AL SHAMALI, UBAY (PGT)

University of Warwick

**FAIDM – Individual Assignment**

[CRISP-DM 3](#_Toc221521278)

[Business Context and Objectives 3](#_Toc221521279)

[Data Mining Goals and Success Criteria 4](#_Toc221521280)

[Constraints and Business Value 4](#_Toc221521281)

[Supervised Learning 4](#_Toc221521282)

[Problem Definition and Rationale 4](#_Toc221521283)

[Problem Definition: 4](#_Toc221521284)

[Rationale for Supervised Learning: 5](#_Toc221521285)

[Practical Relevance: 5](#_Toc221521286)

[Analytical Value: 5](#_Toc221521287)

[Methodology Overview 5](#_Toc221521288)

[Pipeline Design: 5](#_Toc221521289)

[Dataset Overview 5](#_Toc221521290)

[Dataset Source: 5](#_Toc221521291)

[Dataset Size and Scope: 5](#_Toc221521292)

[Features and Variables: 5](#_Toc221521293)

[Target Variable and Characteristics: 6](#_Toc221521294)

[Pipeline 6](#_Toc221521295)

[Exploratory Data Analysis (EDA) 6](#_Toc221521296)

[Data Preprocessing 11](#_Toc221521297)

[Supervised Learning Model Development 13](#_Toc221521298)

[RFE 16](#_Toc221521299)

[Handling Class Imbalance 17](#_Toc221521300)

[18](#_Toc221521301)

[18](#_Toc221521302)

[Hyperparameter Tuning 18](#_Toc221521303)

[Threshold Optimization 21](#_Toc221521304)

[Calibration 21](#_Toc221521305)

[Analysis, Results and Reflection 23](#_Toc221521306)

[Results comparison 23](#_Toc221521307)

[Explainability 25](#_Toc221521308)

[Unsupervised Learning 28](#_Toc221521309)

[Problem Definition and Rationale 28](#_Toc221521310)

[Problem Definition: 28](#_Toc221521311)

[Rationale for Unsupervised Learning: 28](#_Toc221521312)

[Practical Relevance: 28](#_Toc221521313)

[Analytical Value: 28](#_Toc221521314)

[Unsupervised Learning Model Development 28](#_Toc221521315)

[K-Means 28](#_Toc221521316)

[DBSCAN 31](#_Toc221521317)

[Association Rules Mining 36](#_Toc221521318)

[Problem Definition and Rationale 36](#_Toc221521319)

[Problem Definition: 36](#_Toc221521320)

[Rationale for ARM: 36](#_Toc221521321)

[Practical Relevance: 36](#_Toc221521322)

[Analytical Value: 36](#_Toc221521323)

[Model Development 36](#_Toc221521324)

[Feature engineering for ARM 36](#_Toc221521325)

[Transaction Construction 37](#_Toc221521326)

[Training Apriori 37](#_Toc221521327)

[Rule Evaluation and Interpretation 38](#_Toc221521328)

[Deployment and Business Alignment 40](#_Toc221521329)

[ Integrated Analytical Deployment: 40](#_Toc221521330)

[ Supervised Model Deployment: 40](#_Toc221521331)

[ Unsupervised Model Deployment: 40](#_Toc221521332)

[ Knowledge Discovery and Policy Support: 40](#_Toc221521333)

[ Operational Integration and Monitoring: 40](#_Toc221521334)

[ Ethical, Fairness, and Sustainability Considerations: 40](#_Toc221521335)

[References 41](#_Toc221521336)

[Appendix 43](#_Toc221521337)

[Supervised learning 43](#_Toc221521338)

[Unsupervised learning 55](#_Toc221521339)

[DBSCAN 59](#_Toc221521340)

[Association rules mining 62](#_Toc221521341)

CRISP-DM

# Business Context and Objectives

Diabetes is a major public health challenge, associated with increased morbidity, mortality, and long-term healthcare costs. Early identification of individuals at risk and improved understanding of lifestyle and health-related factors are essential for effective prevention and intervention strategies.

In this project, a public health research institute aims to utilise data-driven methods to analyse diabetes risk across the adult population using the CDC Diabetes HealthIndicators Dataset, which contains over 250,000 survey responses. The dataset includes demographic information, lifestyle behaviours, and health indicators, providing a suitable basis for applying artificial intelligence and data mining techniques.

The primary business objectives are threefold. First, the institute seeks to develop a predictive classificationmodel capable of accurately distinguishing between diabetic and non-diabetic individuals, with an additional interest in probabilistic risk estimation to support preventive decision-making. Second, the institute aims to apply clustering techniques to identify meaningful population segments with similar health behaviours and risk profiles, enabling the design of targeted public health interventions. Third, association rule mining is used to uncover interpretable relationships between lifestyle factors, health conditions, and diabetes diagnosis, supporting deeper insight generation.

# Data Mining Goals and Success Criteria

From a data mining perspective, the project aims to implement and evaluate supervised classification models using appropriate performance metrics, apply clustering algorithms and assess their quality using internal validation measures, and extract meaningful association rules. Success is defined by both predictive performance and interpretability, ensuring that results are actionable and relevant to real-world public health contexts.

# Constraints and Business Value

Key constraints include the self-reported nature of the data, potential class imbalance, and ethical considerations related to fairness and responsible use of health predictions. Despite these challenges, the project is expected to provide valuable insights that support evidence-based policy development, targeted health campaigns, and improved understanding of diabetes risk factors through responsible application of AI techniques.

Supervised Learning

# Problem Definition and Rationale

## Problem Definition:

This study addresses a supervised binary classification problem aimed at predicting diabetes diagnosis (diabetic vs. non-diabetic) using demographic, lifestyle, and health-related indicators from the CDC Diabetes Health Indicators Dataset.

## Rationale for Supervised Learning:

Supervised learning is appropriate due to the presence of a clearly labelled target variable, allowing models to learn discriminative patterns from historical health data.

## Practical Relevance:

The developed predictive model supports early identification of individuals at elevated diabetes risk, enabling probability-based predictions that can inform preventative public health strategies.

## Analytical Value:

Supervised methods enable systematic evaluation using standard performance metrics, ensuring reliable, interpretable, and data-driven healthcare insights.

