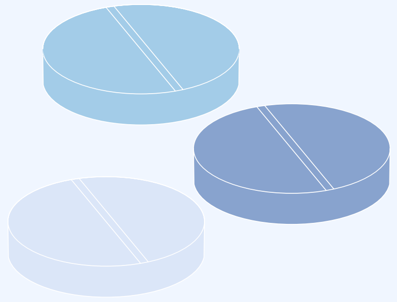
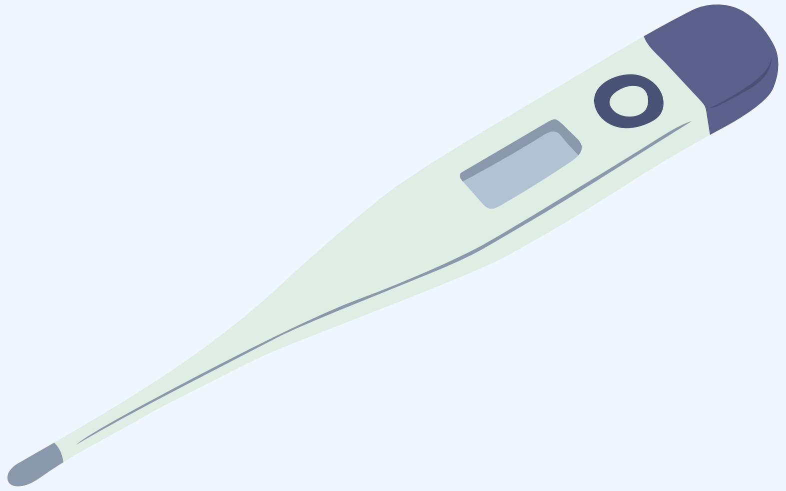
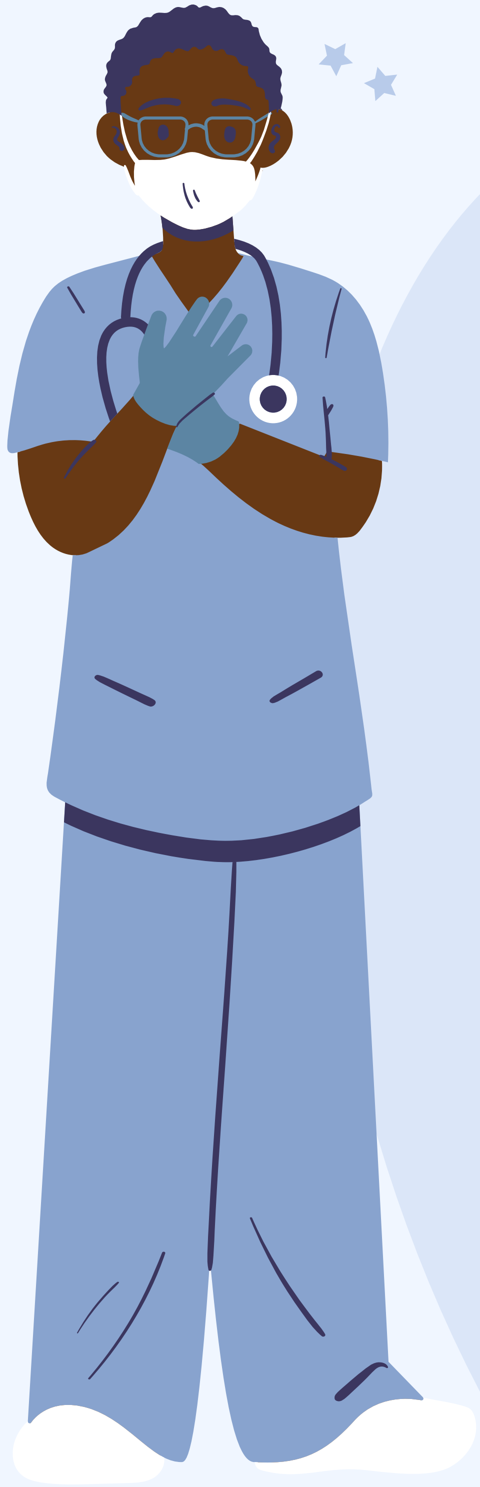
DA 204o: Data Science in Practice

Comparative Model Analysis For Diabetes Prediction





Our Team



Comparative Model Analysis of Diabetes Prediction

Kapil Soni, DSBA, kapilsoni@iisc.ac.in
Mahendra Mahajan, DSBA, mmahendra@iisc.ac.in
Malu Jayachandran, DSBA , maluj@iisc.ac.in
Tanzimur Rahman, DSBA, tanzimrahman@iisc.ac.in

Data Science Canvas

Project: Comparative Model Analysis for Diabetes Prediction

Team: 17

Problem Statement

Business Case & Value Added
Which business case should be analyzed and what added value does it generate?

Enhancing patient outcomes through early detection and personalized care, optimising healthcare resources, and reducing costs. By adopting predictive models, healthcare providers can transform diabetes management, contributing to a healthier population and more sustainable healthcare system. This approach aligns with broader public health goals and provides significant value to patients, providers, and society.

Data Landscape
Which data is required for this and which is already available? Which additional data has to be collected?

