**Documentation**

1. **Dataset Information**

**Dataset name:**

Job Applicant Evaluation Dataset.csv

**Dataset Details:**

This dataset is created for this project. These are about factors that contribute to successful hiring outcomes. By analyzing the features and their relationships with the target variable analysts can gain insights into which characteristics are associated with higher chances of being hired.

**Number of Instances:** 50

**Number of Features**: 5 (including the target)

**Target Variable:** Hired (Binary: 1 - Hired, 0 - Not Hired)

**Dataset Features:**

**Age:** The age of the job applicant.

**Years of Experience:** The number of years of professional experience the applicant has.

**Interview Performance:** A metric indicating the performance of the applicant during the interview process.

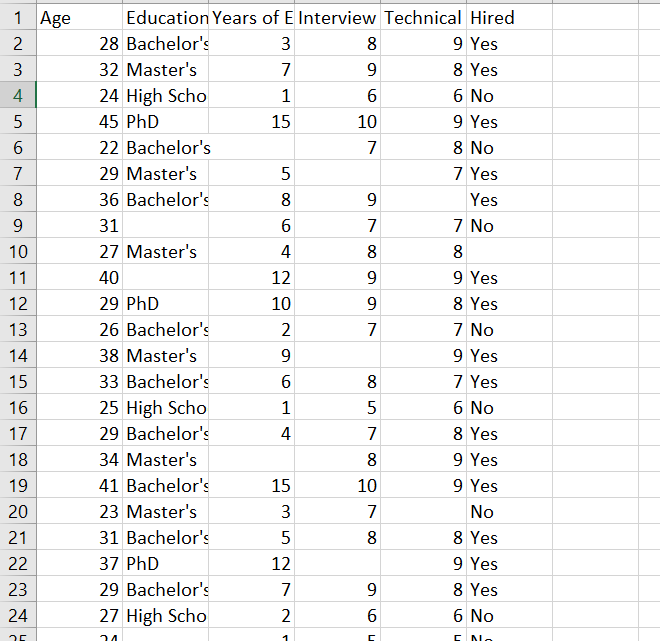
**Technical Skills Rating:** A rating reflecting the technical skills of the applicant.

**Education Level:** The education level of the applicant (e.g., High School, Bachelor's, Master's, etc.).

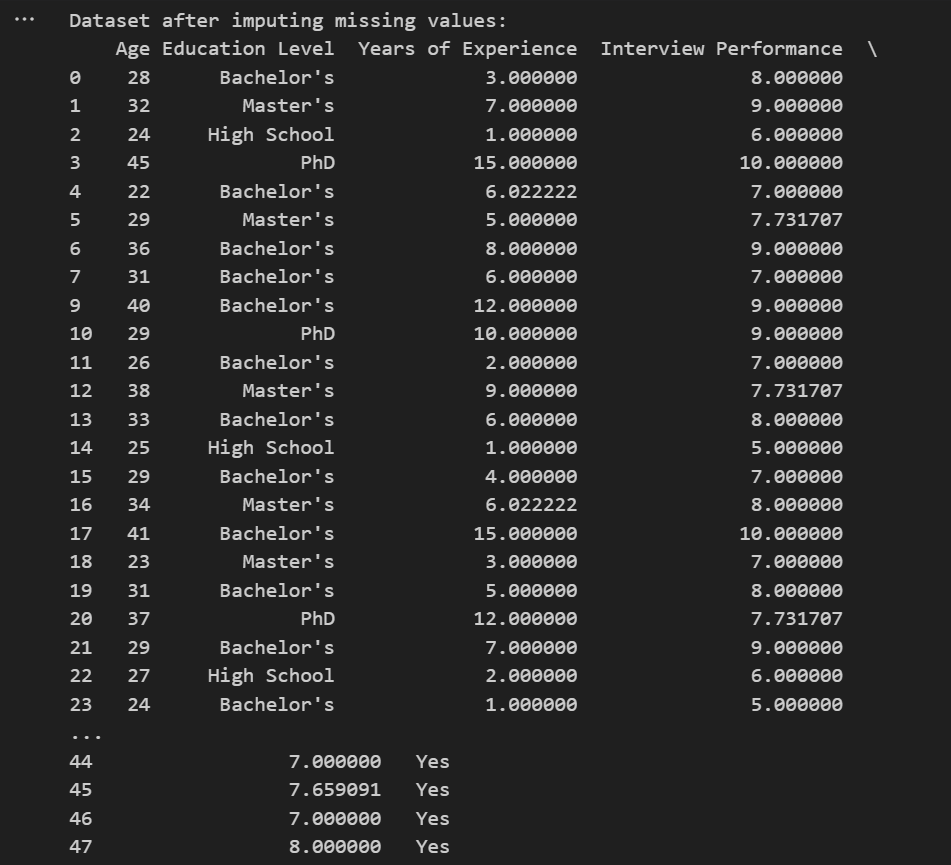
**Target:** A binary variable indicating whether the applicant was hired (1) or not (0).