# Methodology Overview

## Pipeline Design:

A structured machine learning pipeline was implemented to ensure data quality, reproducibility, and robust model evaluation aligned with public health objectives.

# Dataset Overview

## Dataset Source:

The CDC Diabetes Health Indicators Dataset, sourced from the UCI Machine Learning Repository and derived from a national health survey conducted by the CDC.

## Dataset Size and Scope:

Includes over 250,000 adult records, providing broad population coverage and strong statistical reliability.

## Features and Variables:

Covers demographic attributes, lifestyle behaviours, and clinical health indicators such as BMI, blood pressure, general and mental health status.

## Target Variable and Characteristics:

The binary diabetes diagnosis variable is imbalanced, requiring careful evaluation and model design.

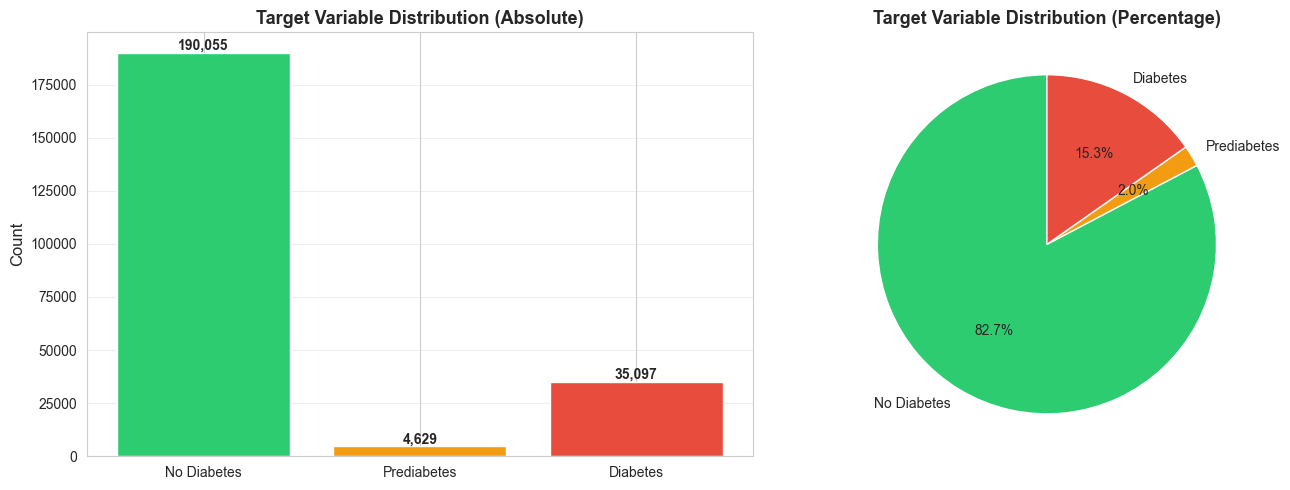
Features are primarily numerical, enabling compatibility with a wide range of supervised learning algorithms.

# Pipeline

## Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase was conducted to understand the structure, distribution, and relationships within the CDC Diabetes Health Indicators Dataset, and to inform preprocessing and modelling decisions.

* **Target Variable Exploration**



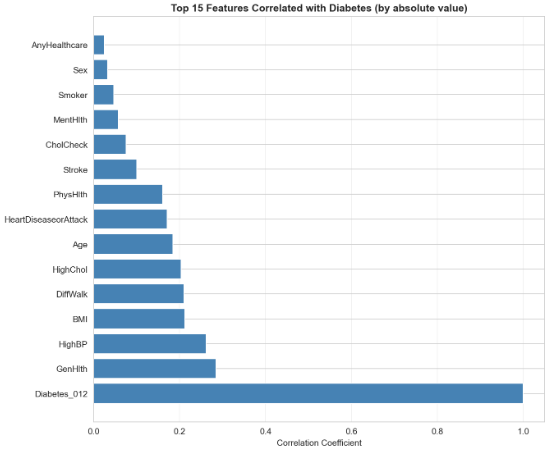
* **Feature Distribution Analysis**

Feature distribution analysis was performed to assess the statistical characteristics of input variables and identify factors that could affect model performance. The dataset includes binary, ordinal, and continuous features representing demographic, lifestyle, and health-related indicators. Binary variables, such as smoking status and physical activity, exhibited highly skewed distributions, reflecting real-world population patterns but posing a risk of biased learning if considered in isolation. Continuous variables, including Body Mass Index (BMI) and age-related measures, showed greater variability, non-normal distributions, and right skewness, with the presence of extreme values. These distributional properties justified the use of feature scaling and supported the selection of both linear and non-linear models. Overall, this analysis ensured that preprocessing and modelling decisions were informed by the underlying data characteristics.

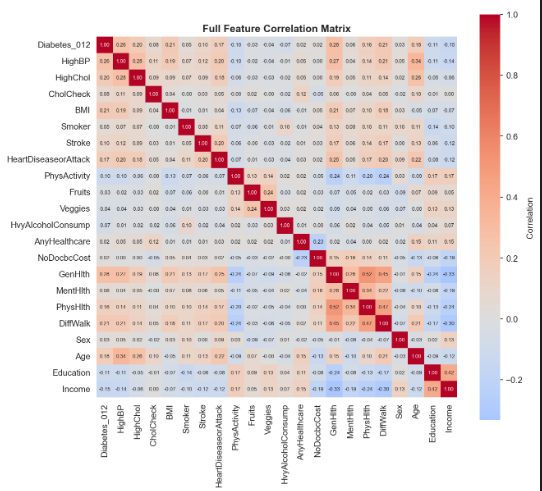
Figure 14-Feature Distribution Analysis

* **Relationship Between Features and Diabetes**

Analysis of feature diabetes relationships showed clear differences between diabetic and non-diabetic individuals. Health indicators such as BMI, general health status, high blood pressure, and cardiovascular disease were more prevalent among those diagnosed with diabetes, with prevalence also increasing with age. Lifestyle factors, including physical activity and diet, exhibited weaker but noticeable associations, reflecting the multifactorial nature of the condition. No single feature was sufficient to explain diabetes diagnosis, supporting the use of multivariate supervised learning models to capture interacting risk factors.