Almost all essential data is available through CDC and related health databases,like BMI, other comorbidilities,general health related data etc. However, additional data collection, particularly concerning lifestyle, family history, and genetic markers, can significantly enhance the predictive model's accuracy and utility.

Model Selection
Which analysis methods can be considered on the basis of the specific data landscape and the business case?

Logistic Regression, Random Forest, and XGBoost are strong candidates due to their balance of performance, interpretability, and flexibility. These models align well with the goal of predicting diabetes for early intervention and prevention, offering insights that are actionable in clinical settings.

Model Requirements
Which model requirements must be complied with in order to obtain a valid model?

Ensure high-quality data through thorough preprocessing, addressing missing values, and ensuring feature relevance. The model must be interpretable, allowing healthcare professionals to understand and trust the predictions made, which is crucial in clinical decision-making. Robustness and generalization need to be prioritized by validating the model's performance on unseen data through techniques like cross-validation. Additionally, the model must comply with ethical standards, safeguarding patient privacy and adhering to regulations to maintain trust and integrity in its application.

Software & Libraries
Which software should be used? Is there already a standard solution? Which libraries are used?

- **Python**
- **Jupyter Notebooks**
- **Pandas**
- **Numpy**
- **Matplotlib**
- **scikit-learn**
- **Seaborn**
- **ydata_profiling**
- **Plotly**
- **SciPy**

Skills
What skills are needed to provide the data and model development?

Data Analysis and Manipulation
Statistical Knowledge
Machine Learning and Model Development
Programming Skills
Data Visualization
Domain Knowledge in Healthcare/Diabetes
Ethical and Regulatory Understanding
Problem-Solving and Critical Thinking

Execution & Evaluation

Model Evaluation
Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?

To ensure the reliability and effectiveness of a diabetes prediction model, key indicators such as accuracy, F1 score, precision, recall, ROC AUC, and log loss require quality control and validation. Accuracy measures overall correctness but should be evaluated alongside other metrics, especially in imbalanced datasets. The F1 score balances precision and recall, indicating the model's ability to manage false positives and negatives. Precision focuses on minimizing false positives, while recall assesses the model's effectiveness in capturing true positive cases. A high ROC AUC suggests strong class discrimination, reflecting the model's ability to differentiate between classes. Log loss evaluates probability calibration, with lower values indicating predictions that closely match actual outcomes. Real-time monitoring can be necessary, especially in clinical settings, to ensure consistent performance and timely predictions, allowing for adjustments as needed to maintain model efficacy.

Data Storytelling
What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?

To effectively communicate the results of a diabetes prediction model to the target group, clarity and simplicity are paramount, ensuring that data is presented in an easily understandable manner without technical jargon. Relevance is key, focusing on insights that directly impact decision-making, such as risk factors and predictive accuracy. Interpretability is crucial, providing context and explanations for the model's predictions to ensure the audience understands their implications. Visual representation through charts and graphs can help convey data trends and patterns quickly and effectively. It's important to present actionable insights, offering recommendations based on the data to guide decision-making. Credibility is essential, so sharing information about data sources and validation processes can build trust in the findings. Finally, using storytelling techniques and real-world examples can engage the audience, making the data narrative compelling and impactful.

Data Collection & Preparation

Data Selection & Cleansing
Which of the available data is relevant? Do the data have to be cleaned up?

In developing a diabetes prediction model, relevant data includes BMI information age, blood pressure, BMI, lifestyle factors are important. This data is crucial for identifying risk factors and patterns associated with diabetes onset. Dataset available was pretty clean without much requirements for data cleaning and pre proceeing

Data Integration
In which system should the data from different sources be migrated?

Not much applicability of data intergration with this dataset and model development.

Data Collection
How and with which methods should additionally required data be collected? What properties has this data to fulfil?

No Additional dataset needs to be collected at this point.

Explorative Data Analysis
Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.

1. Target Feature of the Data is Diabetes_binary.

2. dataset has 15 Discrete type and 7 continuous type feature variables.

3. Dataset dose not have missing values(null values).

4. major feature variables for Diabetes are : HIghBP , HighChol , BMI, PhysicalActivity , GenHlth , MentHlth , PhysHlth , Age , Eduation and Income.

Background Of The Problem

Background of the problem

Early detection of diabetes and management are crucial for preventing many complications and improving the quality of life for individuals with diabetes.

There are many datasets which includes various health metrics and lifestyle factors that can potentially indicate the onset of diabetes.
We can develop predictive models to identify at-risk individuals by analysing these indicators, enabling early intervention and proactive management.

Why is it important?

Health Impact: Early detection of diabetes can prevent or delay complications associated with the disease, thereby improving patient outcomes and quality of life.
Economic Impact: Diabetes management and complications contribute to significant healthcare costs. Early diagnosis and intervention can reduce these costs by minimizing the need for extensive medical treatment.
Public Health: Identifying at-risk populations can help in designing targeted public health strategies and campaigns to promote healthier lifestyles and reduce the prevalence of diabetes.
Personalized Medicine: Predictive models can enable personalized healthcare plans, tailored to an individual's risk profile, thereby enhancing the effectiveness of interventions.

