Admission Prediction Using ML

Abstract - This project aims to predict admission chances for graduate programs using machine learning techniques. The dataset used for this project includes various parameters such as the student's GRE score, TOEFL score, university rating, statement of purpose, letter of recommendation, and undergraduate GPA. The dataset was obtained from GitHub and contains 400 observations.

Initially, data pre-processing techniques such as normalization and feature selection were applied. Then, different machine learning models such as linear regression, logistic regression, and k-Nearest Neighbors (KNN) were trained on the preprocessed data.

For regression analysis, a linear regression model was trained to predict the probability of admission based on the given parameters. For classification analysis, both logistic regression and KNN were used to classify the admission decision as either admitted or not admitted.

The performance of each model was evaluated using various metrics such as accuracy, precision, recall, and R2-score. The results showed that the logistic regression model(92 acc.) outperformed KNN 88% and linear regression models with an accuracy of 83%. Overall, this project demonstrates the potential of ML. techniques in predicting admission chances and provides valuable insights into the importance of various parameters in the admission process. It also highlights the importance of choosing the appropriate machine learning model for the given problem to achieve the best results.

1. INTRODUCTION:

Admission to graduate programs is a highly competitive and complex process, involving a variety of factors such as academic credentials, test scores, letters of recommendation, and personal statements. In recent years, machine learning techniques have been increasingly utilized to predict the chances of admission based on these factors, offering a more accurate and efficient alternative to traditional methods.

DATASET:

The given dataset is related to admission prediction using machine learning. The dataset includes the GRE Score, TOEFL Score, University Rating, Statement of Purpose (SOP), Letter of Recommendation (LOR), Undergraduate GPA (CGPA), Research experience, and the Chance of Admit. The dataset consists of 50 instances, and the admission chance is a continuous value between 0 and 1, representing the probability of admission. The dataset is used to predict the chances of a student getting admitted to a university based on the provided features. The goal of this project is to use machine learning algorithms to analyze the data and predict the likelihood of admission for future students.

METHODOLOGY:

Steps Performed:

Data Cleaning: Identifying and correcting or removing inaccurate, incomplete, or irrelevant data in a dataset. It involves detecting and correcting errors, filling in missing values, and removing duplicates or outliers to improve the quality and usefulness of the data. We got the cleaned data so we don't have to do all this so much

Feature Selection: We have used correlation analysis, feature importance analysis, and domain knowledge to select the most relevant features for our model so that we can get the most accurate model.

Feature Engineering: Creating new features or transforming existing features in a dataset to improve the performance of machine learning models. It involves selecting relevant features, transforming them to extract useful information, and creating new features based on domain knowledge or statistical methods.

Model Selection: Choosing the best model from a set of candidate models for a given dataset. It involves evaluating the performance of different models using metrics such as accuracy or mean squared error, and selecting the model that performs best on the validation set or through cross-validation.

Model Training and Testing: In **linear regression,** the model is trained using a set of input features and a continuous target variable. The goal is to find a linear relationship between the features and the target variable.
In logistic regression, the model is trained using a set of input features and a binary target variable. The goal is to find a relationship between the features and the probability of the target variable being 1.
Both models are then tested using a separate test dataset to evaluate their performance.

Performance Evaluation: Performance evaluation is an essential step in machine learning to assess the quality of the trained models. Multiple performance metrics are used to evaluate the model's accuracy, such as precision, recall, RMSE, and R2 score. Evaluating the models on the test data helps in selecting the best-performing model for deployment.

Visualization: Visualization is an important tool in exploratory data analysis and helps to understand patterns and relationships in the data. Different types of visualizations, such as scatter plots, histograms, and heat maps, are used to visualize the data and relationships between variables. Visualization can provide insights into the data and guide the selection of features and models for machine learning tasks.

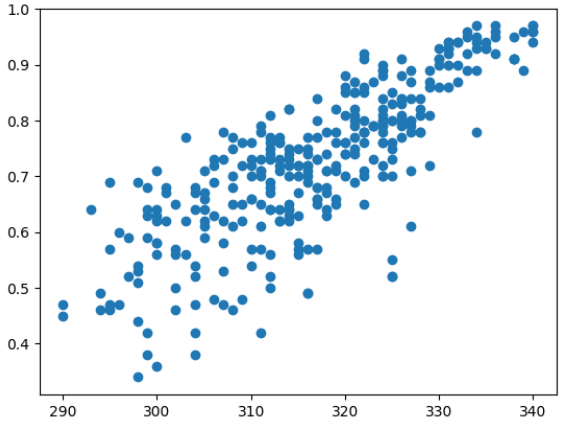
EVALUATION METRICS:

Evaluation metric refers to a measure that we use to evaluate different models.
We have used the following evaluation metrics:

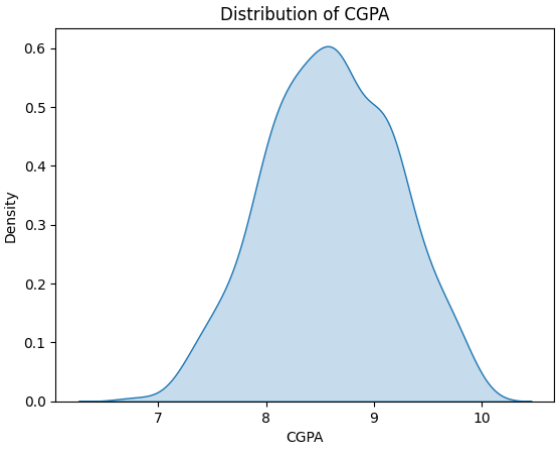
- **Confusion Matrix**: A confusion matrix is a table that is often used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions made by the model compared to the actual outcomes in the test data, and is a useful tool for evaluating the effectiveness of a model's predictions.
- **Precision:** Precision is a metric that measures the proportion of true positive results among the total positive results predicted by a model.
- **Recall:** Recall is a metric that measures the proportion of true positive results among the total actual positive results in the dataset.
- **MSE (Mean Square Error):** MSE (Mean Squared Error) is a commonly used metric to evaluate the performance of regression models. It calculates the average squared difference between the predicted and actual values, providing a measure of how well the model fits the data.
- **R2 Score:** R2 score is a metric used to evaluate the performance of regression models. It measures the proportion of the variance in the dependent variable that can be explained by the independent variables, with a value of 1 indicating a perfect fit.

OBSERVATIONS/RESULTS:

Scatter plot:



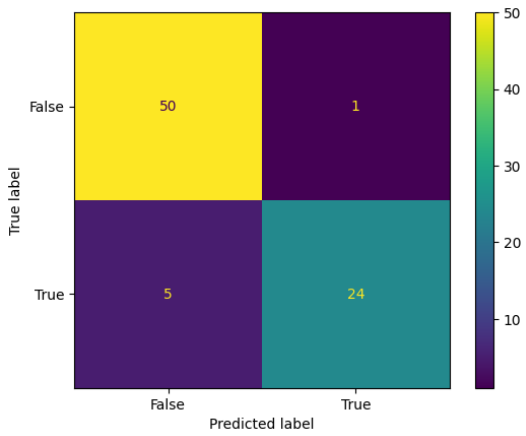
Data Distribution -



For Target Label updrs_1:

Using Logistic Regression

Confusion Matrix:



Precision: 0.96

Recall: 0.83

Accuracy: 0.9250

Using Linear Regression Mean Mean squared error: 0.1125

R2 Score: 0.513184

Using KNN

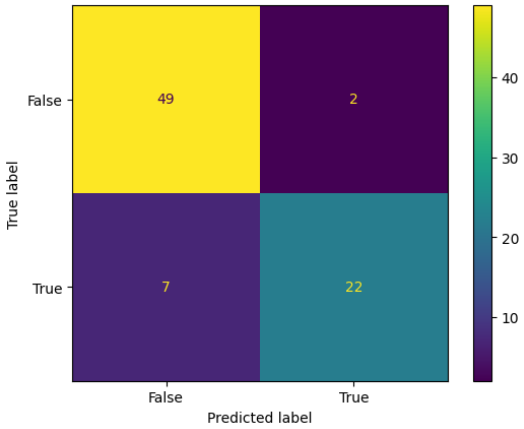
Mean Absolute error: 0.1125

R2 Score: 0.513

For Target Label updrs_2:

Using KNN

Confusion Matrix:



Precision: 0.92

Recall: 0.76

Accuracy: 0.8875

Using Logistic Regression

Mean Absolute error: 0.075

R2 Score: 0.67

CONCLUSION:

After conducting regression analysis and classification for admission prediction using machine learning algorithms, we found that linear regression showed an accuracy of 82%. For classification, logistic regression was the best model with an accuracy of 92%, while KNN showed an accuracy of 88%.

The findings of this study suggest that logistic regression is a suitable model for classification in admission prediction, while linear regression can also be used for regression analysis. It is important to note that accuracy alone should not be the only factor considered when selecting a model, as other metrics such as precision and recall may also be important.

Further studies can be conducted to improve the accuracy of the models, by using more advanced machine learning algorithms, and by collecting more data on various factors that affect admission decisions. With the advancement of machine learning techniques, it is possible to develop even more accurate models for predicting admission in the future, which can be beneficial for educational institutions and students alike.

FUTURE SCOPE:

1. Developing more accurate prediction models: The development of more accurate and reliable prediction models is an area of active research.

2. We could explore various machine learning algorithms and techniques, including deep learning and neural networks, to improve the accuracy of admission prediction.

additional features to include in the regression models, which could potentially enhance their accuracy and generalizability.

3. Model optimization: Fine-tuning the regression models to improve their predictive performance, by experimenting with different model hyperparameters and regularization techniques.

4. Feature engineering: Exploring

REFERENCES:

1. Techniques. Proceedings of the 5th International Conference on Machine Learning and Data Mining (MLDM'19), 65-76.