* **Correlation Analysis**

Correlation analysis examined relationships among numerical features, revealing moderate correlations among health-related variables, especially chronic condition indicators, while lifestyle and demographic features were weakly correlated. No extremely high correlations were found, indicating low multicollinearity risk.

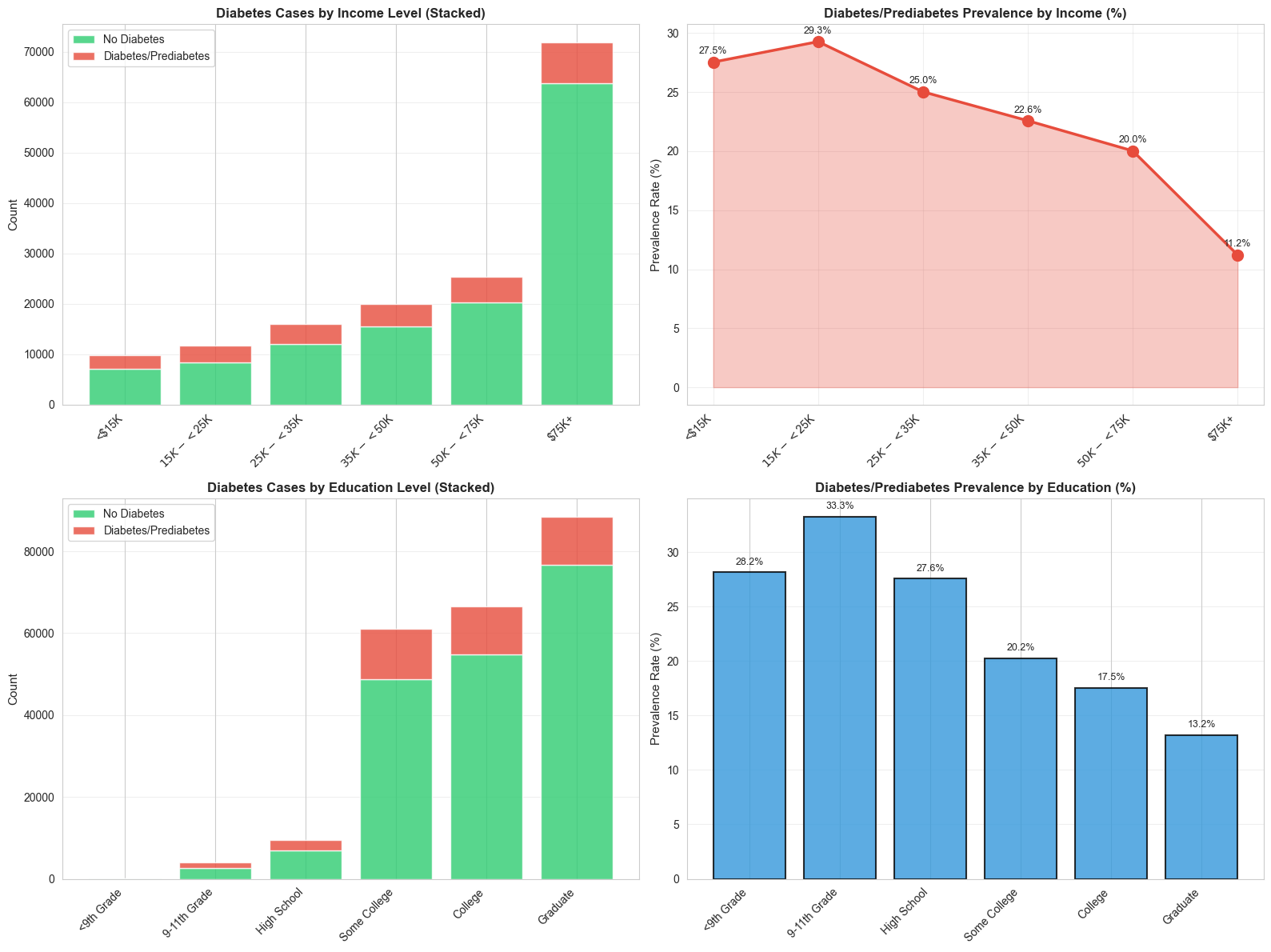
* **Diabetes Cases by Sex**

The dataset contains a higher proportion of male participants overall; however, diabetes prevalence is comparable across both sexes. This indicates that despite the gender imbalance in representation, diabetes affects males and females at similar rates.