Project Objective

- Predict the probability of Diabetes diagnosis from the given data
- To help in quick diagnosis and preventive measures by early prediction
- Compare the performance of different Machine Learning models and find the best model for the objective

Data Collection and Preparation

Data source(s) (where it's from, how it was collected)

1. CDC Diabetes Health Indicators Dataset

Source: UCI Machine Learning Repository
URL: CDC Diabetes Health Indicators
Collection Method:
This dataset is derived from the Behavioural Risk Factor Surveillance System (BRFSS) survey
conducted by the Center for Disease Control and Prevention(CDC).
The data were collected via telephone surveys.

Description of the data (features, size, format)

The Diabetes Health Indicators Dataset contains healthcare statistics
and lifestyle survey information about people in general and their diabetes diagnoses.

Features – 21
Instances – 253680
Format – CSV

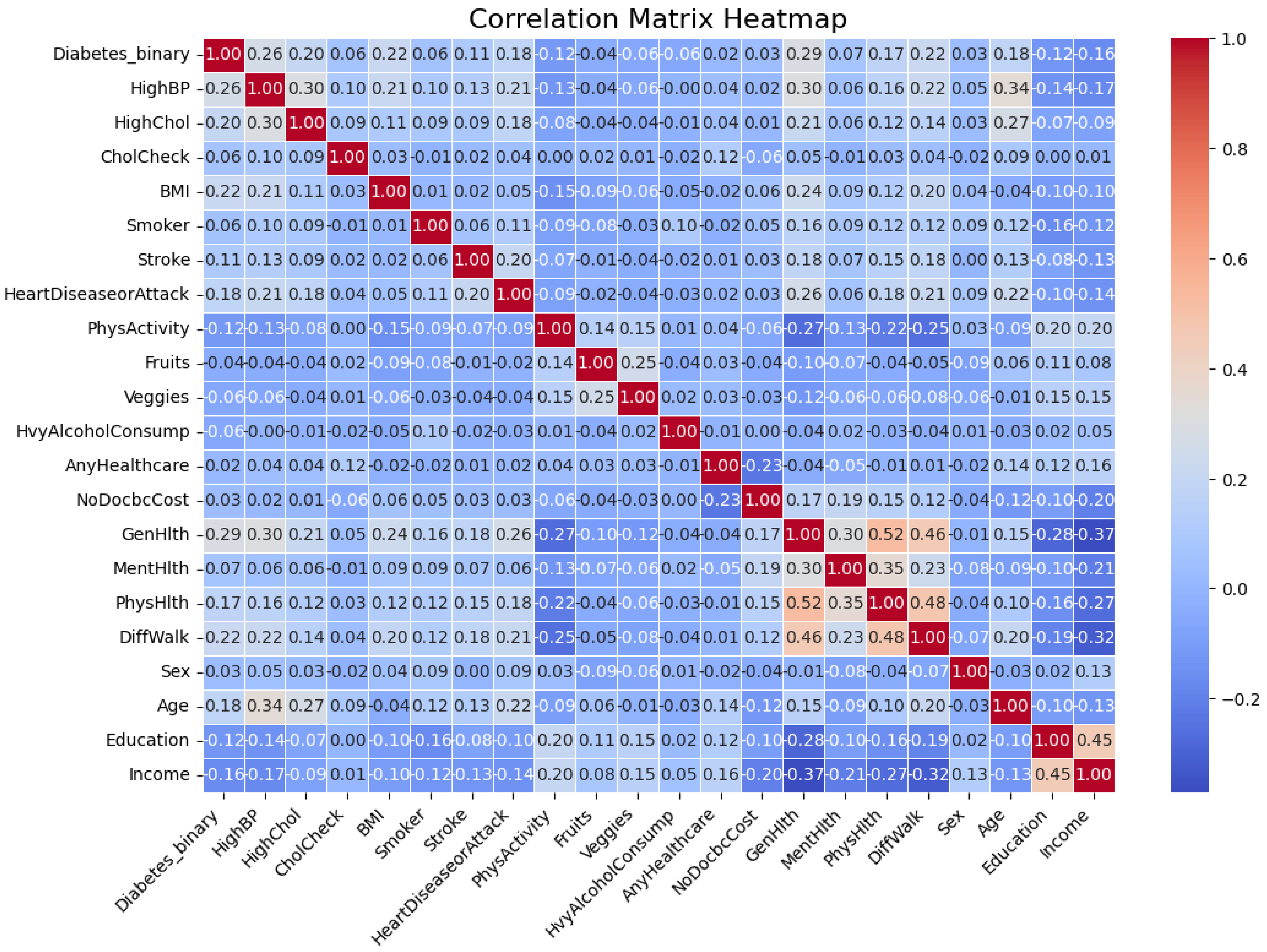
Exploratory Data Analysis

Variable Name	Role	Type	Description	Missing Values
ID	ID	Integer	Patient ID	no
Diabetes_binary	Target	Binary	0 = no diabetes 1 = prediabetes or diabetes	no
HighBP	Feature	Binary	0 = no high BP 1 = high BP	no
HighChol	Feature	Binary	0 = no high cholesterol 1 = high cholesterol	no
CholCheck	Feature	Binary	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years	no
BMI	Feature	Integer	Body Mass Index	no
Smoker	Feature	Binary	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes	no
Stroke	Feature	Binary	(Ever told) you had a stroke. 0 = no 1 = yes	no
HeartDiseaseorAttack	Feature	Binary	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes	no
PhysActivity	Feature	Binary	physical activity in past 30 days - not including job 0 = no 1 = yes	no
AnyHealthcare	Feature	Binary	Have any kind of health care coverage,including health insurance, prepaid plans such as HMO, etc 0 = no 1 = yes	no
NoDocbcCost	Feature	Binary	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes	no
