**Prediction to get an Admission using Machine Learning**

**Problem Statement: What is the chance of Admit of the student in the university.**

**Dataset Overview:**

The dataset consists of 400 entries and 9 rows.

**Serial No.:** This column likely represents a unique identifier or serial number assigned to each applicant.

**GRE Score:** Represents the GRE (Graduate Record Examination) score of the applicant. GRE scores are often used as a standardized measure for assessing readiness for graduate-level academic work.

**TOEFL Score:** Stands for Test of English as a Foreign Language score.It measures the English language proficiency of non-native English speakers applying to English-speaking universities.

**University Rating:** This column could denote the rating or reputation of the university from which the applicant obtained their undergraduate degree. Ratings are often based on various factors such as research output, faculty quality, and academic reputation.

**SOP (Statement of Purpose):** Refers to the applicant's Statement of Purpose, which is a personal statement outlining their academic interests, career goals, and reasons for applying to the program.

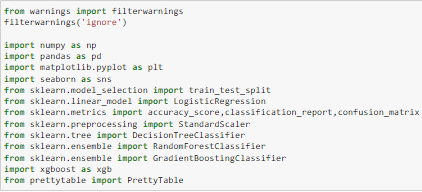
**LOR (Letter of Recommendation):** Indicates the strength of recommendation letters provided by referees who can attest to the applicant's academic abilities, work ethic, and potential for success in graduate studies.

**CGPA (Cumulative Grade Point Average):** Represents the applicant's cumulative GPA from their undergraduate studies. It provides an overall measure of academic performance.

**Research:** Binary variable (0 or 1) indicating whether the applicant has research experience. Research experience is often valued in graduate admissions, particularly for research-focused programs.

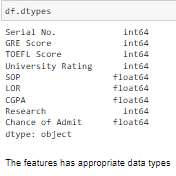
**Chance of Admit:** Represents the probability or likelihood of the applicant being admitted to the graduate program. This column could be derived from historical admission data or predictive models based on the other variables listed.

**Imported Necessary Libraries**

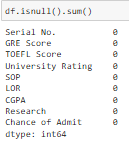


**Exploratory Data Analysis**

**Check for the data types:**



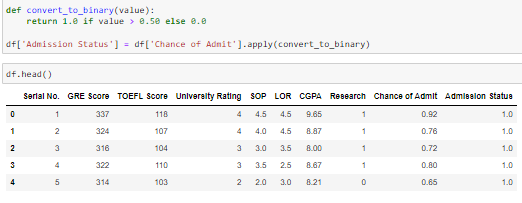
**Check for Null values:**



There are no outliers present in the data

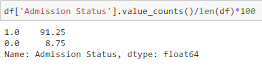
**Binarization:**

I want to convert the target column ‘Chance of Admit’, which is currently represented as a probability, into binary form. Applicants with a probability greater than 0.50 will be considered to have a chance of admission, while those with a probability less than or equal to 0.50 will be considered to have no chance of admission.



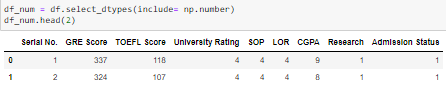
Note: Drop the column “Chance of Admit”.

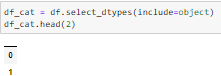
**Assess for imbalance in the target variable**



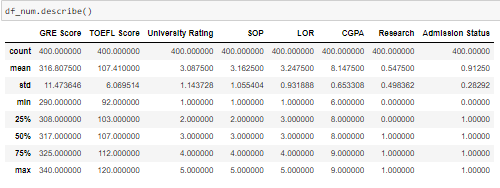
There is a huge imbalance in the data.

**Distinguishing between categorical and numerical data:**





**Summary Statistics**

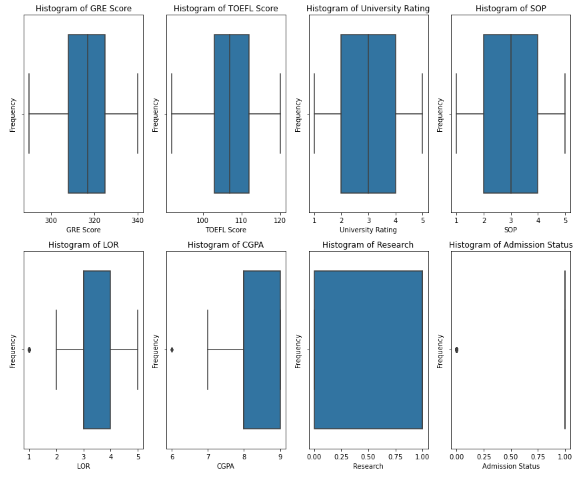


The average GRE score among applicants is 316.