****

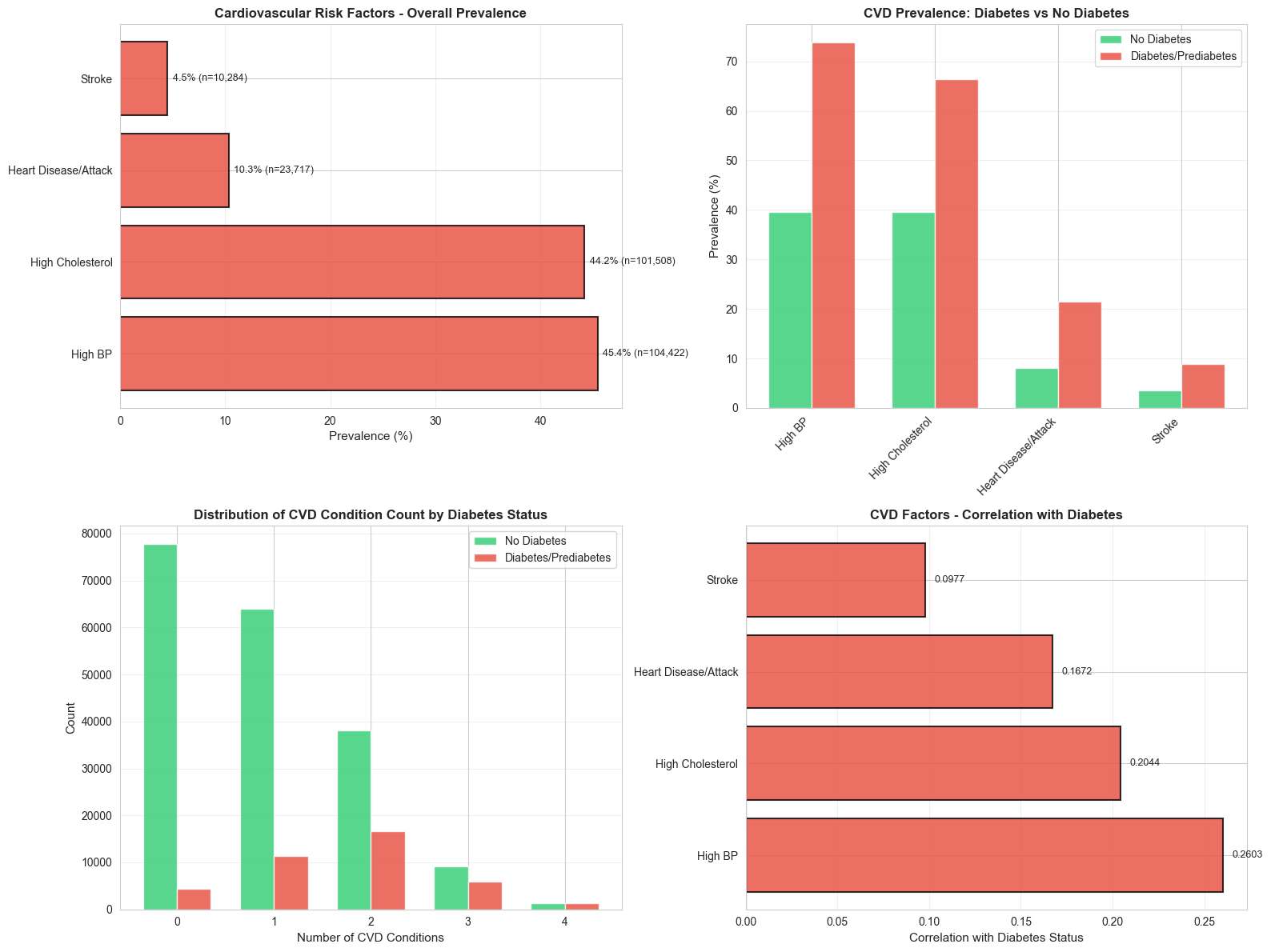
* **Diabetes Prevalence by Income and Education**

Individuals with lower income levels and lower educational attainment exhibit higher diabetes prevalence, with rates reaching approximately 25–30% among those earning less than $15,000 annually or possessing minimal formal education. In contrast, higher income and education levels are associated with reduced diabetes risk. These findings highlight the significant influence of socioeconomic factors on diabetes prevalence, underscoring the need for targeted public health interventions focused on vulnerable populations.

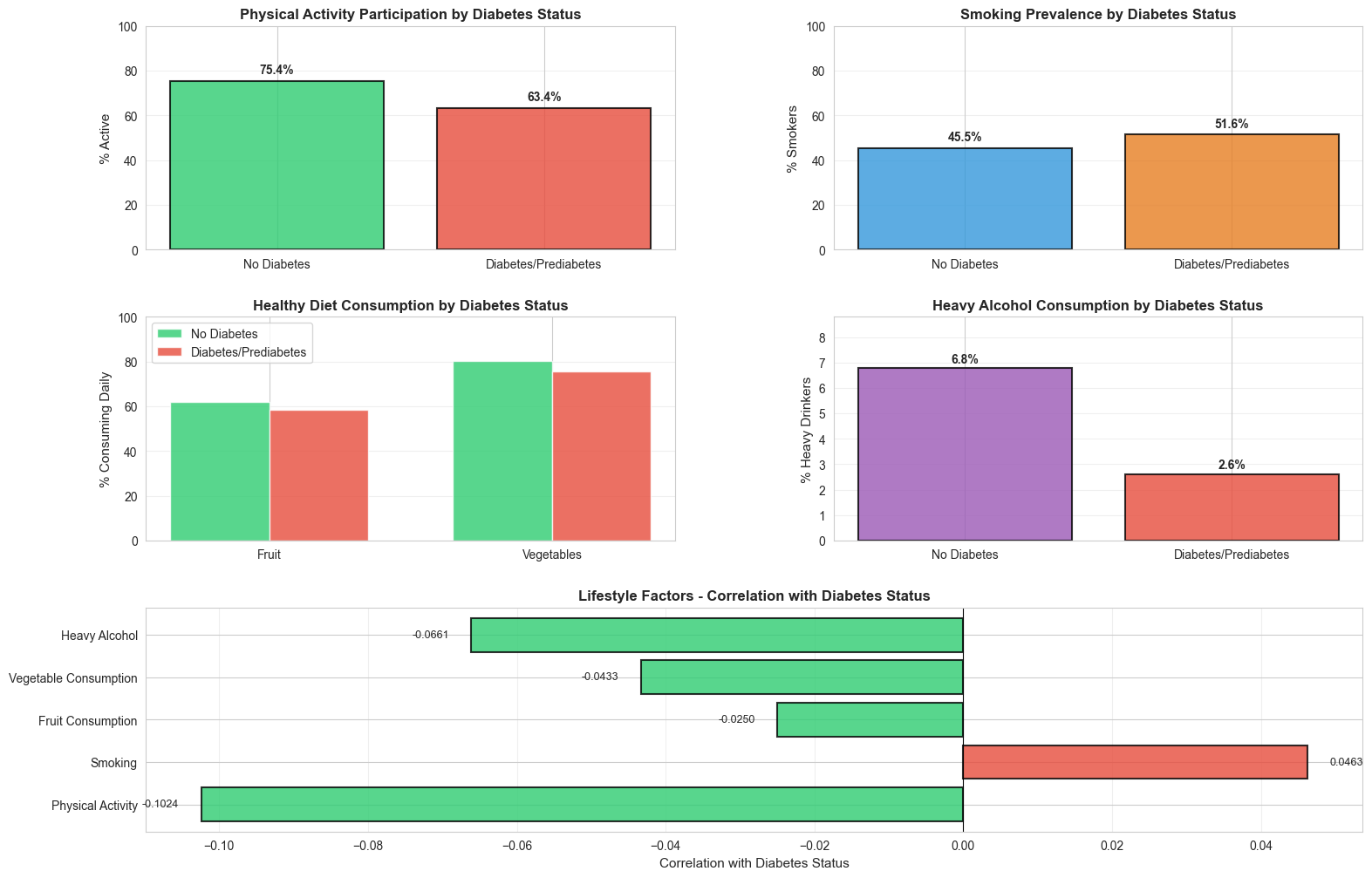


* **Cardiovascular Health Conditions**

High blood pressure (45.4%) and high cholesterol (44.2%) are the most common cardiovascular risk factors. Individuals with diabetes or prediabetes show higher prevalence across all conditions, including heart disease and stroke, are more likely to have multiple conditions, and exhibit strong positive correlations between diabetes and cardiovascular risk.