GenHlth	Feature	Integer	Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor	no
MentHlth	Feature	Integer	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days	no
PhysHlth	Feature	Integer	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days	no
DiffWalk	Feature	Binary	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes	no
Sex	Feature	Binary	Sex 0 = female 1 = male	no
Fruits	Feature	Binary	Consume Fruit 1 or more times per day 0 = no 1 = yes	no
Veggies	Feature	Binary	Consume Vegetables 1 or more times per day 0 = no 1 = yes	no
HvyAlcoholConsump	Feature	Binary	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes	no
Age	Feature	Integer	Age 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older	no
Education	Feature	Integer	Education Level Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)	no
Income	Feature	Integer	Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,0005 = less than$35,000 8 = \$75,000 or more	no

Exploratory Data Analysis

Steps Taken:

- No missing data, no duplicate data
- No outliers Detected
- Univariate and Multivariate analysis to identify the relationships
- Plotted the distribution of variables and relations
- Correlation analysis: Examine and discuss correlations between features and target variables



Exploratory Data Analysis

Conclusions From EDA

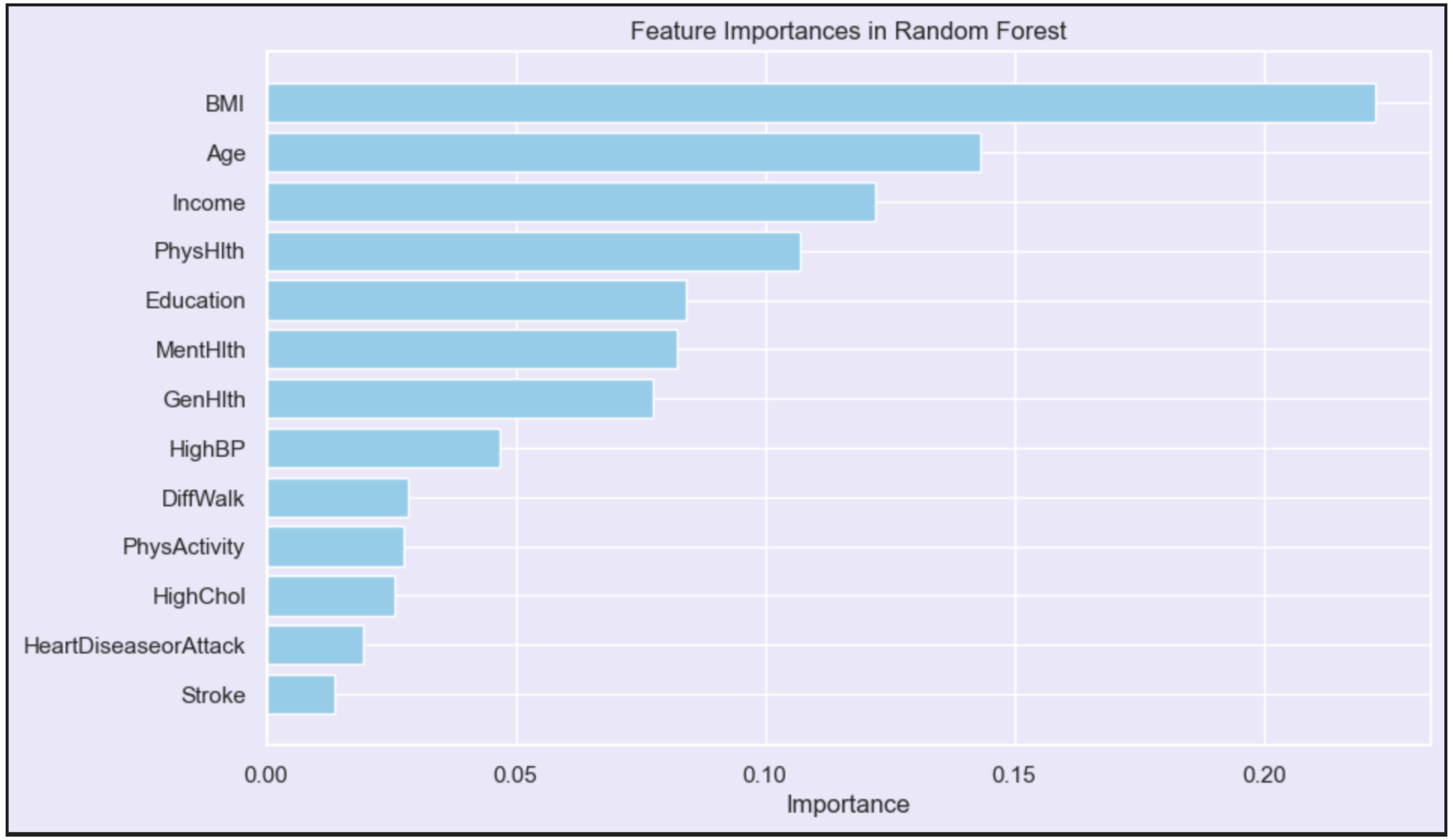
1. Target Feature of the Data is Diabetes_binary.
2. dataset has 15 Discrete type and 7 continuous type feature variables.
3. Dataset dose not have missing values(null values).
4. major feature variables for Diabetes are : HIghBP , HighChol , BMI, PhysicalActivity , GenHlth , MentHlth , PhysHlth , Age , Eduation and Income.
5. Feature variables which increases the risk of Diabetes togather are : Smoking and HvyAlcoholConsump , Stroke and HeartDiseaseorAttack , HighBP and HighChol.
6. Feature variable Which is least effective on Diabetes , but they can help in decreasing the risk Diabetes are : Fruits , Veggies , AnyHealthcare , CholChek.

