Problem Statement

An Admission Prediction Dataset typically contains information about students applying for admission to universities or academic programs, with features (variables) related to their academic performance, extracurricular activities, demographics, and other factors that may influence the admission decision. The dataset is often used for machine learning tasks such as predicting whether an applicant will be admitted or rejected based on various input features.

Below is a general description of the features commonly found in an Admission Prediction Dataset:

1. GRE Score

Type: Numerical

Description: The score obtained by the applicant in the Graduate Record Examination (GRE), a standardized test used for admissions in graduate schools.

Range: Typically 260-340 (Higher scores indicate better performance).

2. TOEFL Score

Type: Numerical

Description: The score obtained in the Test of English as a Foreign Language (TOEFL), which assesses the English language proficiency of non-native speakers.

Range: Typically 60-120 (Higher scores indicate better English proficiency).

3. University Rating

Type: Ordinal (Categorical)

Description: A rating given to the university or institution from which the applicant is applying. The rating could be on a scale from 1 to 5 or any other scale that reflects the university's overall reputation.

Values: 1, 2, 3, 4, 5 (Higher values represent better-rated universities).

4. SOP (Statement of Purpose) Strength

Type: Ordinal (Categorical)

Description: A score indicating the quality of the applicant's Statement of Purpose (SOP), which is a critical document in the admission process.

Values: Typically rated on a scale from 1 to 5 (1 being weak and 5 being strong).

5. LOR (Letter of Recommendation) Strength

Type: Ordinal (Categorical)

Description: A score or rating indicating the strength of the applicant's Letter of Recommendation (LOR).

Values: Typically rated on a scale from 1 to 5 (1 being weak and 5 being strong).

6. CGPA (Cumulative Grade Point Average)

Type: Numerical

Description: The applicant’s undergraduate CGPA or final grade point average. This is a measure of the applicant's academic performance during their undergraduate studies.

Range: Typically between 0.0 and 10.0 (Higher CGPA indicates better academic performance).

7. Research Experience

Type: Binary (Categorical)

Description: Whether the applicant has research experience or not. This is an important factor for many graduate programs.

Values: 0 (No research experience) or 1 (Has research experience).

8. Admission Decision (Target Variable)

Type: Binary (Categorical)

**Important points to remember**

* Use only classical machine learning models. Strictly no deep learning models.
* Structure and organize the code into sections. Provide your observations at the end of each section.

**Assignment workflow**

1. **Import Libraries/Dataset** 
   1. Download the dataset.
   2. Import the required libraries and packages.
2. **Data Visualization and Exploration [1M]**
   1. Print 5 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   2. Provide appropriate data visualizations to get an insight about the dataset.
   3. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks.**
3. **Data Pre-processing and cleaning [2M]**
   1. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Perform feature encoding. Mention the pre-processing steps performed in the markdown cell.
   2. Apply appropriate feature engineering techniques (if applicable). Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**. Explore a few techniques for identifying feature importance for your feature engineering task.
4. **Model Building [5M]**
   1. Split the dataset ([KDDTrain+.txt](https://github.com/jmnwong/NSL-KDD-Dataset)) into training and test sets. **Answer without justification will not be awarded marks.** [1M]
      1. Train = 80 % Test = 20%
      2. Also, try to split the dataset with different ratios of your choice.
   2. Build model using Logistic model and decision tree  [4 M]
      1. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.
5. **Performance Evaluation [2M]**
   1. Compare the performance of the Logistic Regression and Decision Tree models using appropriate evaluation metrics.
   2. Provide insights into which model performs better and why. **Answer without justification will not be awarded marks.**

***For clarifications, contact Srinivasa Rao Kongara (kongaras@wilp.bits-pilani.ac.in)***