* **Lifestyle Factors and Diabetes**

****

* **Healthcare Access**

****

## Data Preprocessing

Data preprocessing was conducted to ensure data quality, consistency, and suitability for supervised machine learning models. The following steps were systematically applied:

1. **Data Inspection and Understanding**

* Loaded the dataset and examined its overall structure, including:
  + Number of samples and features
  + Data types of each feature
  + Distribution of the target variable
* Reviewed feature descriptions to understand their relevance to diabetes prediction.
* Identified potential issues such as inconsistent values or abnormal distributions.

1. **Data Cleaning**

* Checked for missing or null values across all features.
  + Confirmed that no critical features contained missing data requiring imputation.
* Removed duplicate records to prevent bias in model training.
* Verified that all feature values were within reasonable and medically plausible ranges.

1. **Target Variable Encoding**

Ensured the target variable was properly encoded for binary classification:

**0** → non-diabetic

**1** → Diabetic / Pre-diabetic

The pre-diabetic class was highly underrepresented, providing insufficient data for reliable pattern learning.

Treating it as a separate class increased class imbalance and introduced noise during training.

Merging diabetes and pre-diabetes increased the effective sample size of positive cases.

This resulted in more stable learning, improved recall for at-risk individuals, and better overall model generalization, as we are going to see the results of the three-category target feature later.

Clinically, both classes represent elevated risk, making the merged class appropriate for early detection objectives and classification.

Confirmed correct class labelling to avoid data leakage or misclassification.

1. **Feature Engineering**

Feature engineering was conducted to enhance model performance by expanding and refining the feature space. Additional features were intentionally created to capture potential patterns and interactions that may contribute meaningful predictive information. These engineered features were included on an exploratory basis, with the understanding that many could later be removed through correlation analysis, feature importance evaluation, or recursive feature elimination. Numerical features were standardized to ensure balanced learning across models, while redundant and weakly informative features were discarded to reduce noise and prevent overfitting. This exploratory approach allowed meaningful features to be retained while systematically eliminating those that did not improve model performance.

Exploratory features been added for potential enhancements as been shown in the code:

Figure 15 - new features

1. **Feature Selection**

* Reviewed correlations between features to identify redundancy.
* Removed features with negligible variance or minimal predictive contribution.
* Prepared the dataset for advanced feature selection techniques (e.g., RFE).

1. **Train–Test Split**

Split the dataset into training and testing subsets using stratified sampling, preserved the original class distribution of the target variable, and ensured fair and unbiased evaluation of model performance.

1. **Feature Scaling and transforming**

Applied standardization (StandardScaler) to numerical features:

Mean = 0

Standard deviation = 1

Ensured scaling was fit only on the training set to prevent data leakage, and this have been applied on the

Here is the code for data splitting with feature scaling and transforming:

Figure 16 - Data splitting transforming & scaling

## Supervised Learning Model Development

**Base line models**

1. Logistic regression

* Implementation: LogisticRegression(max iter=1000, random state=42).
* Training uses the log-transformed, standardized feature set.
* Rationale: strong baseline, fast to train, coefficients support interpretability.

1. Random forest

* Implementation: RandomForestClassifier(n estimators=200, max depth=15, random state=42, n jobs=-1).
* Training uses the scaled feature set.
* Rationale: Captures non-linear relationships and feature interactions, handles tabular data well, and provides feature importance while reducing overfitting compared to single decision trees.

1. XGboost

* Implementation: XGBClassifier(n estimators=200, max depth=6, learning rate=0.1, random state=42n jobs=-1, scale pos weight=2, eval metric=’mlogloss’).
* Training uses the scaled feature set.
* Rationale: strong performance on tabular data and supports explicit positive-class weighting.

Figure 17 - Baseline models code

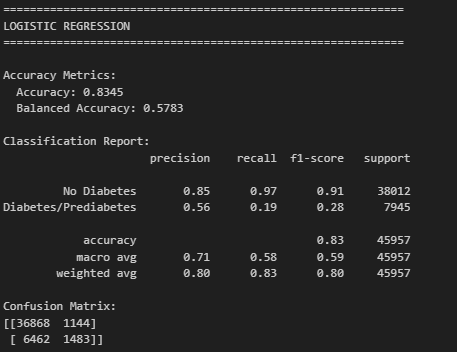
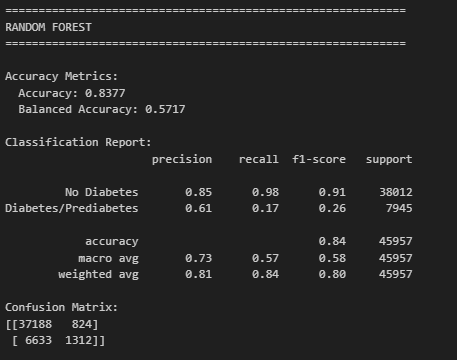
**Baseline models results & analysis**