Feature Engineering

Steps Taken:

- Correlation Analysis
- Feature Importance Analysis from Initial Iterations of the Models
- From the analysis, we have selected 13/21 Features for model training which seems to have the most impact
- Selected Features

'HighBP', 'HighChol', 'BMI', 'PhysActivity’,'GenHlth', 'MentHlth', 'PhysHlth', 'DiffWalk’, 'Age', 'Education' ,'Income','Stroke','HeartDiseaseorAttack'



Model Development

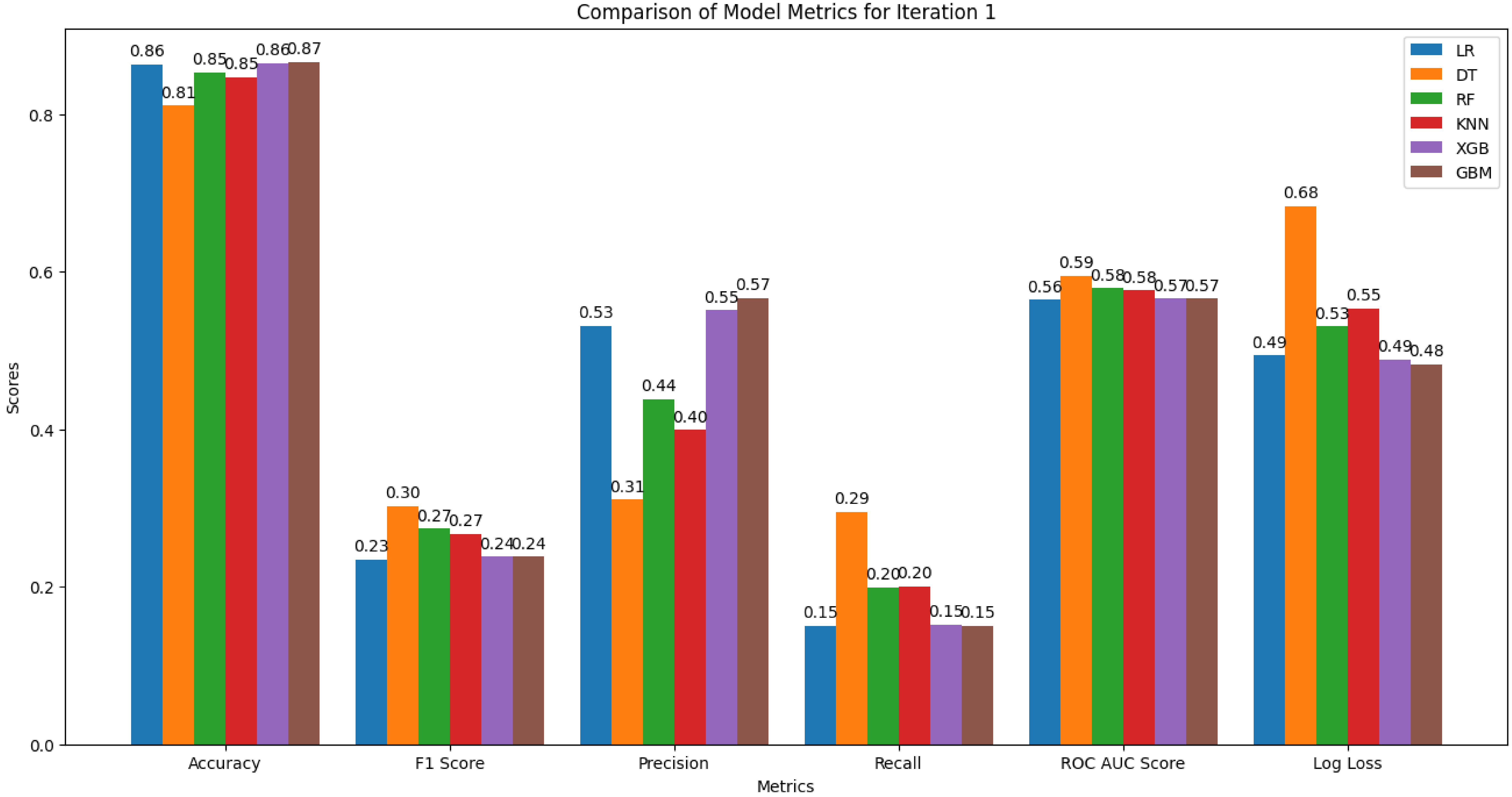
Considerations

- Target Variable: Diabetes Binary : 0 = no diabetes 1 = prediabetes or diabetes
- Modelling As : Classification Problem
- Machine Learning Algorithms Considered

1. Logistic Regression
2. Decision Tree
3. Gradient Boosting Machines (GBM)
4. Random Forest
5. K Nearest Neighbours (KNN)
6. XGBoost

