# CIS550 - Group 5 - Project Proposal

Forecasting Graduate Admissions Outcomes Using Regression Models

Date: 18.04.2025

## Group Members

Alta Santosh Satyasai Harshith (Roll No: 02, Student ID: 02197872)

Maddineni Rishindra Chowdary (Roll No: 19, Student ID: 02179038)

Thatikonda Kinshuk (Roll No: 31, Student ID: 02182480)

## 1. Introduction

Graduate admissions can be challenging and opaque, leading many students to question their chances for success. Machine learning presents an excellent opportunity to offer data-driven clarity to this method. Our work aims to create a regression-based machine learning pipeline that can predict a student's likelihood of admission to residential or industrial settings using quantitative profile factors such as the GRE, TOEFL, CGPA, and research experience.

Based on previous data trends, this model could be a useful application screening tool for students, allowing them to make better informed decisions and more realistically target universities.

## 2. About the Dataset

We make use of the 500-student "Graduate Admissions" dataset that is accessible on Kaggle. It has the following characteristics:   
GRE Score   
TOEFL Score   
University Rating   
SOP (Statement of Purpose)   
Strength LOR (Letter of Recommendation) Strength   
CGPA Research Experience (0 or 1)   
Chance of Admit (Target Variable)   
  
We aim to predict the Chance of Admit using the above variables through a regression framework.  
Dataset: Admission\_Predict.CSV

## 3. Tools and Technologies Used

1. Python
2. Jupyter Notebook
3. Libraries: NumPy, Pandas, Matplotlib, scikit-learn, Seaborn, XG-Boost.

## 4. Implementing the ML Pipeline

**Data Cleaning / Preprocessing:**

* Handle missing values (if any)
* Normalize numerical values using MinMaxScaler
* Encode binary fields
* Visualize correlations to guide feature selection

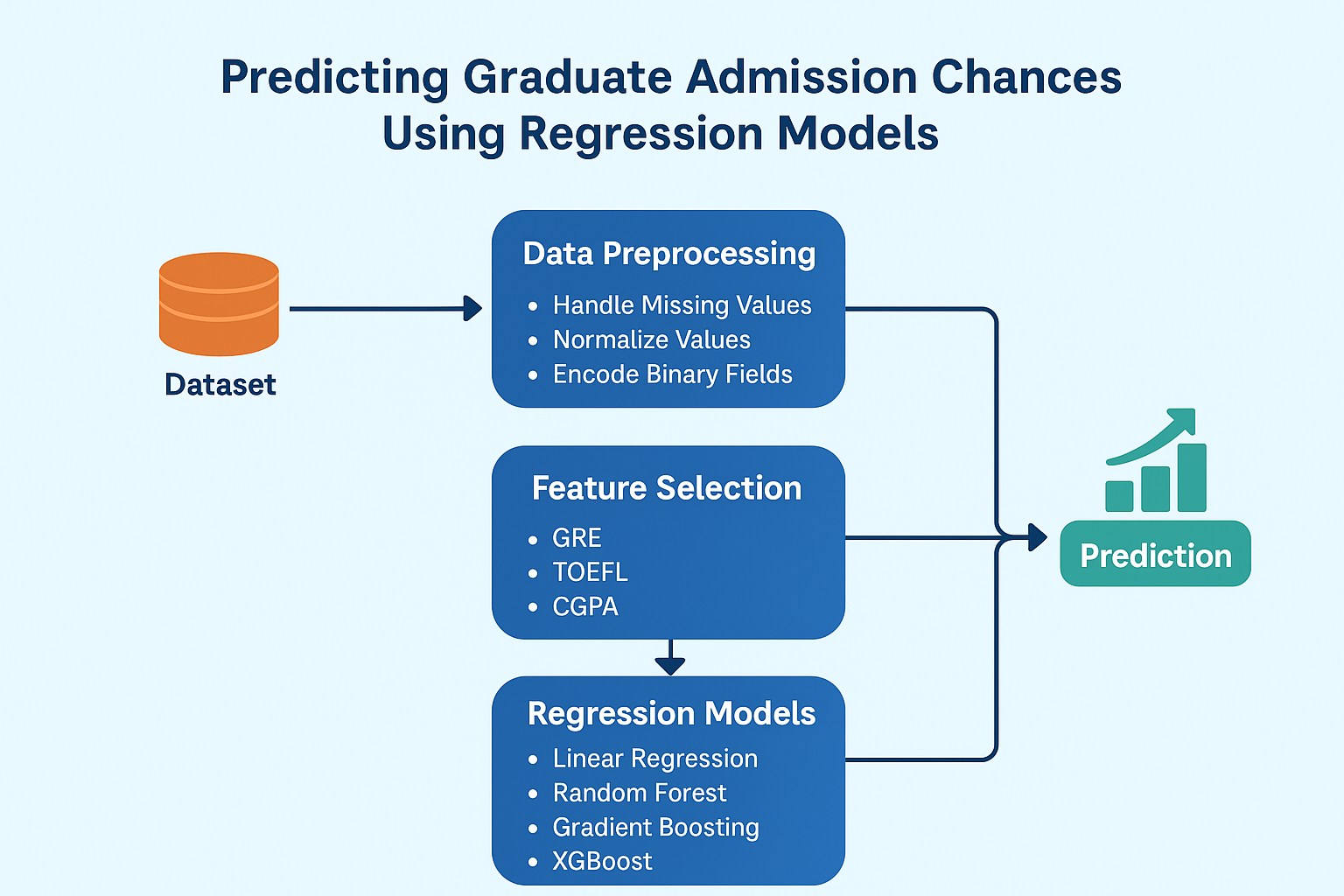
**Feature Selection:**

* Focus on key predictors: GRE, TOEFL, CGPA, and Research
* Drop features with weak correlation or redundancy if necessary

**Model Building:**  
We compare multiple regression models:

* Linear Regression
* Random Forest Regressor
* Gradient Boosting Regressor
* XGBoost Regressor

## Architecture



Models Used:  
1) The Linear Regression is a baseline model that is easy to understand. aids in comprehending the linear connections between the target (Chance of Admit) and   
characteristics.

2) The Random Forest decision tree-based ensemble learning technique. avoids overfitting by bagging and captures non-linear correlations.

3) Boosting gradients builds a progressive, step-by-step additive model. aims to reduce residual errors in order to produce forecasts that are more accurate.

4)XGBoost is a gradient boosting implementation that is optimized. Extremely effective and frequently produces cutting-edge outcomes in assignments involving structured or tabular data..=

## 5. Model Evaluation

We split the dataset into 80% training and 20% testing. Evaluation is performed using:

* Root Mean Squared Error (RMSE)
* R² Score
* Mean Absolute Error (MAE)
* Scatter plots of actual vs predicted values

This helps assess both the accuracy and consistency of each model.

## 6. Evaluation Metrics

* RMSE / MSE: Measures prediction error magnitude
* R² Score: Measures explained variance
* MAE: Measures average absolute prediction error
* Visual Comparison: Plots for prediction trends and outliers

## 7. Outcome and Impact

We expect the final model to predict admission chances with high reliability, enabling:

* Students to gauge reach/safety schools more accurately
* Consultants to guide applications using data-backed insights
* An interactive tool for quick self-assessment

We may further enhance the model by tuning hyperparameters, including interaction terms, and comparing with neural networks for more complex pattern capture.