Figure 1 - Random Forest Original Data

Figure 2 - Logistic Regression Original Data

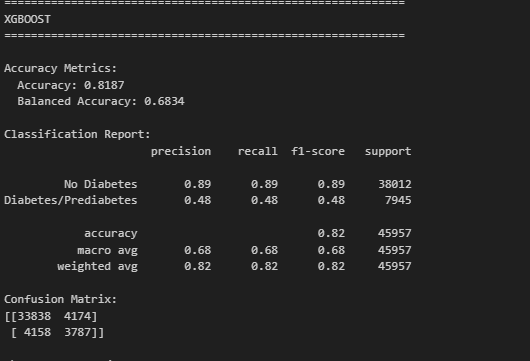
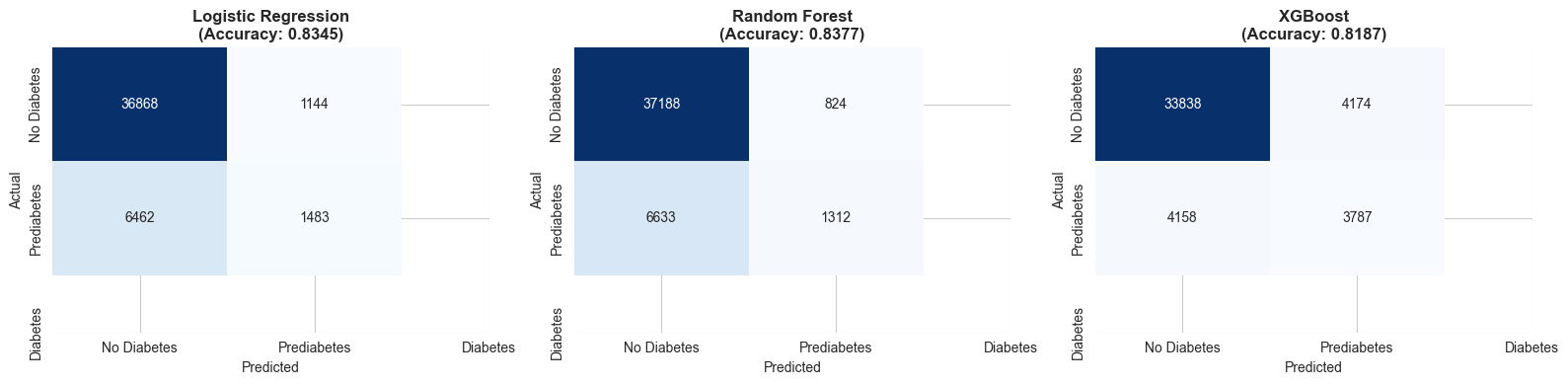
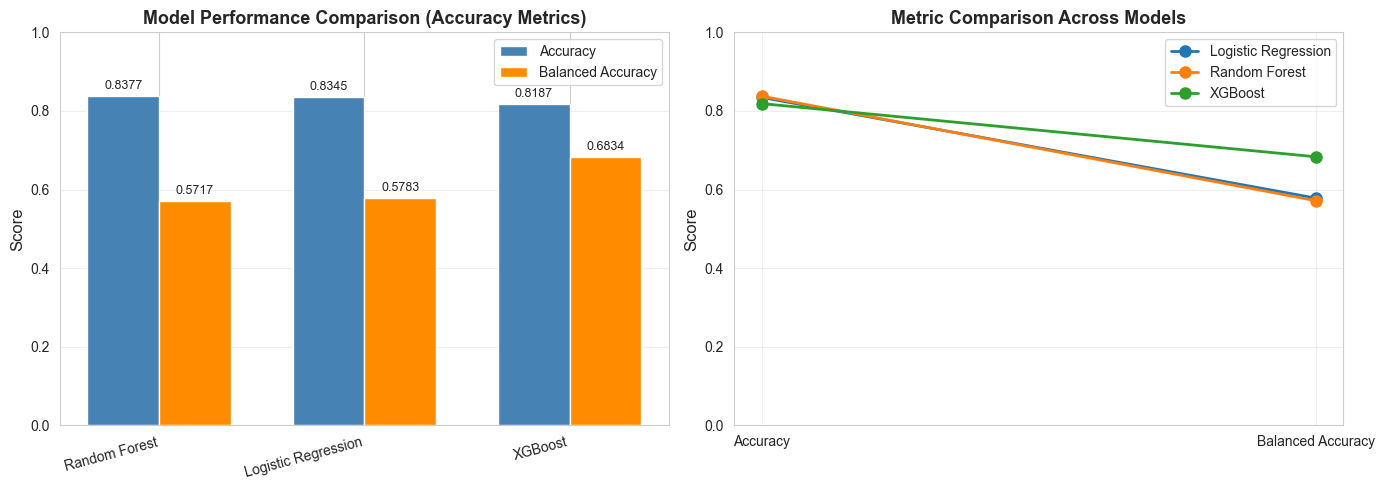
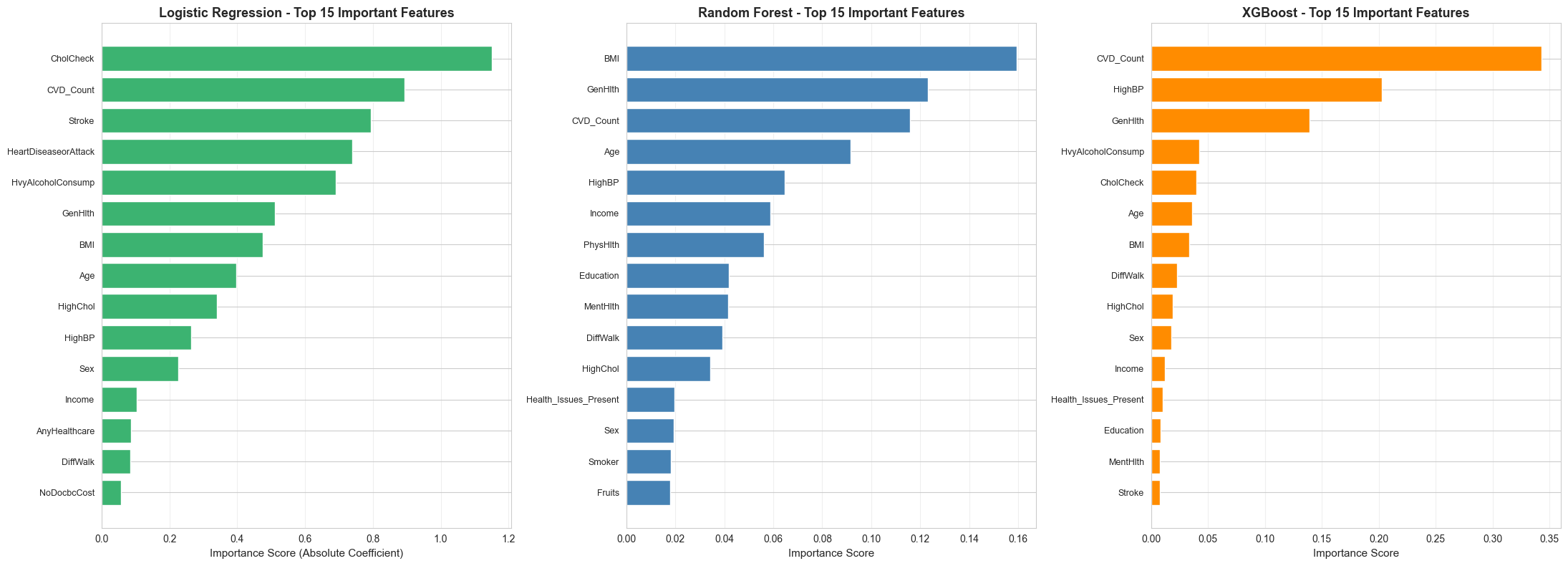
****