Model Development

First Iteration: Without any Optimizations



Model Development

First Iteration: Without any Optimizations

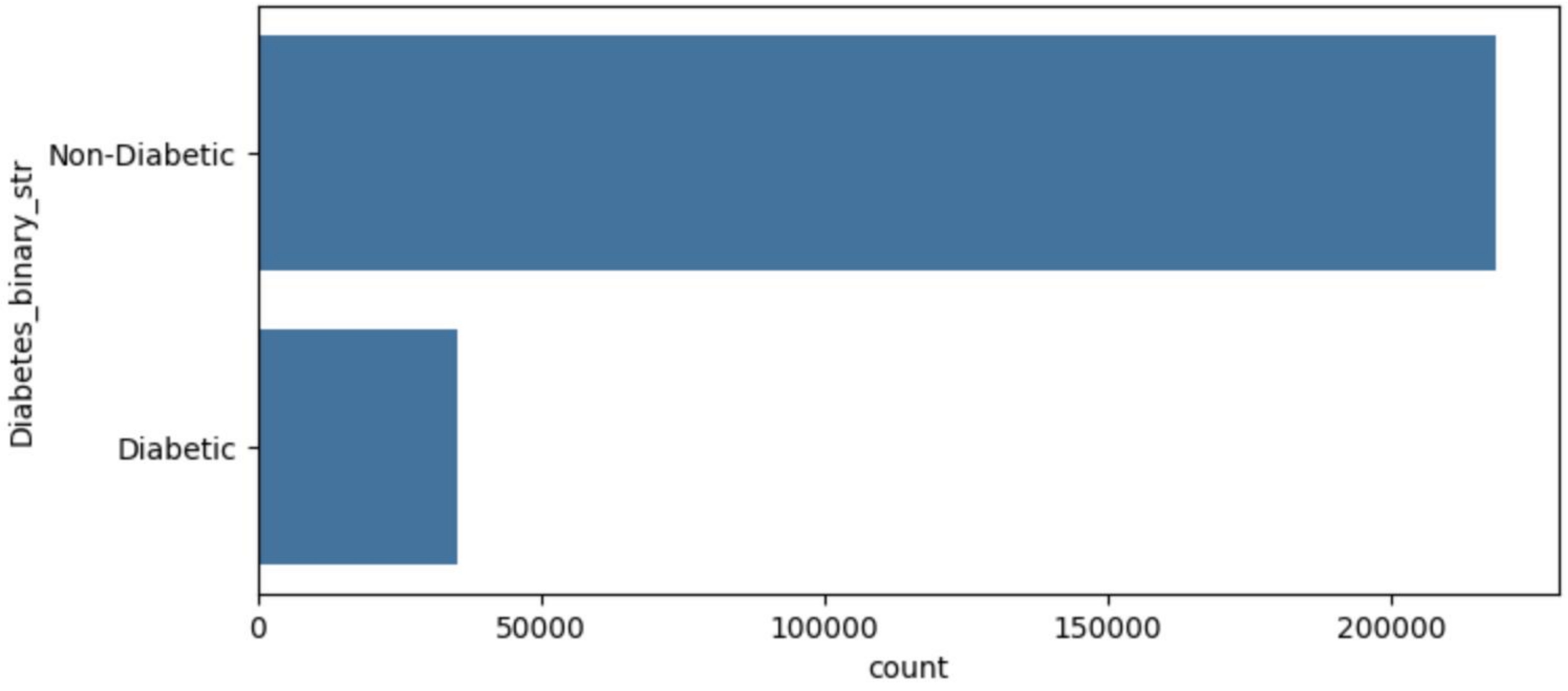
- Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB) have been selected for further iterations due to their strong performance metrics and suitability for healthcare prediction tasks.
- LR is valued for its balanced precision and recall, crucial for early detection, and well-calibrated probabilities.
- RF is chosen for its robustness and ability to model complex interactions, while XGB stands out for its high accuracy, efficiency, and scalability, making it ideal for real-time clinical applications.
- Decision Tree (DT) and K-Nearest Neighbors (KNN) were excluded due to lower performance metrics, and although GBM's results are comparable to XGB, XGB's faster training times and slightly better log loss make it preferable.

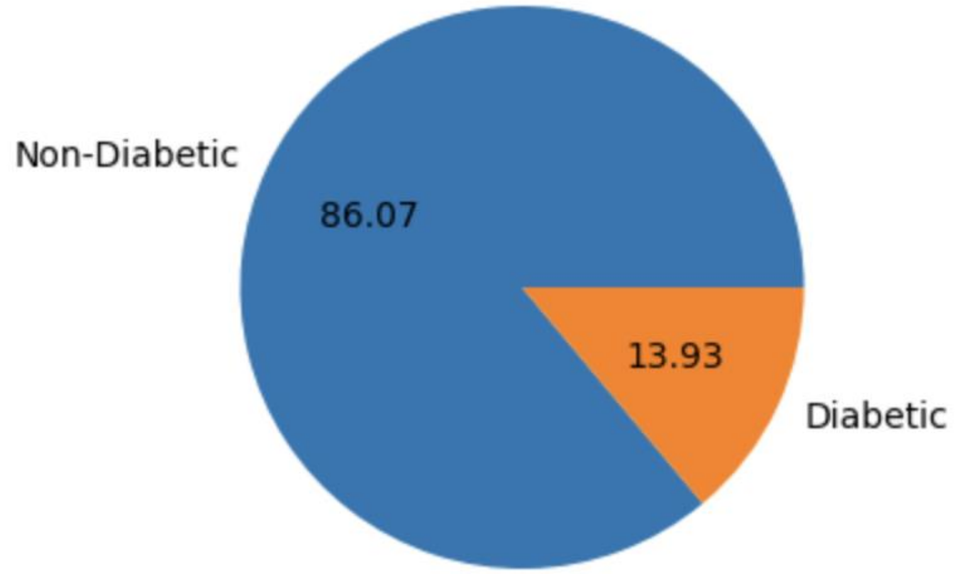
Selected Models for Further Iterations: LR, RF, XGB

