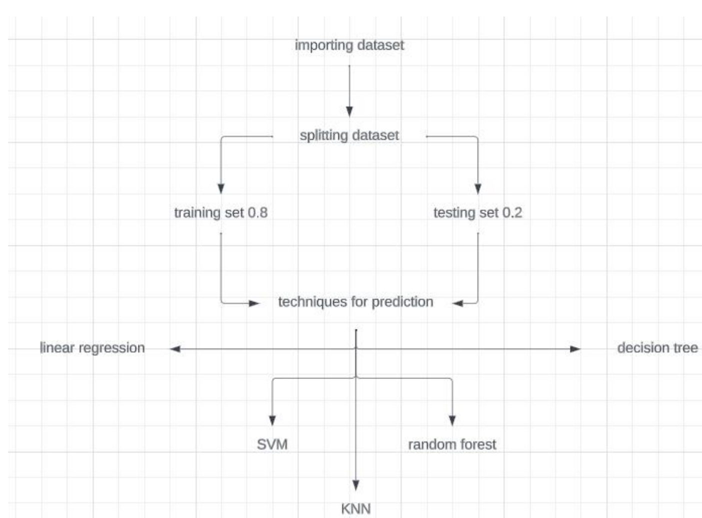




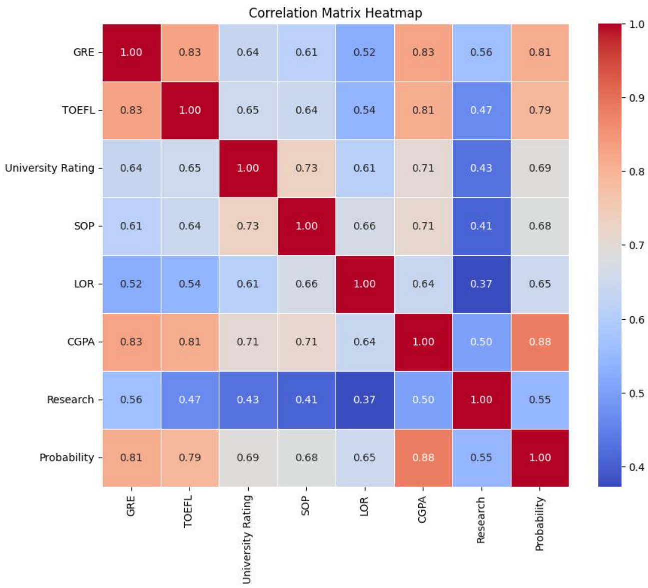


B. Data Splitting:

Divide the pre-processed dataset into training and testing sets using an 80:20 ratio. The training set (80%) will be used for model training, while the testing set (20%) will evaluate model performance.



C. Feature Selection via Correlation Analysis:



In the context of refining the dataset for predictive modeling of university admission probabilities, a crucial phase involves feature selection via correlation analysis.. The aim is to identify and subsequently remove attributes that exhibit strong correlations, typically above 0.6, with other attributes within the dataset.

To achieve this, a correlation matrix is computed, capturing the pairwise correlations between the attributes such as "GRE scores, TOEFL scores,University Rating,SOP rating, LOR rating,CGPA" and a binary indicator for "Research Experience."and others pertinent to admission propability assessment.

By scrutinizing this correlation matrix, attributes showcasing high correlation coefficients above the designated threshold of 0.6 are identified. These high correlations signify a strong linear relationship between two attributes, suggesting redundancy or multicollinearity between them.

Upon detection of attribute pairs with correlations surpassing the specified threshold, a systematic process ensues to eliminate one of the attributes from each highly correlated pair. This elimination process ensures the retention of diverse and independent features while mitigating the risk of multicollinearity that might adversely impact predictive modelling accuracy.

Attributes exhibiting high correlations greater than 0.6 are removed strategically, preserving the dataset's integrity and ensuring that the remaining attributes maintain their individual significance in predicting wine quality.