Figure 3 - XGBoost Original Data

**Confusion matrix**

**Comparison**

****

**Feature importance**

## RFE

**Recursive Feature Elimination (RFE)** was applied to identify the most informative subset of features and reduce model complexity. A **Random Forest classifier** was used as the base estimator due to its ability to capture non-linear relationships and feature interactions, robustness to noise, and inherent feature importance estimation. RFE iteratively removed the least important features until the top 15 features were retained. The selected features were then used to retrain the Random Forest model and evaluate performance relative to the full feature set.

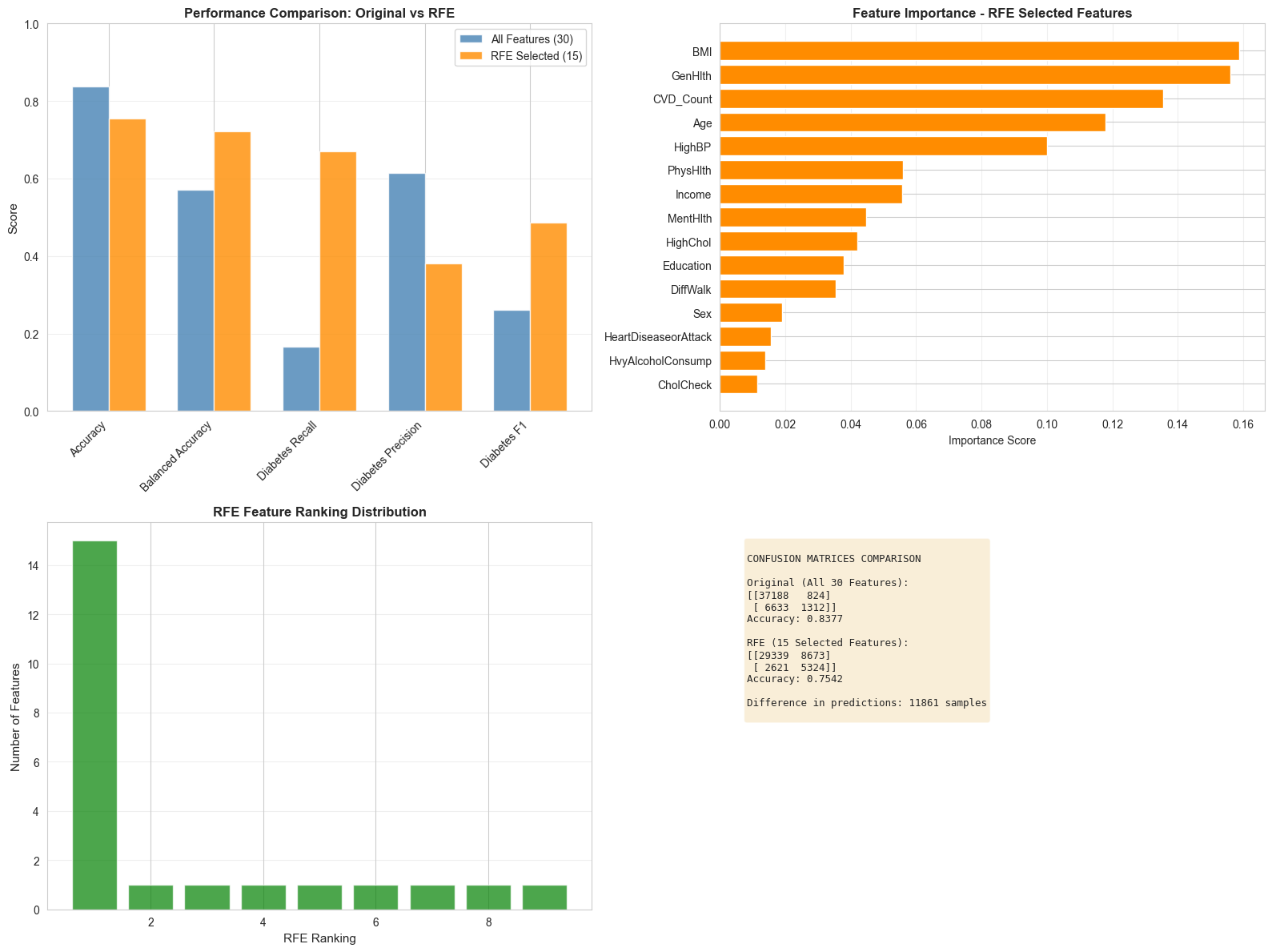
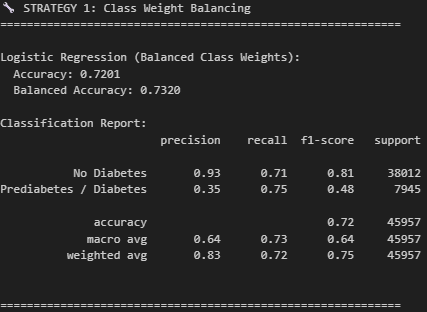


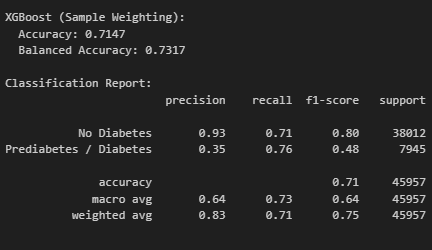
Figure 4 - RFE

## Handling Class Imbalance

* **Class weight balancing**

To address class imbalance, class weights were incorporated directly into the learning process for Logistic Regression and Random Forest using the class\_weight='balanced' option, while XGBoost was trained using sample level weighting derived from inverse class frequencies. This approach improved minority class sensitivity, as reflected in higher balanced accuracy and recall for the diabetes class.