Applicants have an average TOEFL score of 107

The average CGPA of applicants is 8.5

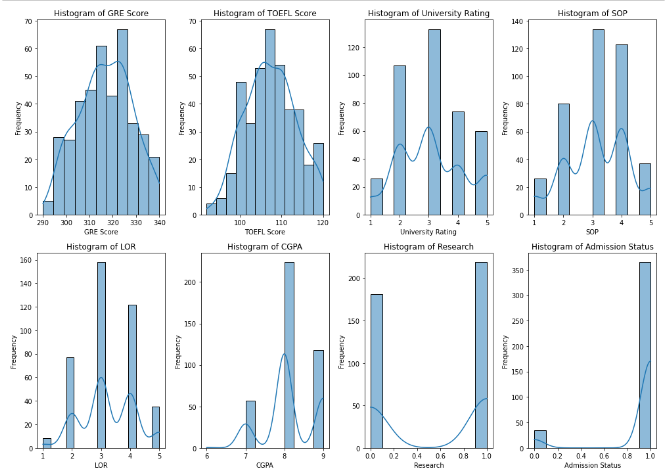
**Check for the outliers in the data:**



**Univariate Analysis:**

**Numerical Columns:**

* **Histogram:**



**Interpretation:**

**GRE:** Over 60 students scored 320 or higher on the GRE.

**TOEFL:** Over 60 students scored 110 or higher on the TOEFL.

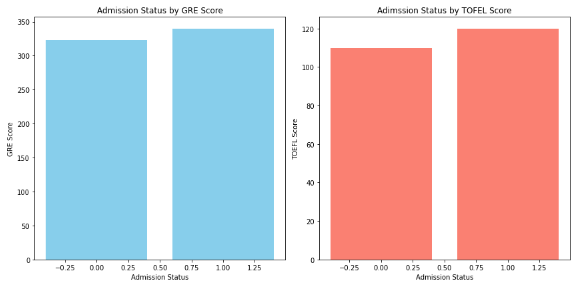
**University Rating:** The majority of university ratings fall between 3 and 3.5.

**SOP:** Over 60 students received SOP scores of 4 to 5.

**LOR:** Over 60 students received LOR scores of 3.

**CGPA:** Most students have CGPA scores ranging from 8.0 to 9.0.

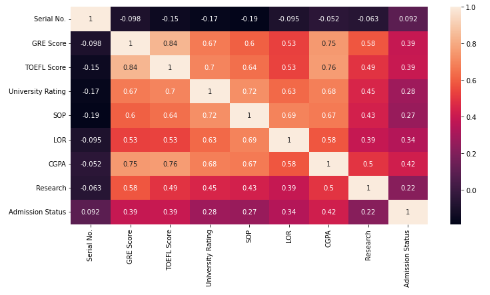
**Bivariate Analysis:**



The students who have 300 to 350 GRE score are having 50% to 100% chance of getting admission.

The students who have 100 to 120 TOEFL score are having 50% to 100% chance of getting admission.

**Multivariate Analysis:**



**Base Model Evaluation**

**Initial Logistic Regression Model**

The base model employed for initial evaluation was a logistic regression model without any preprocessing steps such as data scaling or outlier removal. This approach aimed to establish a benchmark performance to assess subsequent improvements. The model yielded 36% of f1\_score for class 0 and 95% of f1\_score for class 1, suggesting huge imbalance in the training.

Confusion Matrix:



True Positive: 2 instances correctly predicted as positive.

False Positive: 3 instances incorrectly predicted as positive.

False negative: 4 instances incorrectly predicted as negative.

True negative: 71 instances correctly predicted as negative.

**Exploration of Data Preprocessing techniques**

To mitigate imbalance observed in the base model, various preprocessing techniques were explored. This included feature scaling to enhance the model’s capabilities.

**Ensemble Models:**

**Random Forest Classifier:**

The model employed for evaluation was a random forest model with preprocessing steps such as data scaling. This approach aimed to establish a benchmark performance to assess subsequent improvements. The model yielded 0% of f1\_score for class 0 and 94% of f1\_score for class 1, suggesting huge imbalance in the training.

Confusion Matrix:



True Positive: 0 instances correctly predicted as positive.

False Positive: 5 instances incorrectly predicted as positive.

False negative: 4 instances incorrectly predicted as negative.

True negative: 71 instances correctly predicted as negative.

**Gradient Boost:**

The model employed for evaluation was a gradient boost model with preprocessing steps such as data scaling. This approach aimed to establish a benchmark performance to assess subsequent improvements. The model yielded 31% of f1\_score for class 0 and 94% of f1\_score for class 1, suggesting huge imbalance in the training.



True Positive: 2 instances correctly predicted as positive.

False Positive: 3 instances incorrectly predicted as positive.

False negative: 6 instances incorrectly predicted as negative.

True negative: 69 instances correctly predicted as negative.

**Final Evaluation**

After thorough evaluation, the Gradient Boost model scaled input data was chosen as the final model. The decision was based on its ability to maintain strong predictive performance while mitigating imbalance issues observed in earlier stages. The model’s performance metrics and stability make it suitable for set attributes.

