DATA*6300 - ANALYSIS OF BIG DATA

Dr. Taiwo Omomule

PROJECT 1

PREDICTING RE-ADMISSION OF DIABETES PATIENT

NAME : Nidish Murugan
Student ID : 1295078

Problem Statement:
We aim to predict patient readmission of patients using the "*Diabetes 130-US hospitals dataset spanning the years 1999-2008*". The dataset includes 101,766 entries and 50 features, encompassing patient details, drug information, diagnostic results, and readmission status.

Executive Summary:
The project focuses on predicting patient readmission and involves several stages: Exploratory Data Analysis, Data Pre-processing, Data Cleaning, Feature Engineering, Modeling, Model Selection, Comparative Analysis and Conclusion

1. Data Pre-Processing:

1. **Addressing Missing Values**:
   a. *'Weight'*, *'payer_code'*, *'medical_specialty'* were dropped due to significant missing values.
   b. *'citoglipton'* and *'examide'* columns with constant values were removed.
   c. Rows with missing *'diag_1'*, *'diag_2'*, *'diag_3'* were grouped using ICD9 codes and *'gender'* values were dropped.

2. Feature Engineering:

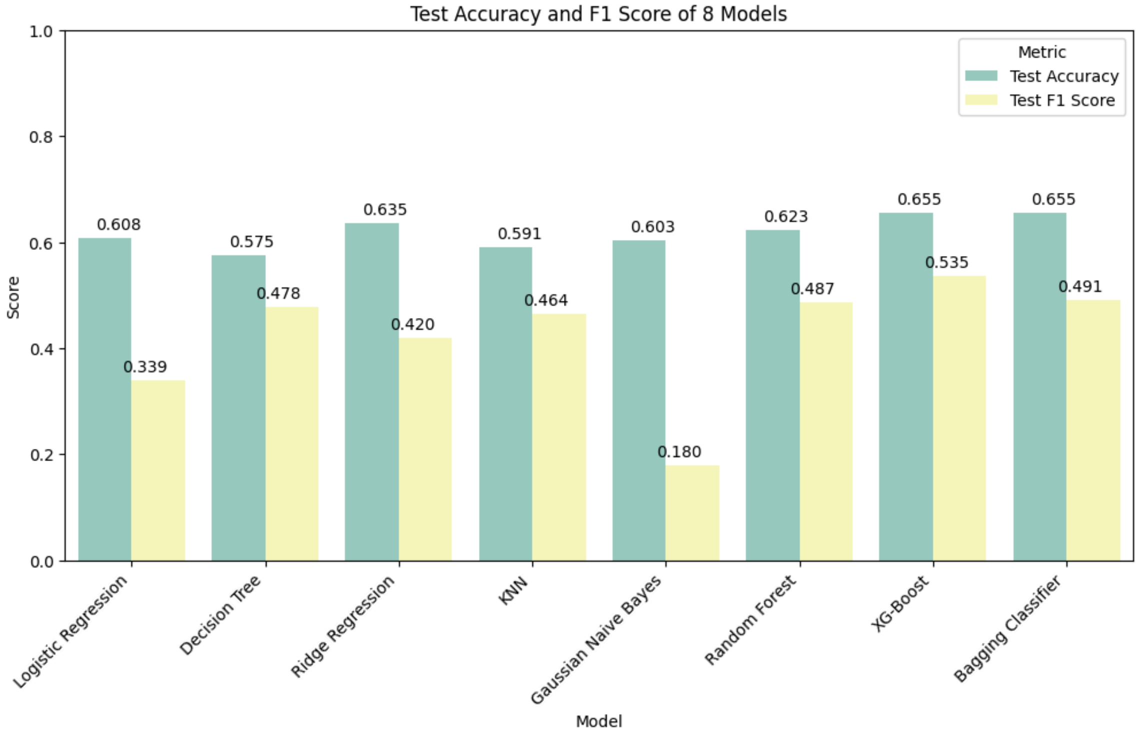
1. **Age** : Ranges in *'age'* were replaced with median values.
2. **Adding Feature** : *'hospital_visits'* was introduced by summing 'number_outpatient', 'number_emergency', and 'number_inpatient'.
3. **Reducing Unique Values** : *'discharge_disposition_id'*, *'admission_source_id'*, 'admission_type_id' were grouped and mapped.
4. **Binary Conversion** : *'diabetesMed'*, *'gender'*, *'change'*, and drug columns were converted into binary values.
5. **Dropping Duplicates** : Duplicate entries based on *'patient_nbr'* were removed.
6. **Class Imbalance:** The class imbalance was addressed by converting 'NO' to '0' and '>30' and '<30' to '1'.
7. **Grouping and Mapping 'Diag'** : *'diag'* columns were grouped based on research papers for ICD9 code mapping.

3. Modeling:

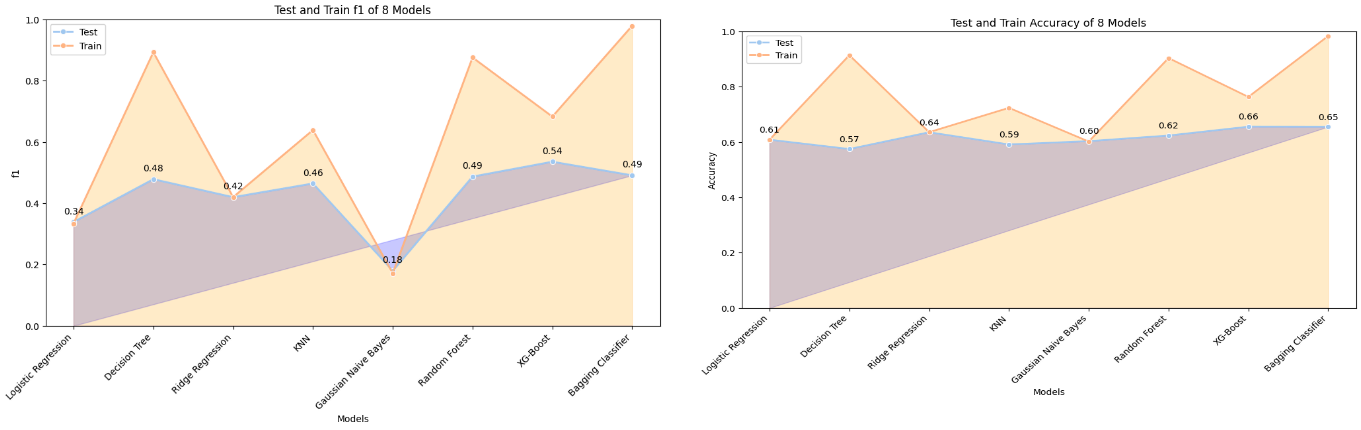
1. **Single Models**: Logistic Regression, Decision Tree, Ridge Regression, KNN, Gaussian Naive Bayes.
2. **Ensemble Models**: Random Forest, XG-Boost, Bagging Classifier.

4. Model Selection & Score Analysis :

Accuracy And F1-Score of the 8 Models



TEST - TRAIN ACCURACY & F1-SCORE FOR 8 MODELS



5. Conclusion:

Based on the comprehensive analysis conducted, XG-Boost emerges as the optimal choice with 66% accuracy and 53% F1 Score, showcasing superior performance in both test accuracy and F1-Score. This harmonious balance between accuracy and F1-Score positions XG-Boost as the preferred model for the current dataset and establishes it as a robust choice for future datasets within similar contexts.