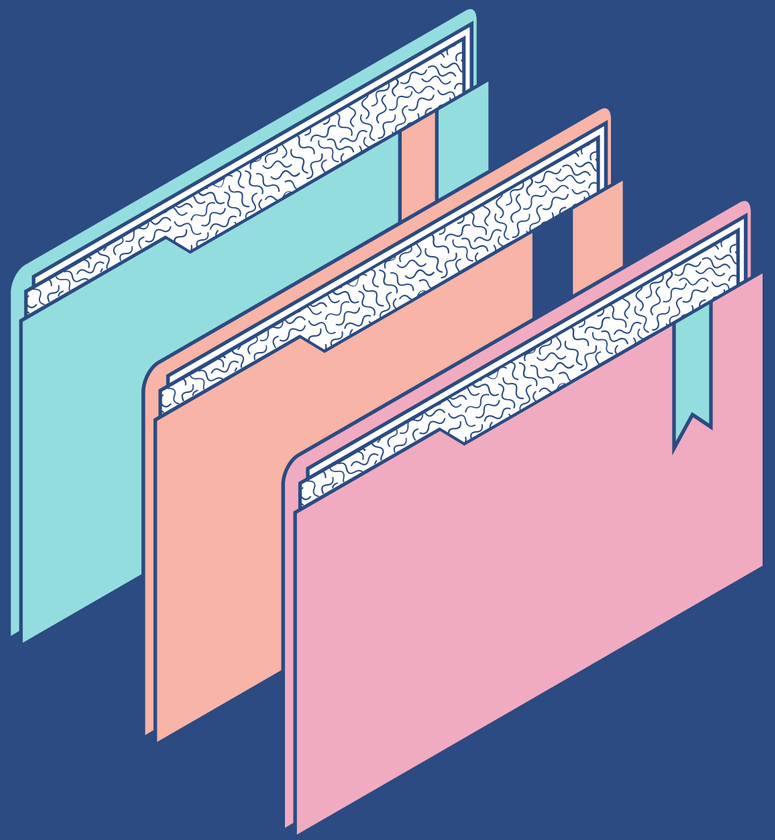
CSCI 111 - K

Catalysts of Choice: A Comparative Analysis of Graduate Admissions Models

Dizo, Salazar, Yumol



Presentation Flow

- Introduction
- Dataset
- Classification Models
  - KNN
  - Decision Trees
  - Random Forest
- Regression
- Demonstration
- Insights
- Recommendations

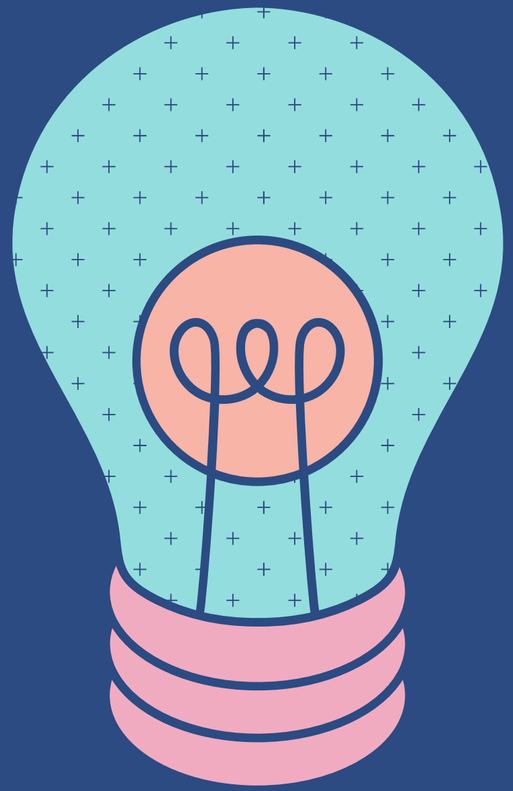
The Problem:

In an era of intensifying global competition for college admissions, educational institutions face **the challenge of sifting through an influx of applicants**. Ranging from quantitative metrics to qualitative insights, the task of equitably identifying deserving students has never been more critical.

Objective:

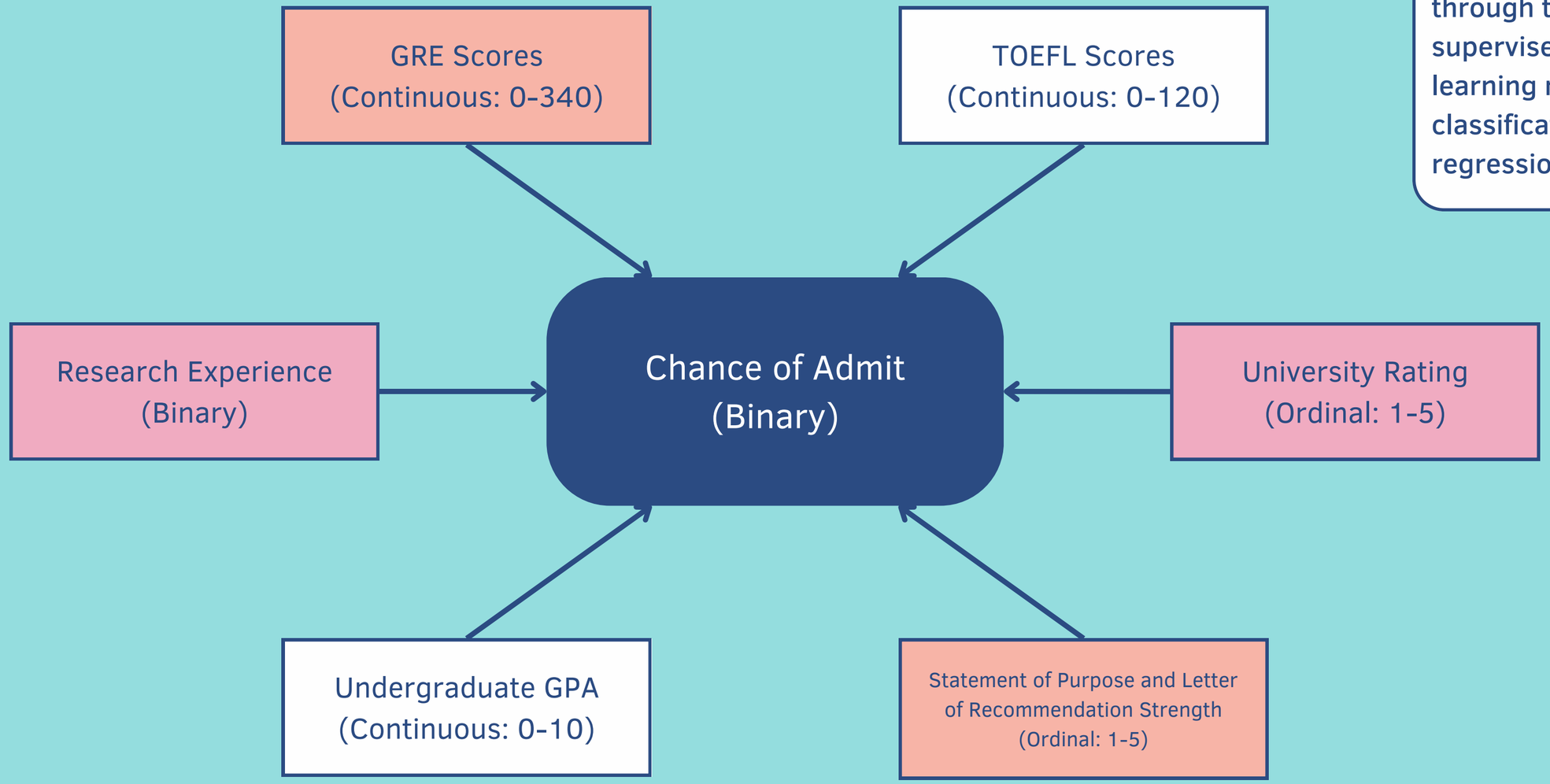
This project aims to address this challenge by developing models through **supervised machine learning methods** (K-Nearest Neighbors, Search Tree, and Random Forest).

The objective is to create a classification system that efficiently **determines whether a student qualifies for admission**, thereby contributing to a more streamlined and equitable admissions process



Dataset Parameters

We want to determine what parameters are important for a student to get into a graduate school through the utilization of supervised machine learning methods, namely classification and regression.



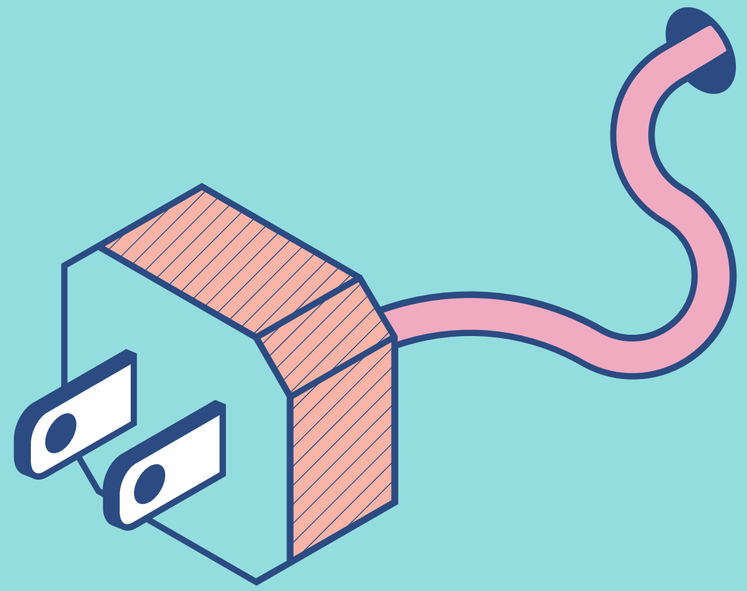
Simple Exploratory Data Analysis

STEP 1	STEP 2	STEP 3	STEP 4	STEP 5
Outlier Detection	Correlation Heatmap and Pairplots	Distribution Graphs	Cutoff Analysis	Average Scores

Outlier Detection

- **Tukey Method**: identified outliers based on the interquartile range (IQR)
- **Z-Score Method**: identified values deviating significantly from the mean

After applying the two methods, we found **no extreme values**. All data points are within a certain range, indicating a relatively consistent distribution.