1. **Data Cleaning**

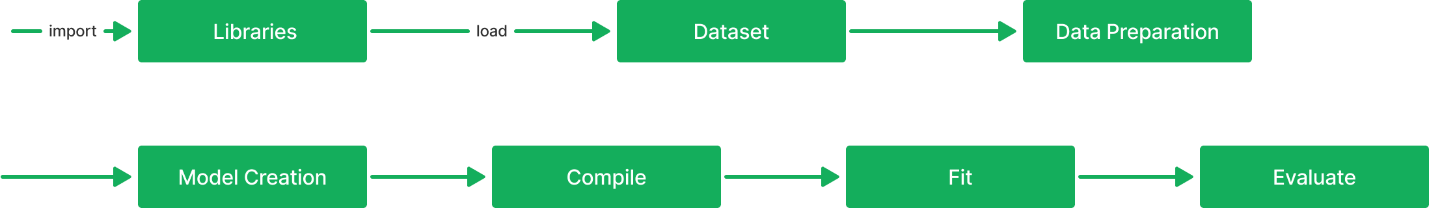
**Dataset before Data Cleaning:**

****

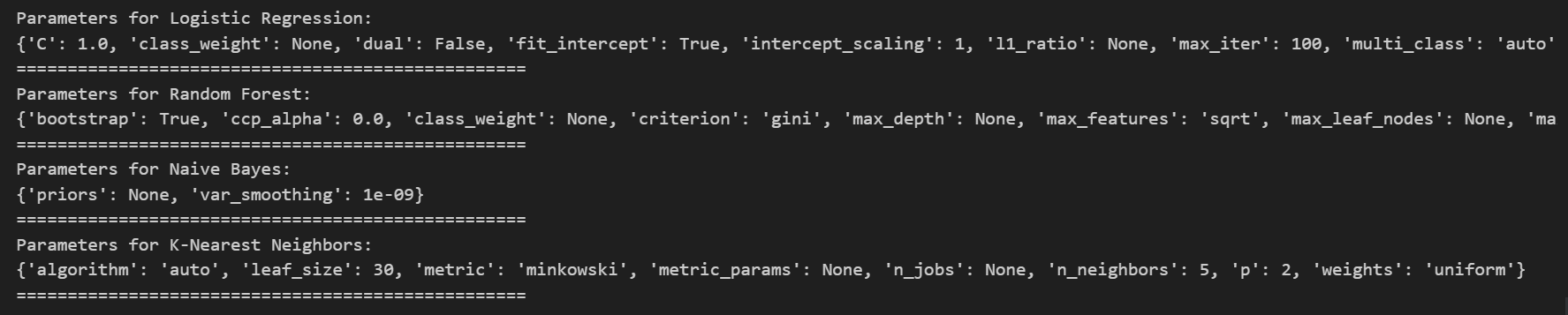
**Dataset after Imputation using Python:**

****

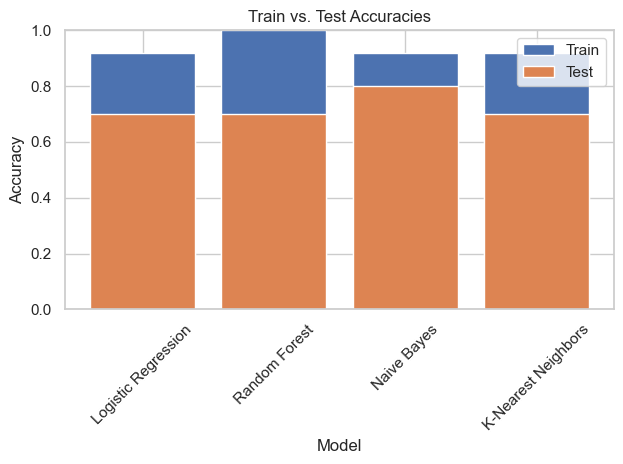
1. **Flow Chart of Processes**

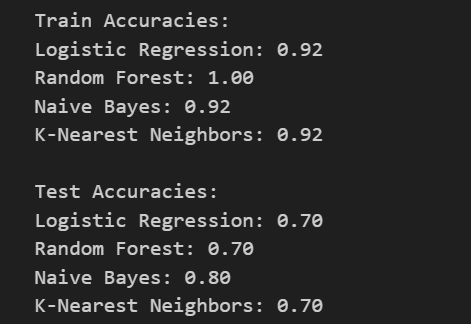
****

First, we need to import the necessary libraries (e.g., pandas, seaborn, sklearn, etc.). Then we load our dataset and prepare it for exploration and analysis. In this case, we impute the missing values with the mean for numerical features and mode for the categorical feature. After data preparation, we now proceed to creating the Machine learning models and evaluating the train and test performances of each. For this task, I have tried four algorithms which are: Logistic Regression, Random Forest, Gaussian Naïve Bayes, and KNN.

These are the parameters for each model:

1. **Results**

****

****

As seen above, the algorithm which achieved the highest accuracy in testing is the Random Forest Classifier. However, Gaussian NB achieved the highest accuracy in testing. The testing accuracies are significantly higher than the testing accuracies. This indicates overfitting. However, this is a lot due to the number of instances in the dataset which I just created. The model will have significant improvements with increased size of the dataset or using data augmentation, hyperparameter tuning, or using other machine learning algorithms that were not included in this model.