**Name: Mohammad Sakibul Islam ID: 1265299**

**TAKE HOME ASSIGNMENT**

The dataset comprises a decade of clinical records from 130 US hospitals, with a goal to determine the early patient readmission.

**Data Preprocessing Approach:**

- **Removing Duplicate Records:** The dataset consists 0f 101,766 rows, with 71,518 unique patients identified by patient numbers (patient_nbr). To maintain data integrity, duplicate entries based on patient numbers are removed to ensure each patient is represented only once in the dataset.
- **Handling Missing Data:** Columns with high proportion of missing values, including weight, payer code and medical speciality, are dropped and missing race data is imputed with the most frequent race category.
- **Removing Uninformative Features:** Irrelevant features, such as encounter id, patient number, and highly imbalanced features (where over 98% of data points with the same value) are removed to optimize classification towards target variable.
- **Handling Diagnostic Features:** The diagnostic features (‘diag_1’, ‘diag_2’,’diag_3’) are categorized using the ICD-9 codes (https://en.wikipedia.org/wiki/List_of_ICD-9_codes), which actually improves the interpretability and relevance.
- **Mapping for Admission types, Admission Sources and Discharge Types:** The admission type IDs, Admission Sources and Discharge Type IDs are replaced with corresponding descriptive categories, while also consolidating similar meanings into one category during mapping. For Instance, the conversion includes mapping similar terms like ‘Urgent' to ‘Emergency’, and grouping ‘NaN’ and ‘Not Mapped’ as ‘Not available’.
- **Handling Age Feature:** The age ranges are mapped to numerical formats, by assigning each range a corresponding numeric value, with the lower range starting at 0 and incrementing accordingly as higher age correlates with a greater likelihood of readmission.
- **Handling Outliers for numeric features:** Based on boxplot visualizations, outliers exceeding specified thresholds are removed to ensure data consistency.
- **Handling drug medication features:** From visualization, it is observed that certain drug medication features predominantly represent one class, significantly outnumbering instances of the other class. Due to this imbalance, these columns are dropped.
- **Handling Glucose Serum Feature:** Individuals with higher glucose serum levels(‘>200’ and ‘>300’) are more likely to be readmitted than those with normal levels or None, as inferred from the visualization. Therefore, ‘max_glu_serum’ column is mapped to numeric values, with ‘None’ is mapped to 0, ‘Norm’ is mapped to 1, ‘>200’ is mapped to 2 and ‘>300’ is mapped to 2..
- **Handling diabetes medication feature:** A potential correlation between the likelihood of readmission and diabetes medication feature is indicated by higher readmission rates observed from visualization when diabetes medication is prescribed. So, ‘diabetesMed’ column is mapped to numeric values, with ‘No’ is mapped to 0 and ‘Yes’ is mapped to 1.
- **Handling change of medications feature:** Higher readmission rates are observed from visualization when there is change in the diabetes medication. So, ‘change’ feature is transformed to numeric, with ‘No’ is mapped to 0 indicating no change and ‘Yes’ is mapped to 1 indicating change in medication.

- **Handling Target Variable:** The target readmitted variable is converted into numerical values, where 'NO' is mapped to 0 indicating no readmission, '>30" to 1 indicating readmission after 30 days, and '<30' to 2 indicating readmission within 30 days.
- **Handling Categorical Features:** One-hot encoding is applied to the categorical data to transform categorical variables into numeric format.
- **Visualization:** A heatmap is used to display the correlations among numerical columns and categorical features are visualized with countplot to observe their relationship with the target variable.
- **Addressing class imbalance with SMOTE:** SMOTE (Synthetic Minority Over-Sampling Technique) is applied to mitigate the class imbalance. SMOTE generates synthetic samples of the minority class to ensure a balanced class distribution, which actually helps to prevent bias towards the majority class. So, oversampling using SMOTE improves the generalization and also prediction capabilities.

The dataset is divided into training and testing subsets, with 10% of the data allocated for testing purposes.

**Model Accuracy Analysis and Comparison:**

| Model Name | Training Accuracy | Testing Accuracy | F1-score(micro) |
|---|---|---|---|
| **Logistic Regression** | **0.634** | **0.622** | **0.622** |
| **Decision Tree** | **0.638** | **0.572** | **0.572** |
| **K Nearest Neighbour** | **0.719** | **0.546** | **0.546** |
| **Naive Bayes** | **0.576** | **0.531** | **0.531** |
| **Support Vector Machine(SVM)** | **0.675** | **0.617** | **0.617** |
| **Random Forest** | **0.785** | **0.612** | **0.612** |
| **Gradient Boosting** | **0.736** | **0.609** | **0.608** |
| **Bagging Classifier** | **0.689** | **0.604** | **0.604** |
| **Adaboost Classifier** | **0.775** | **0.589** | **0.589** |

Through the visualization of ROC curves and calculating AUC(Area Under Curve) values, the model's ability to correctly classify instances from each class is assessed in this project. A heatmap representation of the confusion matrix is displayed which provides a visual assessment of the model's effectiveness, by showing the counts of correctly and incorrectly predicted instances. Finally, in order to compare the effectiveness of different models, two bar plots are displayed: one demonstrating testing accuracy and the other illustrating F1-scores.