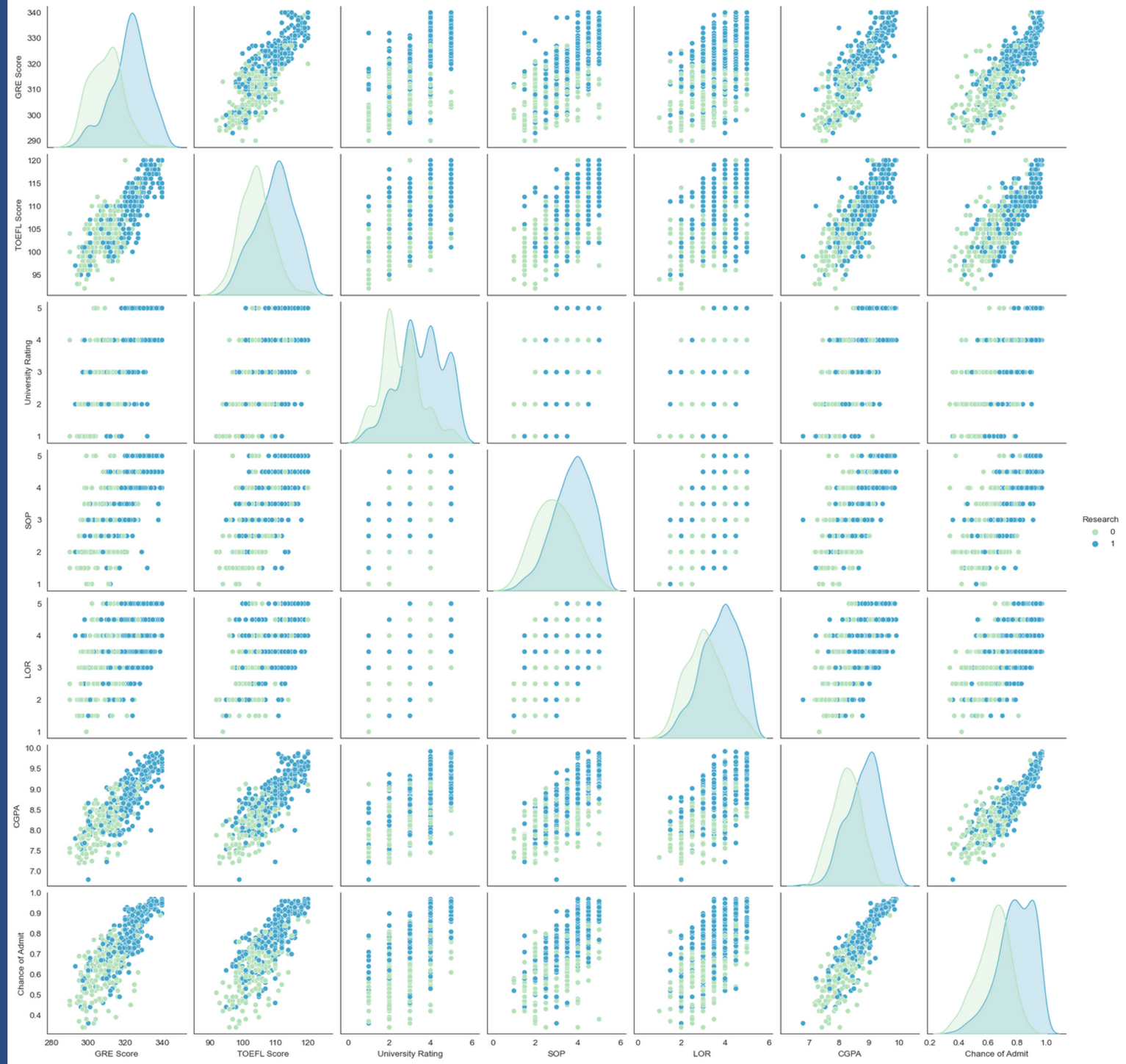


Correlation Heatmap

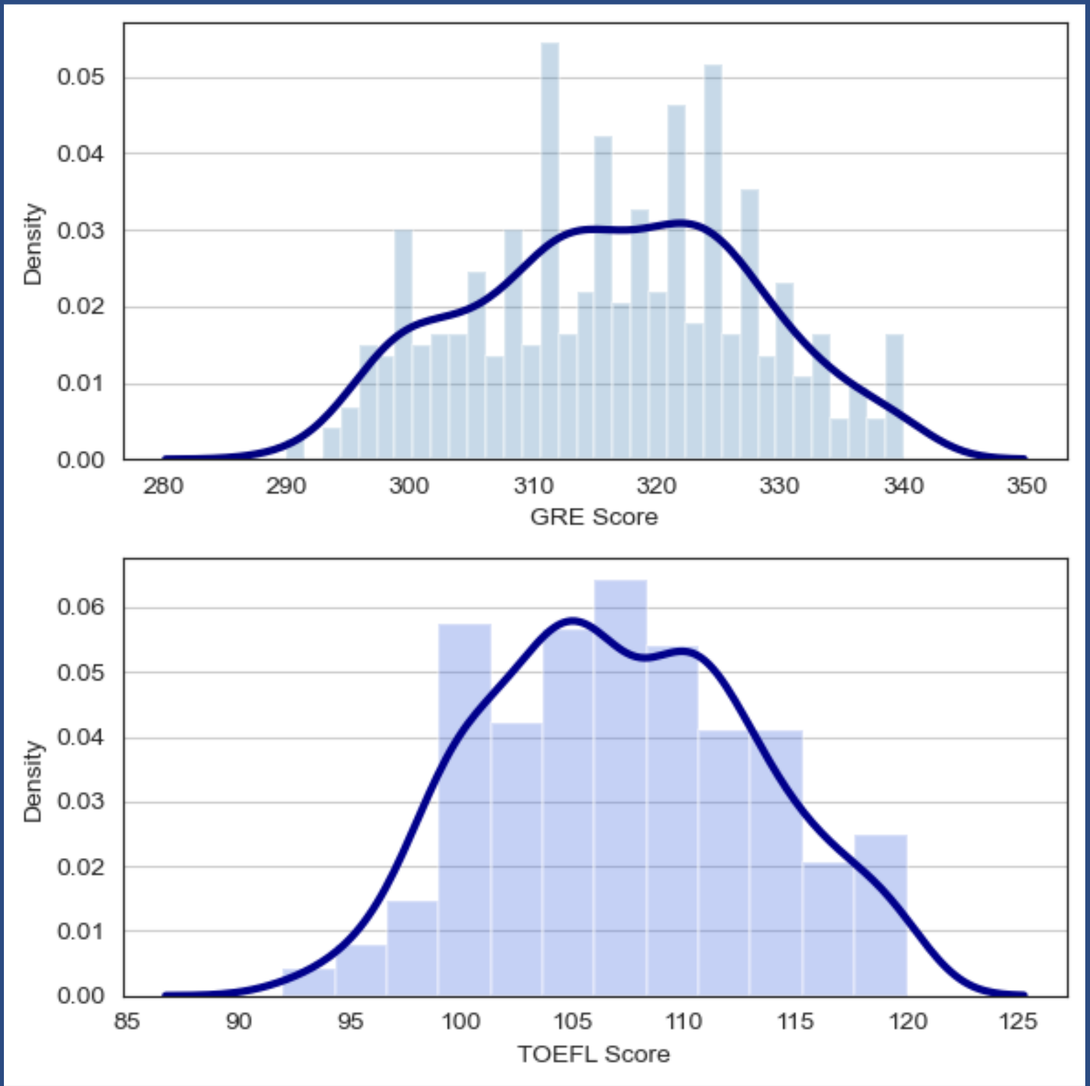
Our correlation heatmap revealed strong correlations between the chance of admission and features such as CGPA, GRE scores, and TOEFL scores.

