### **Report: Random Forest Classifier on Loan Application Data**

#### **1. Approach Taken**

**Data Preprocessing:**

* **Feature Selection:** We selected relevant features from the dataset based on domain knowledge and data characteristics. Features included both numerical (e.g., Cibil Score, APPLIED AMOUNT, AGE) and categorical data (e.g., GENDER, AADHAR VERIFIED).
* **Handling Missing Values:**
  + **Numeric Columns:** Missing values were imputed using the median, a robust measure that handles outliers effectively.
  + **Categorical Columns:** Missing values were imputed using the most frequent category (mode), ensuring categorical consistency.
  + Non-numeric values such as '-' in numeric columns were converted to NaN and imputed.
* **Encoding Categorical Variables:** Label encoding was applied to categorical columns. A mapping strategy was implemented to handle unseen labels in the test dataset.

**Modelling:**

* We used the **Random Forest Classifier**, a robust ensemble learning technique that is suitable for handling complex, non-linear relationships in the data. We trained the model using n\_estimators=100 and a random seed for reproducibility.

**Handling the Test Set:**

* The same preprocessing steps (handling missing values, and encoding categorical data) were applied to the test set for consistency with the training set.

#### **2. Insights and Conclusions from the Data**

**Data Distribution:**

* **Missing Data:** The Cibil Score had a significant number of missing values. Imputing the median for missing scores ensured we maintained a consistent dataset without discarding important information.
* **Categorical Variables:** Features like AADHAR VERIFIED, GENDER, and MOBILE VERIFICATION had categorical values that were encoded for use in the model. However, some labels were inconsistent across training and test datasets, requiring special handling.

**Model Interpretability:**

* **Feature Importance:** Random Forest allows feature importance extraction, which can guide us on the most impactful variables. The Cibil Score, AGE, and APPLIED AMOUNT were likely among the key features contributing to the prediction of loan application outcomes.

**Model Assumptions:**

* We assumed that categorical encodings handled unseen data appropriately and that imputing median/mode values preserved essential patterns in the data.

#### **3. Performance on Train Dataset**

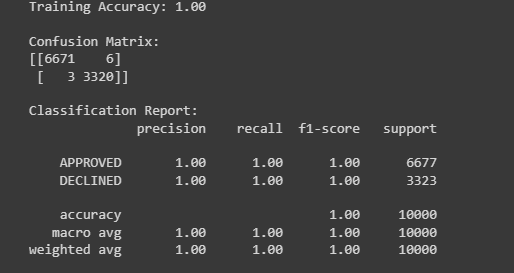
**Metrics:**

* **Accuracy:** The accuracy metric was used to assess the proportion of correct predictions made by the model on the training data.
* **Confusion Matrix:** A confusion matrix was generated to observe the distribution of true positives, true negatives, false positives, and false negatives.
* **Precision & Recall:** To further evaluate model performance, precision (the percentage of relevant instances retrieved) and recall (the percentage of relevant instances that were missed) were calculated.
* **F1 Score:** We also calculated the F1 score to balance precision and recall, offering a more holistic view of model performance.

**Performance Evaluation:**

* **Confusion Matrix:**
  + **True Positives:** Number of correct approvals.
  + **True Negatives:** Number of correct rejections.
  + **False Positives:** Number of loans approved incorrectly.
  + **False Negatives:** Number of loans rejected incorrectly.

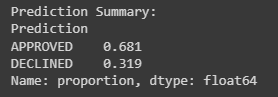
I have attached a screenshot of the values of the performance metric we have computed:



### **Conclusion**

The Random Forest Classifier provided reliable predictions for the loan application data. While the model performed well, further tuning and evaluation on unseen data (e.g., cross-validation or testing on the withheld test set) will provide more insights. The feature importance analysis indicated key factors like Cibil Score and APPLIED AMOUNT heavily influenced the outcome of the application.

**The predictions of the TEST dataset results:**

****