Model Development

Considerations: Further Iterations

- **Major Issues Noticed After First Iterations**
  - **Class Imbalance: The disparity between high accuracy and low F1 score, recall, and RO-AUC suggests class imbalance.**
  - **The model is likely biased towards predicting the majority class.**
  - **Model Calibration and Discrimination:**
  - **The high Log Loss and low RO-AUC indicate that the model's probability estimates are poorly calibrated and it struggles to discriminate between classes.**





Model Development

Considerations: Further Iterations

- **Methods Followed**

  1. StratifiedKFold for cross-validation

  2. SMOTE (Synthetic Minority Over-sampling Technique)

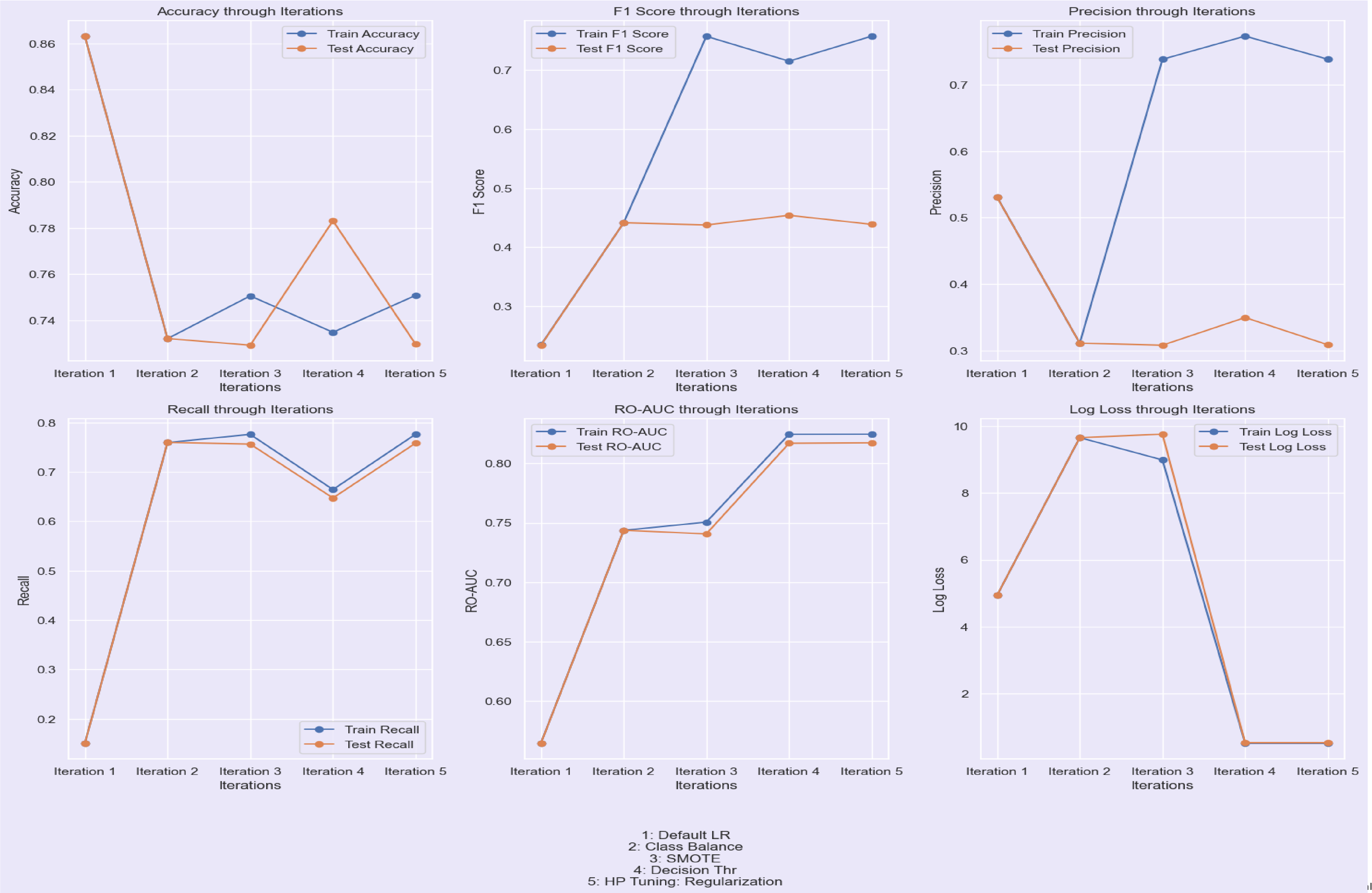
  3. Model specific tuning for class balance

  eg: LogisticRegression model with class_weight='balanced'