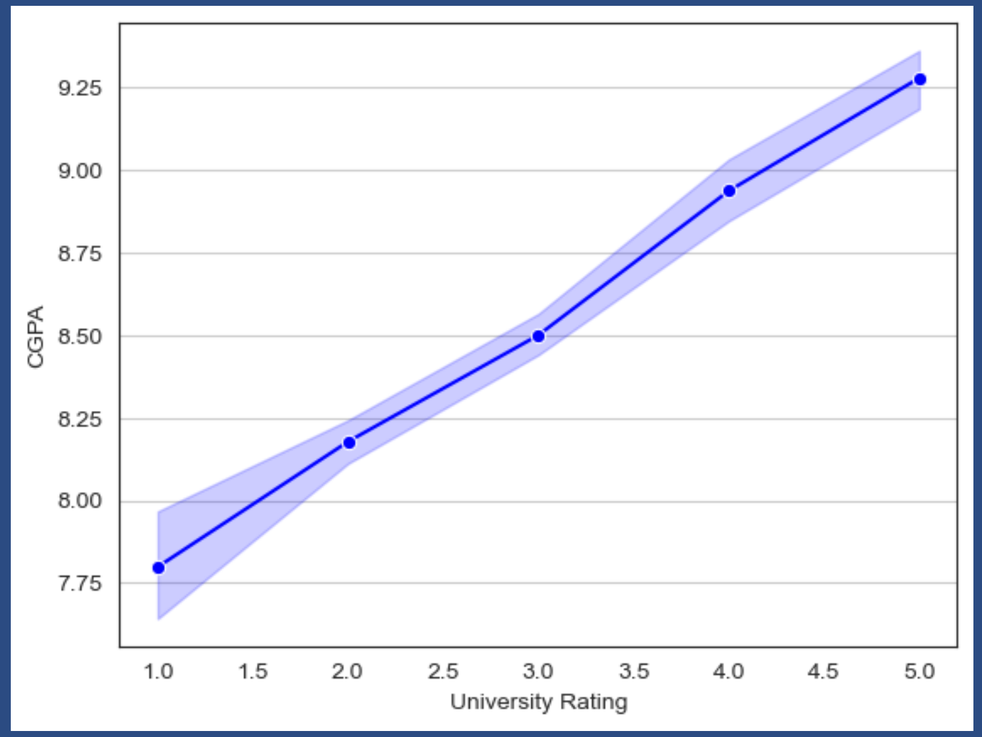
Pairplots

Pairplots illustrated linear relationships between GRE scores, TOEFL scores, and CGPA, reinforcing the idea that these variables are interconnected.



Distribution Graphs





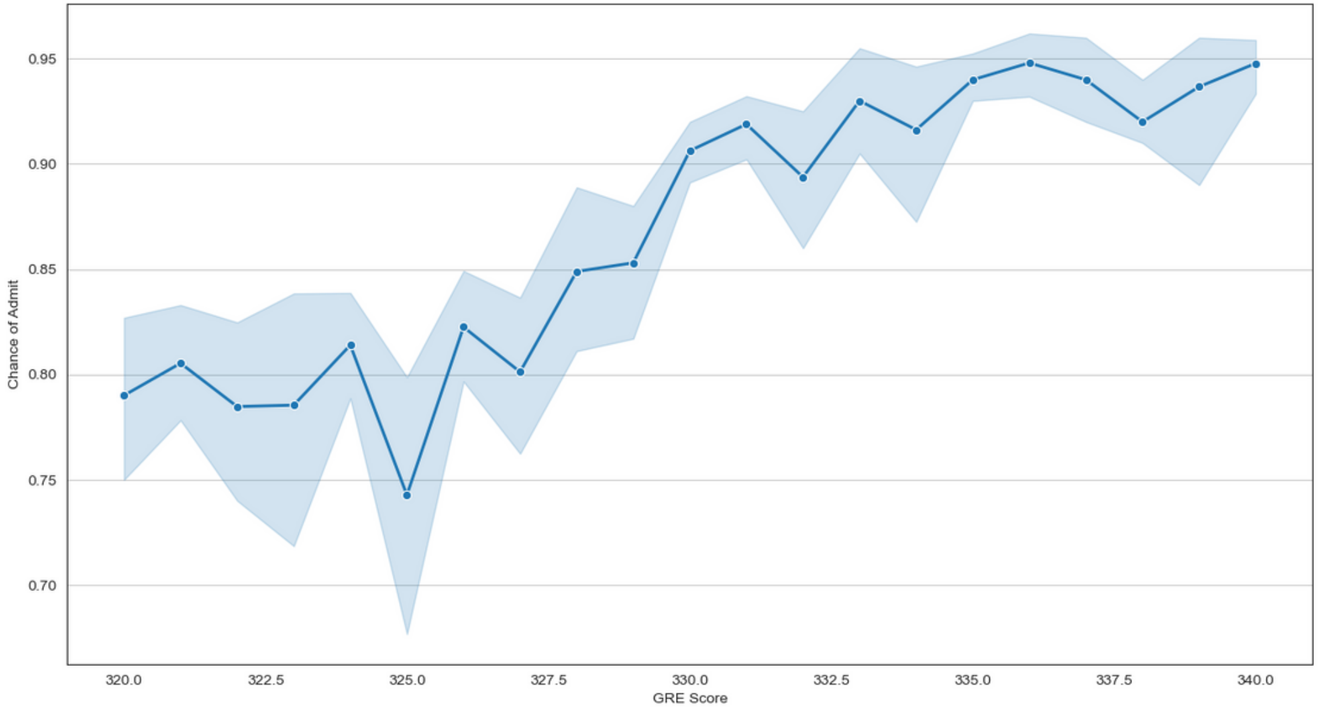
- Individuals in our dataset tend to score above 310 in GRE
- Individuals in our dataset tend to score above 100 in TOEFL
- University Ratings increased with higher CGPA