**Conclusion:** By analyzing the testing accuracy and F1- scores, it is observed that Logistic Regression, SVM, Random forest, Gradient Boosting and Bagging Classifier exhibit better performance compared to Decision Tree, KNN, Naive Bayes and Adaboost Classifier. The best accuracy that I obtained is 0.622 .

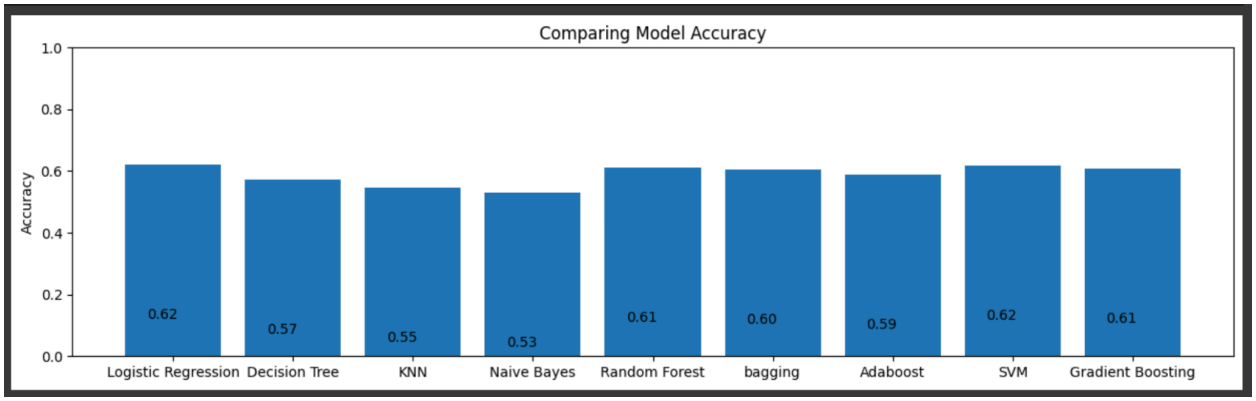


Fig: Model Testing Accuracy Comparison

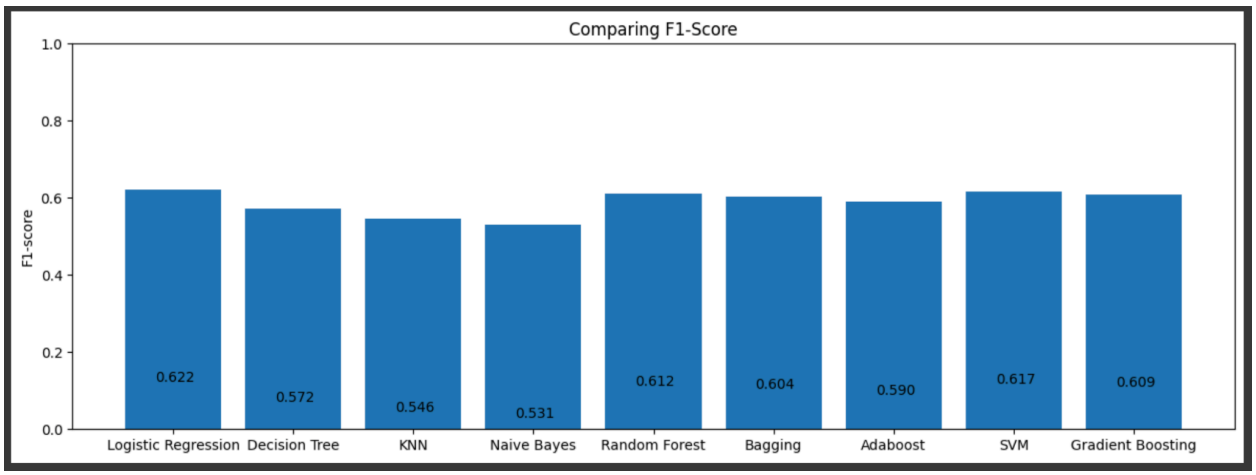


Fig: F1-Score Comparison