PROBLEM STATEMENT: KNOWLEDGE REPRESENTATION AND INSIGHTS GENERATION FROM A STRUCTURED DATASET

The primary objective of this project is to develop an AI-based solution that can effectively represent knowledge and generate insights from any structured dataset. The solution is capable of processing and analysing structured data, identifying patterns, and generating meaningful insights that can aid in decision-making processes.

1. INTRODUCTION:

AI for Engineering College Fee Categorization

To find a perfect college within a specified fee category is tough nowadays, to tackle with this solution I come up with this idea. This report details the development and evaluation of a machine learning model designed to classify engineering colleges in India based on their fee structures. The model uses a random forest classifier to predict whether a college's fees are high, medium, or low based on various input features. Additionally, it provides a list of colleges corresponding to each predicted fee category. The method leverages a carefully curated dataset and employs advanced feature engineering techniques to create a robust machine learning model.

2. DATASET DESCRIPTION:

The primary dataset used in this analysis is derived from the file engineering colleges in India.csv, by Kaggle (https://www.kaggle.com/datasets/shrirangmhalgi/engineering-colleges-in-india)which contains detailed information about 5446 engineering colleges in India. Each entry represents a unique college, encompassing key attributes such as:

- College **Name**: The name of the college.
- Genders **Accepted**: The gender policy of the college.
- Campus **Size**: The physical size of the campus.
- Total **Student Enrolments**: Number of students enrolled.
- Total **Faculty**: Number of faculty members.
- Established **Year**: The year the college was established.
- Rating: The college's rating.
- University: The affiliated university.
- Courses: The courses offered.
- Facilities: Available facilities.
- City: The city where the college is located.
- State: The state where the college is located.
- Country: The country (India).
- College **Type**: The type of college (public/government or private).
- Average **Fees**: The average fees for courses offered.

- **Key Features (**Top 5 Most Important Features):
    - Average Fees
    - Cluster #cluster is obtained by k mean analysis
    - University
    - Total Faculty
    - State

3. METHODOLOGY:

The methodology for developing the machine learning model involves the following steps:

- **Data Preprocessing**: Data preprocessing involves several steps, including:

Tools: Pandas, NumPy, Scikit-learn

- ✓ **Handling Missing Values**: Imputing or removing missing values in the dataset. Consider dropping columns with high percentage of missing values (Rating and Campus Size were dropped).NA value was filled in categorical columns with NaN values ('Genders Accepted', 'University', 'College Type'). NaN values in 'Facilities' was filled with "Unknown". NaN values in numerical columns was filled with 0.
  ('Total Student Enrollments', 'Total Faculty', 'Established Year', 'Average Fees')
- ✓ **Converting Data Types:** Converting columns from categorical to numerical. ('Total Student Enrollments', 'Total Faculty', 'Established Year', 'Average Fees')
- ✓ **Encoding Categorical Variables**: Converting categorical variables into numerical formats using LabelEncoder technique. ('Genders Accepted', 'University', 'City', 'State', 'Country')

- **Knowledge Representation:** Visualization of data through bar plots, pie chart, scatter plots, heatmaps.

Tools: Matplotlib, Seaborn

Data Structures: Arrays, Lists, Data Frames

- ✓ **Univariate Analysis:** Bar plot of top 10 states by number of colleges. Pie chart of top 10 city by number of colleges. Line Graph of top 10 established years by number of colleges.
- ✓ **Bivariate Analysis:** Bar Graph for subset data which gives first 10 colleges according to the average fee with respect to Total Student Enrollments and Total Faculty. Box Plot between average fees and college type for detecting the outliers
- ✓ **Multivariate Analysis:** Heatmap of every numerical column. Scatter plot of every numerical column.

- **Feature engineering:**

- ✓ Dividing colleges Avg fee in three categories (Low, High, Medium) on the basics of quantile and plot pie chart for the distribution.
- ✓ Creating a new feature: Student-to-Faculty Ratio

- **Pattern Identification:**

Tools: Scikit-learn

- ✓ Clustering of Colleges by K Means (Cluster 0: Colleges with moderate enrollments and moderate fees. Cluster 1: Colleges with high enrollments and high fees. Cluster 2: Colleges with lower enrollments and lower fees.)
- ✓ Anomaly Detection by Isolation Forest (The data point is considered an anomaly, outlier.1: The data point is considered a normal observation, inlier.)
- ✓ **Model Development and Training:**
  - Implemented Random Forest Classifier, decision tree, SVM Classifier to predict the fee category.
  - The model was trained using a portion of the dataset, typically 80%, with the remaining 20% used for validation.