  4. HyperParameter Tuning With GridSearchCV

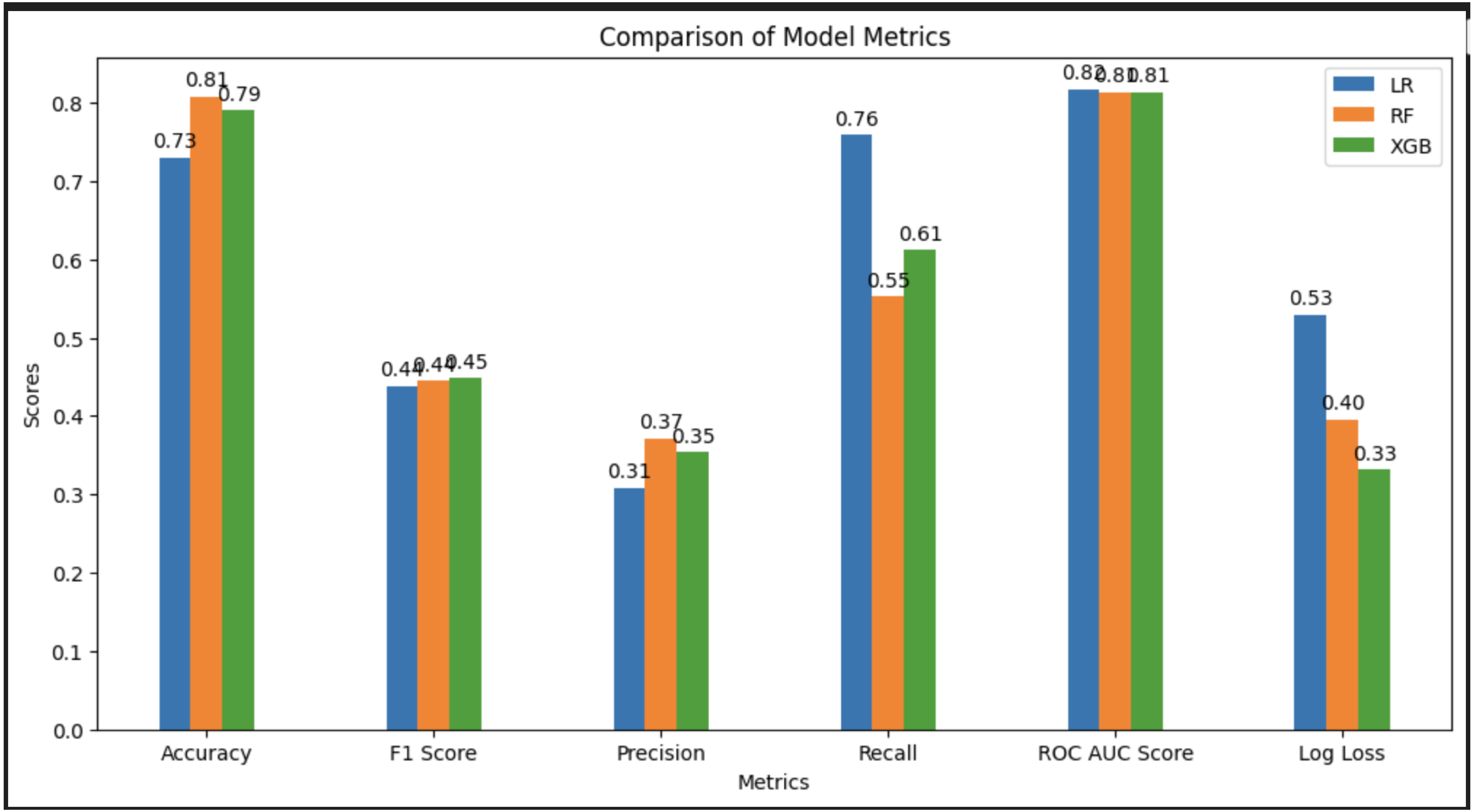
Model Development

LR Model Metrics through iterations



Model Development: Results

Model Metrics after final iteration



Conclusion

Best Model Selection

- **Objectively Best Model - XGB**
  - **Best model in terms of accuracy and probability calibration, making it a robust choice for general prediction tasks.**
- **Best Model For Diabetes Prediction – Logistic Regression**
  - **High Recall and ROC AUC Metric**
  - **Crucial in healthcare settings to ensure that all potential diabetes cases are flagged for further testing, minimising the risk of false negatives.**
  - **Ensure that fewer cases of diabetes are missed, which is vital for effective intervention and management.**

Learnings and Future Scope

- **Learnings**
  - Real life data is mostly imbalanced, cannot be solved by data collection methods
  - Evaluating the model across all metrics is very important
  - Select the metric which is closer to the problem domain needs

- **Future Scope**
  - Have more complete dataset which have more relevant features, like genetic disposition, hormonal imbalances etc that may affect Diabetes Prediction
  - Use Ensemble methods to achieve better performance by combining multiple algorithms and its strengths.
  - Use probability calibration techniques like Platt Scaling or isotonic regression.

Thank you