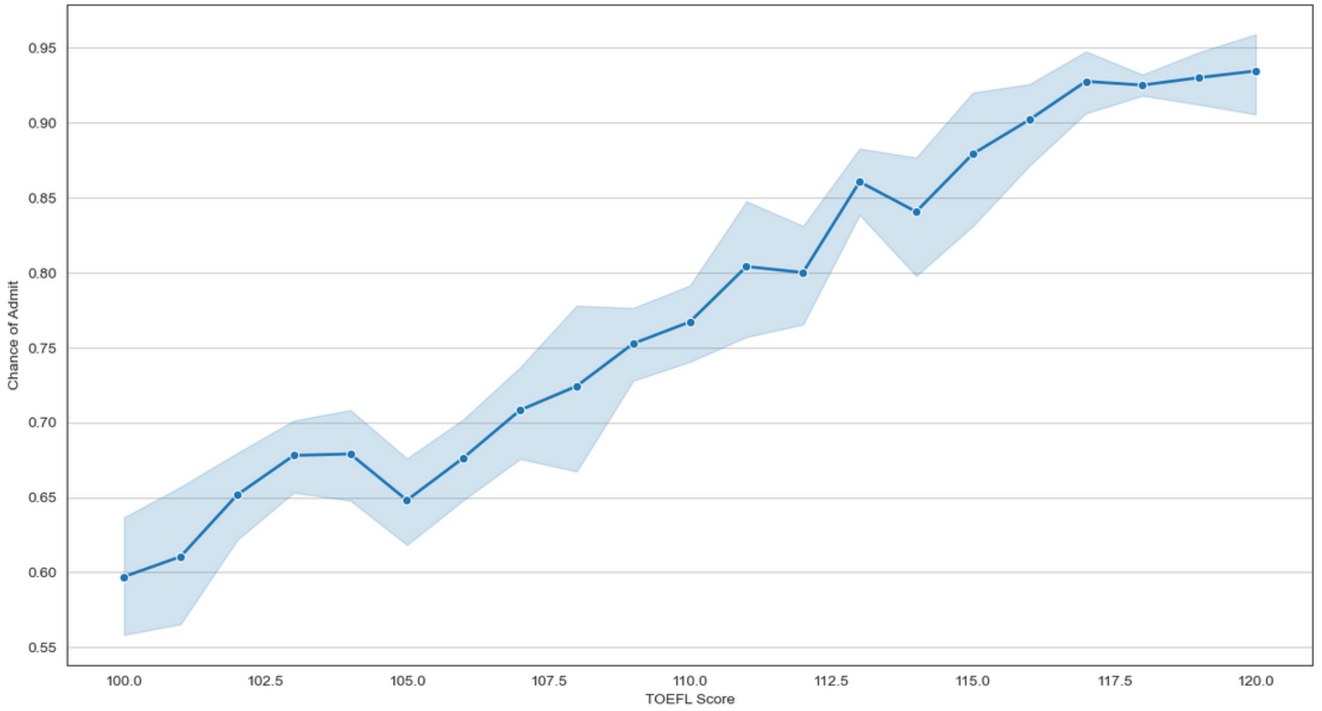
Cutoff Analysis

Setting cutoff scores based on external sources, we analyzed data for scores above these cutoffs.

- **GRE scores greater than or equal to 320**
- **TOEFL scores greater than or equal to 100**

Line graphs for GRE and TOEFL scores against the chance of admission clearly showed a **positive correlation.**

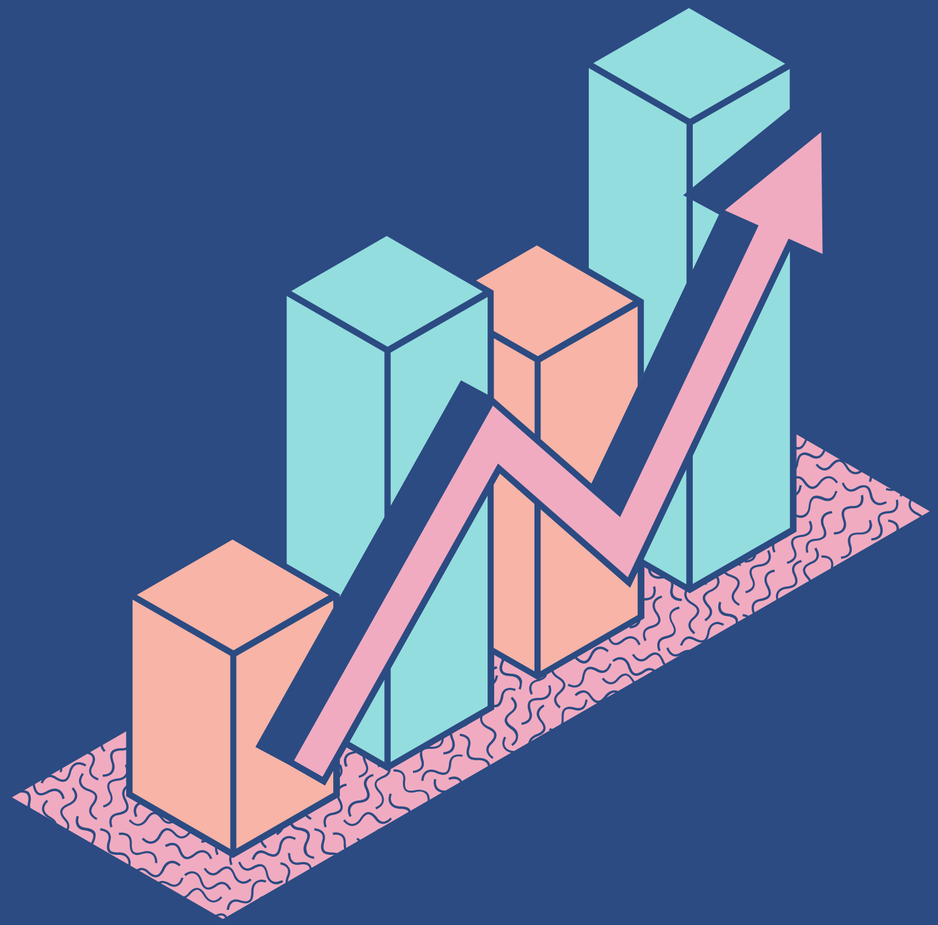




Average Scores

Average scores for **admitted candidates** were calculated, providing a benchmark for comparison:

- Average GRE Score: 316.47 out of 340
- Average TOEFL Score: 107.19 out of 120
- Average CGPA: 8.58 out of 10
- Average Chance of Admission: 72.17%



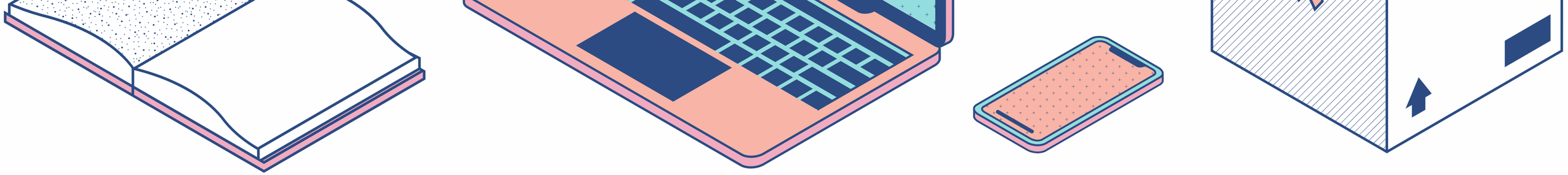
Machine Learning Application

Preprocessing

For our machine learning models, we performed a 75-25 split for training and testing the data. The Chance of Admission was binarized, classifying it as 1 if it was greater than 75%, and 0 otherwise.

Model Selection

We used K-Nearest Neighbors, Decision Trees, and Random Forests for classification (predicting whether the chance of admission is greater than 75%) and Linear Regression, Decision Tree Regression, Random Forest Regression, and K-Neighbors Regression for regression (predicting the exact chance of admission).



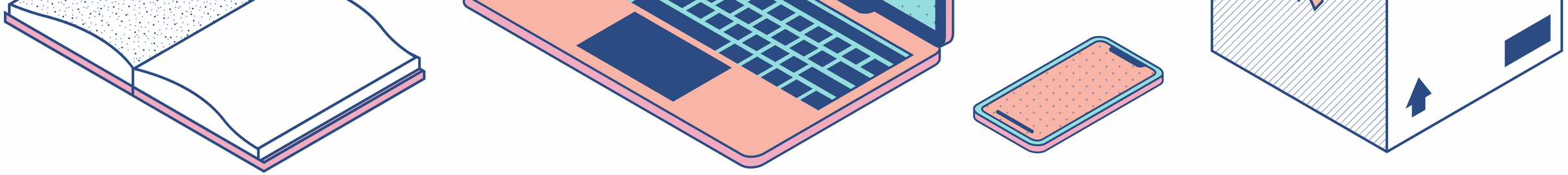
K-Nearest Neighbors (KNN)

Advantages

- Simple and Intuitive
- Non-Parametric
- Adaptability to Complex Boundaries

Disadvantages

- Computational Cost
- Sensitivity to Noise and Outliers
- Need for Feature Scaling