- **Model Evaluation**:

- ✓ The performance was evaluated using the Accuracy (The overall correctness of the model's predictions), Classification Report and feature importances.
- ✓ A random forest classifier was chosen due to its robustness and ability to handle complex datasets with multiple features

- **Prediction and Listing**:

- ✓ The final trained model was used to predict the fee categories (high, medium, low) for all colle ges in the dataset.
- ✓ The colleges were then listed according to their predicted fee category. predict_fee_category F unction was made to predict fee category for a given college.

- **User-friendly Interface:**

Tools: Flask (Python), HTML

- ✓ For Front-end we have used HTML.
- ✓ API development was done by Flask.
- ✓ JSON module was used in flask to interchange data between front-end and back-end.

4. RESULT AND DISCUSSION:

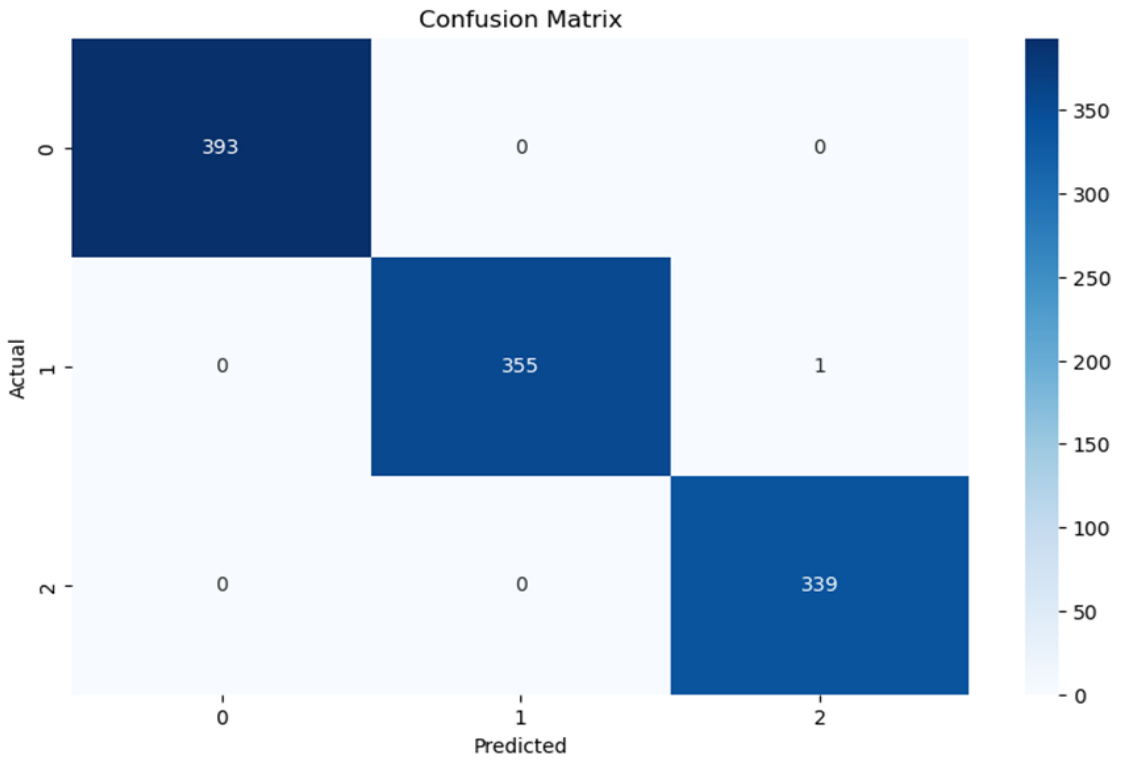
The random forest classifier achieved the following performance metrics on the validation dataset:

- **Accuracy:** 99.90808823529412

- **Classification Report:**

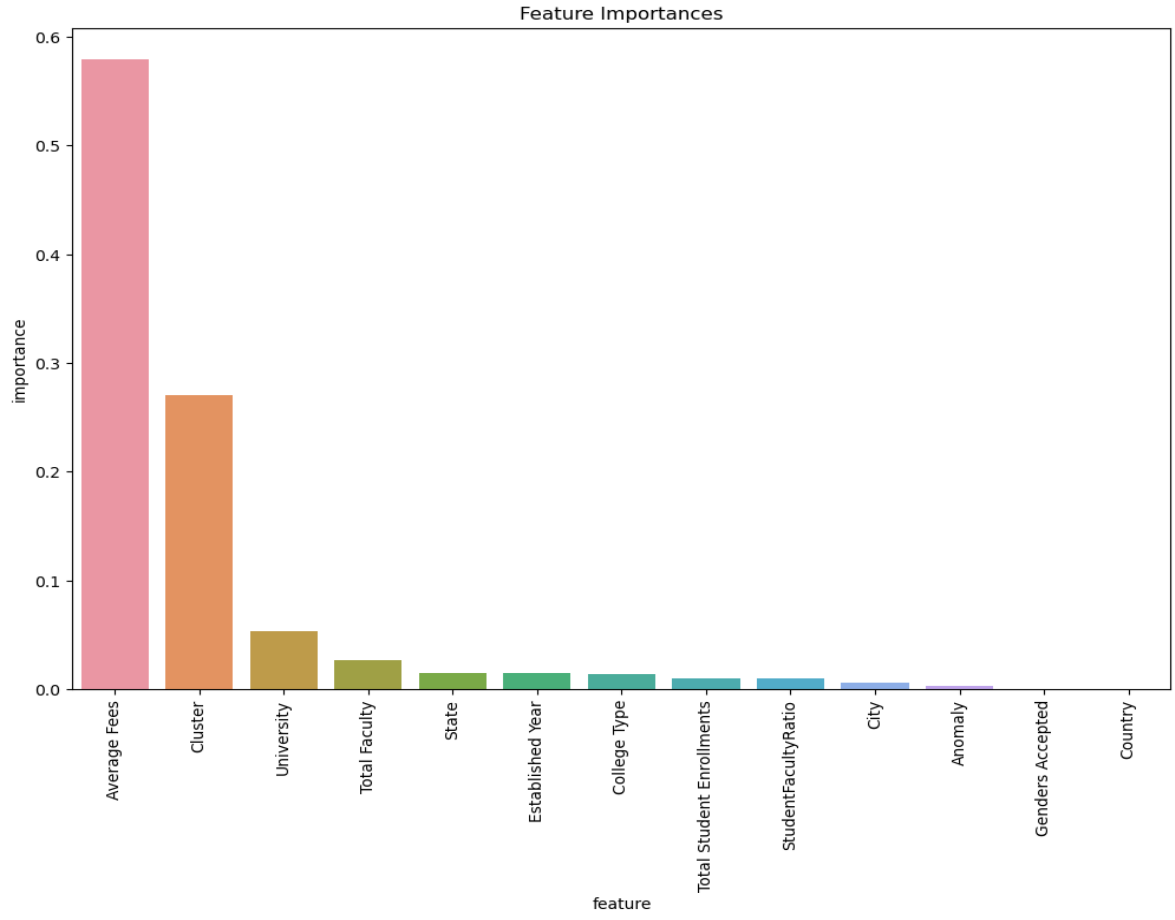
	precision	recall	f1-score	support
High	1.00	1.00	1.00	393
Low	1.00	1.00	1.00	356
Medium	1.00	1.00	1.00	339
accuracy			1.00	1088
macro avg	1.00	1.00	1.00	1088
weighted avg	1.00	1.00	1.00	1088

- **Confusion Matrix:** This is a table used to evaluate the performance of a classification algorithm. It provides a summary of the prediction results on a classification problem by comparing the actual t arget values with those predicted by the model.