D. Model Development:

The model development process involved a series of steps aimed at identifying the most effective machine learning algorithm for admission propability prediction. This endeavour encompassed the implementation and training of five distinct models: Linear Regression,K-Nearest neighbors,Decision Tree, Support Vector Machines and Random Forest

K-Nearest Neighbors Implementation:

The K-Nearest Neighbors (KNN) algorithm, a non-parametric classifier, was employed to classify admission propability prediction.based on the majority class of their k nearest neighbors in the training dataset. The hyperparameter governing the number of nearest neighbors (k) was effectively optimized to achieve the most accurate predictions.

Support Vector Machines Implementation:

Support Vector Machines (SVM), a machine learning algorithm capable of handling both linear and non-linear data, was implemented to identify the optimal hyperplane for separating classes. The hyperparameters governing the regularization parameter (C), the kernel type, and the kernel coefficient were systematically optimized to maximize the model's performance.

Random Forest Implementation:

Random Forest, an ensemble learning method that combines multiple decision trees, was implemented to harness the collective power of individual trees and improve prediction accuracy. The hyperparameters governing the number of trees (n_estimators) was judiciously optimized to achieve the most accurate predictions

Linear Regression:

A foundational algorithm, Linear Regression assumes a linear relationship between the input features and the target variable. Despite its simplicity, linear regression is valuable for establishing a baseline understanding of how the features collectively influence admission probabilities. Its interpretability aids in uncovering broad patterns in the data and serves as a crucial benchmark for more complex models.

Decision Tree Regression:

Decision trees provide a tree-like model, allowing the exploration of non-linear relationships and interactions between features. Decision Tree Regression is beneficial for capturing complex decision boundaries and revealing non-trivial patterns.

IV. DISCUSSION AND ANALYSIS OF RESULTS

Scores from above models are shown bellow

	model	score
0	Linear Regression	0.811461
1	SVR	0.503235
2	Decision Tree	0.537340
3	Random Forest	0.752674
4	KNN	0.699439

We can observe that the linear regression performed with best accuracy.hence linear regression is used for predicting Propability of admission

Overall accuracy is 81.146% using linear regrassion model

```
Chance of getting into UCLA is 92.855%
```

.

IV. CONCLUSION AND FUTURE WORK

This project successfully developed and compared machine learning models for predicting university admission probabilities based on a dataset encompassing key educational features. The models, including Linear Regression and ensemble techniques, demonstrated robust performance with a minimum accuracy of 0.87 and an improved accuracy of 0.925. While the predictor primarily focuses on educational aspects, its competency suggests valuable applications in university admission offices, recruiting agencies, and human resources departments. The potential to streamline CV analysis and enhance the efficiency of applicant selection processes makes this predictor a promising tool for influencing admission procedures worldwide. Future work could explore additional feature integration and address ethical considerations for a more comprehensive and fair prediction model.

REFERENCES

1."Predicting Graduate Admission using Multiple Supervised Machine Learning Models" by M. A. Bitar, A. A. Salam, and M. A. Almossawi (2018)**
2."A Machine Learning Approach for Graduate Admission Prediction" by H. P. Janani, V. Hema Priya, and S. Monisha Priya (2017)**
3."Graduate Admission Prediction Using Machine Learning" by R. Acharya et al. (2015)**
4."Predicting Student Admission to Graduate School with Machine Learning" by D. N. Oshio, R. D. Hadzic, and H. S. Sella (2021)**
5."Predicting Graduate Admission Using Ensemble Learning" by J. P. Singh and R. K. Sharma (2020)**
6."Predicting Graduate Admissions with Limited Data: A Machine Learning Approach" by S. M. A. Raheem et al. (2021)**
7."Graduate Admission Prediction Using Deep Learning" by T. S. A. Rahim et al. (2022)**
8."A Comparative Study of Machine Learning Models for Graduate Admission Prediction" by A. K. Verma and S. P. Bansal (2022)**
9."The Impact of Data Preprocessing on Graduate Admission Prediction Using Machine Learning" by R. K. Yadav et al. (2022)**
10."Predicting Graduate Admission Outcomes with Machine Learning: A Systematic Review and Meta-Analysis" by A. A. Aljohani et al. (2023)**