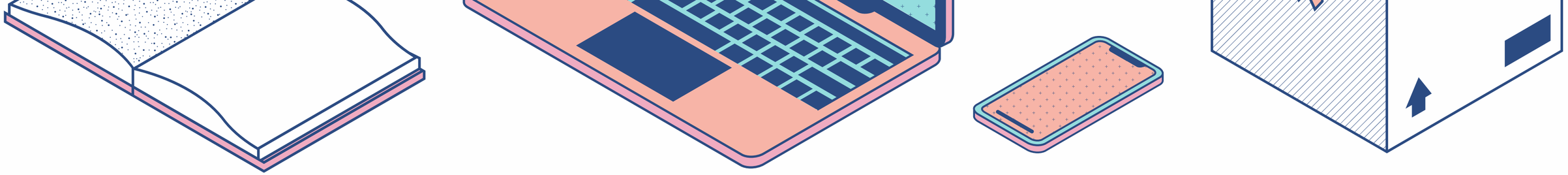
Decision Trees

Advantages

- Interpretability
- Handles Mixed Data Types
- Automatically Selects Features

Disadvantages

- Overfitting
- Instability
- Limited Expressiveness



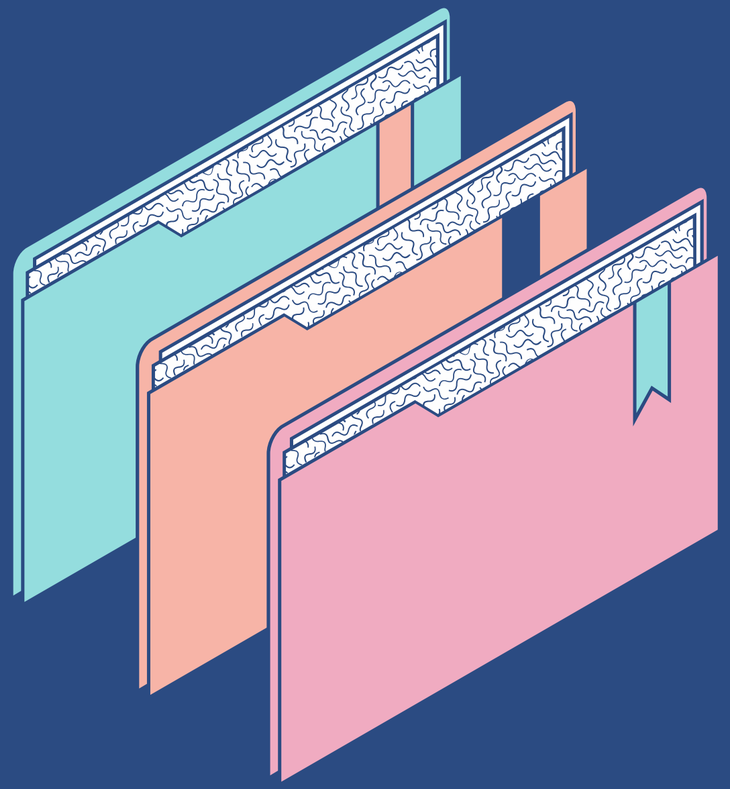
Random Forests

Advantages

- Ensemble Learning
- High Accuracy
- Handles Missing Values

Disadvantages

- Computational Complexity
- Less Interpretability
- Black Box Nature



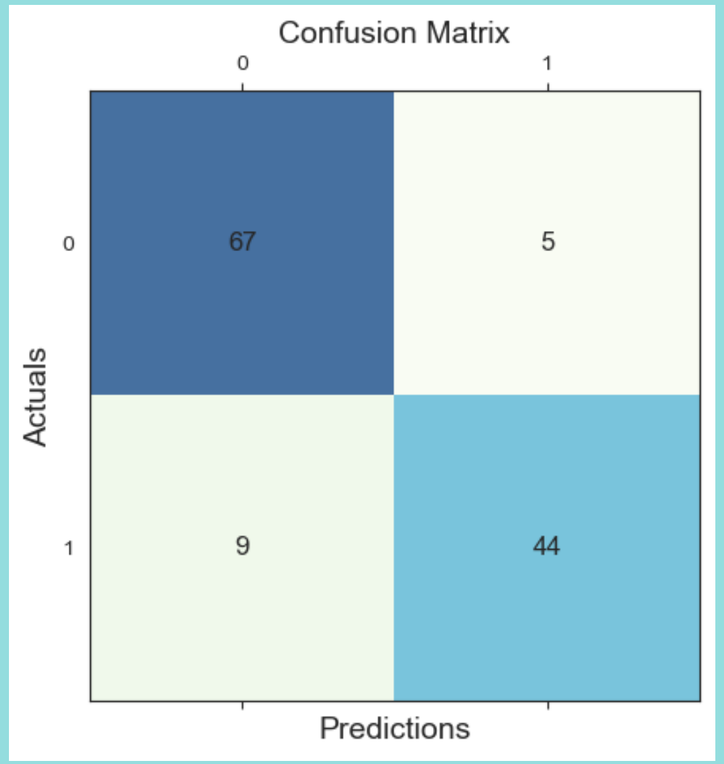
Evaluation

- **Accuracy Score:** Measures the overall correctness of the model, the proportion of correctly classified instances.
- **Precision Score:** Indicates the ability of the model not to label as positive a sample that is negative.
- **Recall Score:** Measures the ability of the model to capture all the positive instances.
- **F1 Score:** Harmonic mean of precision and recall.

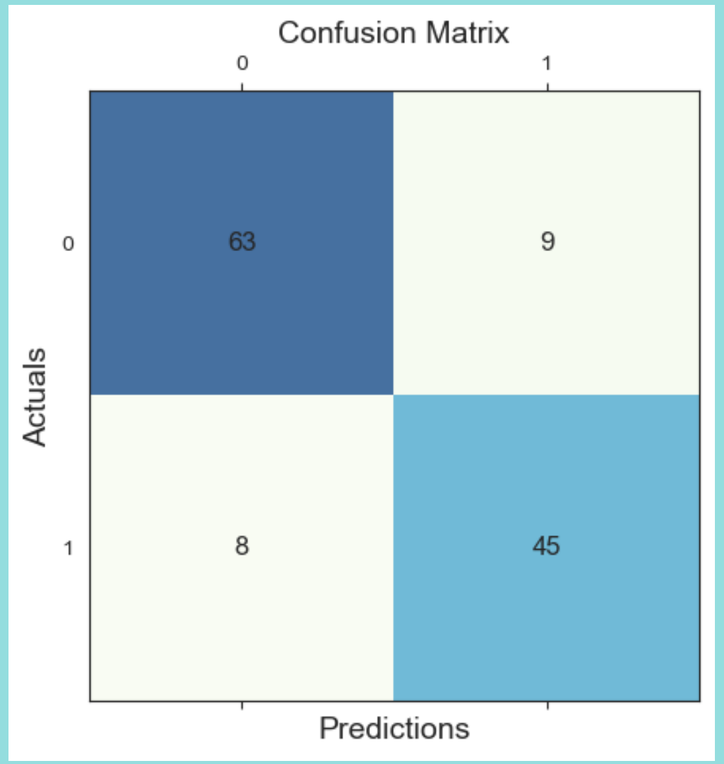
Summary of Evaluation Metrics for Classification Models

	Accuracy Score	Precision Score	Recall Score	F1 Score
KNN (k=7)	0.888	0.898	0.830	0.863
Decision Trees	0.872	0.849	0.849	0.849
Random Forests	0.920	0.939	0.868	0.902

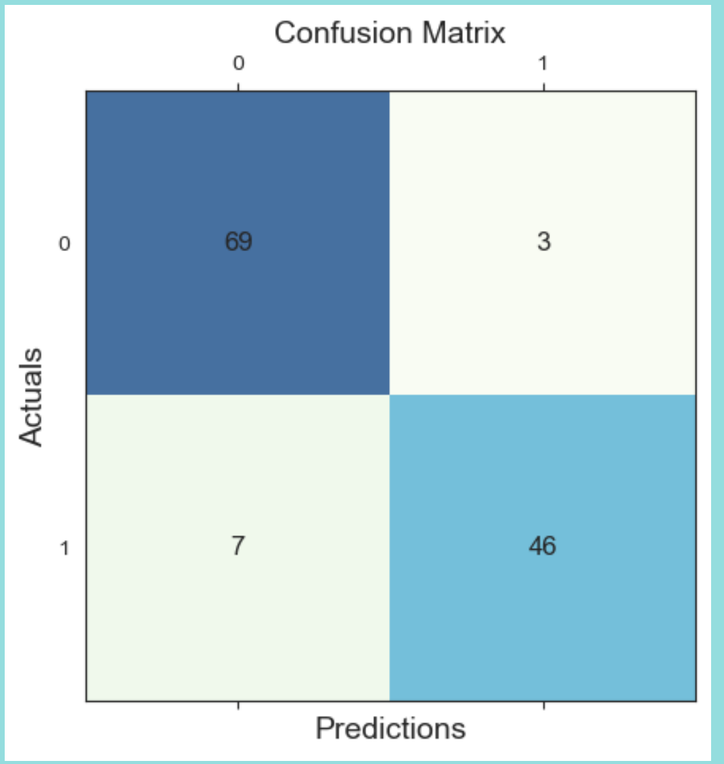
Confusion Matrix



KNN (K=7)