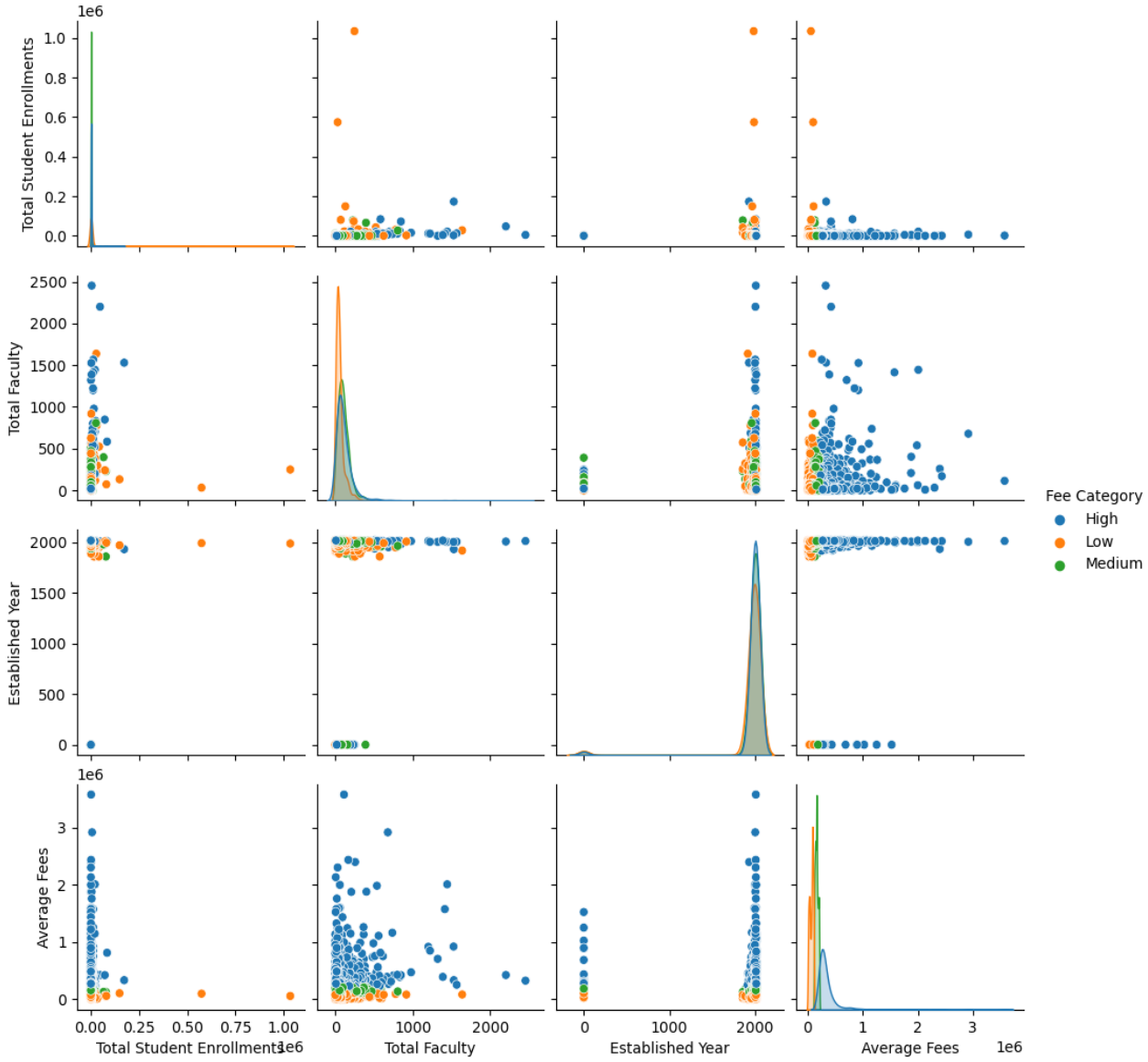
- **IMPORTANT FEATURES:**

feature	importance
Average Fees	0.579052
Cluster	0.270518
University	0.053132
Total Faculty	0.026162
State	0.014526
Established Year	0.014446
College Type	0.013952
Total Student Enrollments	0.010280
StudentFacultyRatio	0.009425
City	0.005689
Anomaly	0.002507
Genders Accepted	0.000313
Country	0.000000



These metrics indicate that the model performs well in classifying colleges into the fee categories.

- **EDA ANALYSIS OF NUMERICAL COLUMNS:**



1. CONCLUSION:

This study has demonstrated the feasibility of using random forest classifier which effectively categorize engineering college fees into high, medium, and low categories. The developed model can a ccurately predict fee categories based on a comprehensive set of features, providing valuable insights for students and institutions alike. This categorization provides valuable insights for students, educators, and policymakers in understanding the financial landscape of engineering education in India