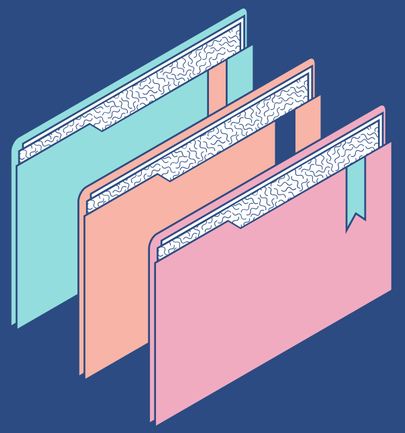


DECISION TREES



RANDOM FORESTS

Regression



- **Linear Regression :** Simple, interpretable, and suitable when the relationship between variables is linear. May not perform well in the presence of non-linear relationships.
- **Decision Tree Regression :** Useful for capturing non-linear relationships, easy to interpret, but prone to overfitting.
- **Random Forest Regression :** Combines the strengths of decision trees while addressing overfitting. Offers high accuracy but is computationally more demanding.
- **K-Neighbors Regression :** Simple and flexible but computationally expensive, especially for large datasets. Sensitive to noise and requires proper scaling.

Evaluation Metrics

Mean Absolute Error (MAE)

- Average absolute difference between the predicted and actual values
- Smaller values indicate better performance

Mean Squared Error (MSE)

- Average squared difference between the predicted and actual values
- Generally penalizes larger errors more heavily than MAE

R-squared (R^2)

- Proportion of the variance in the dependent variable that is predictable from the independent variables
- Ranges from 0 to 1, where 1 indicates a perfect fit

Summary of Evaluation Metrics for Regression Models

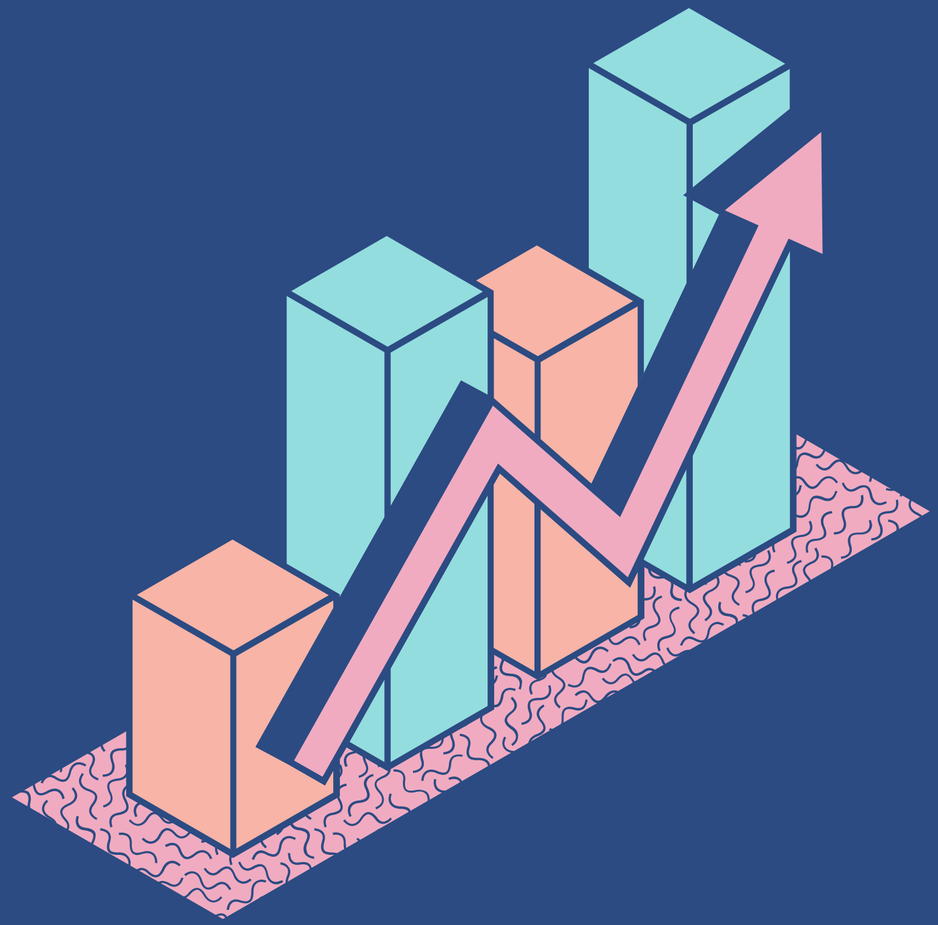
	Mean Absolute Error	Mean Squared Error	R-squared
Linear Regression	0.264	0.103	0.583
Decision Tree	0.150	0.150	0.394
Random Forest	0.142	0.077	0.690
K-Neighbors	0.212	0.094	0.619

Code Demo

Note

In the end, these classification models should only serve as a guide for people behind the admission processes of a university.

Additionally, this may vary from school to school and across varying courses due to differences in the importance of each metric.



Recommendations

- To develop similar models for other datasets from **varying schools** to observe whether there is a similar pattern for certain universities or other possible classification bases
- To **improve the accuracy and precision** of these models through the use of deep learning
- To apply models to the **Philippine context**