* **SMOTE (Synthetic Minority Oversampling)**

SMOTE was applied to the training data to synthetically oversample the minority class and create a more balanced class distribution before model training. Models trained on the SMOTE augmented dataset showed improved minority class recall, with some trade-offs in overall accuracy, highlighting the effectiveness of data level imbalance correction.

## 

## 

Class weight balancing proved to be the more effective and stable approach in this study. It improved minority-class recall and balanced accuracy without altering the original data distribution, whereas SMOTE introduced synthetic samples that occasionally increased variance and slightly reduced generalization performance. As a result, class weight balancing was preferred for the final models.

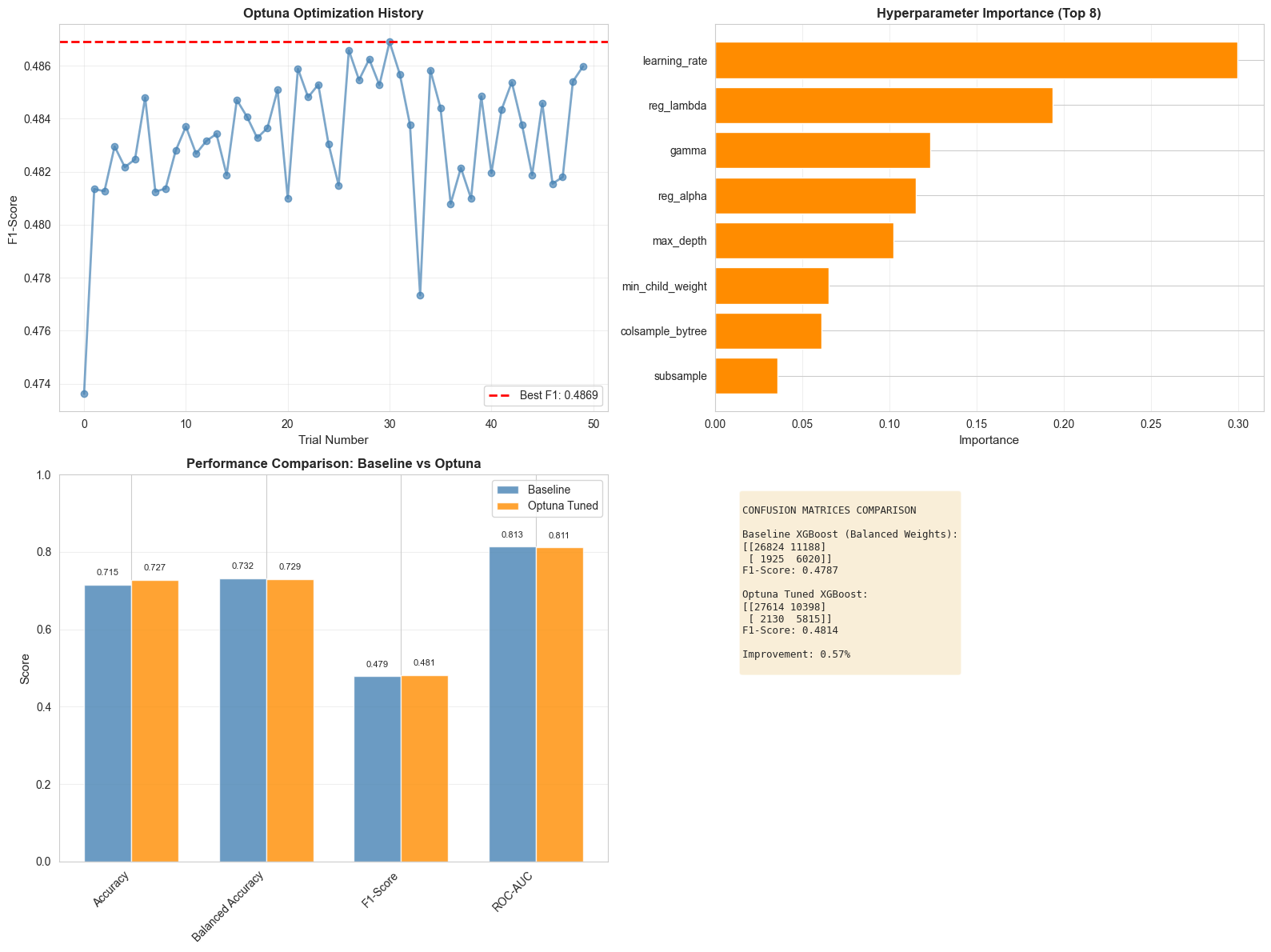
## Hyperparameter Tuning

Hyperparameter tuning was applied to Random Forest, XGBoost, and Logistic Regression models to improve performance under class imbalance, using F1-score as the primary optimization metric. Random Forest was tuned using GridSearchCV with 3-fold cross-validation, achieving a cross-validated F1-score of 0.4787 and a balanced accuracy of 0.7295, with stable test performance and improved recall for the diabetes class.  
XGBoost was optimized using Optuna with Bayesian search and balanced sample weighting, resulting in consistent improvements over the baseline in F1-score, balanced accuracy, and ROC-AUC.  
Logistic Regression was also optimized using Optuna, primarily tuning the regularization strength and penalty term. The tuned model achieved an F1-score of 0.4812, balanced accuracy of 0.7321, and ROC-AUC of 0.8073, yielding modest but consistent improvements over the baseline while maintaining strong interpretability.

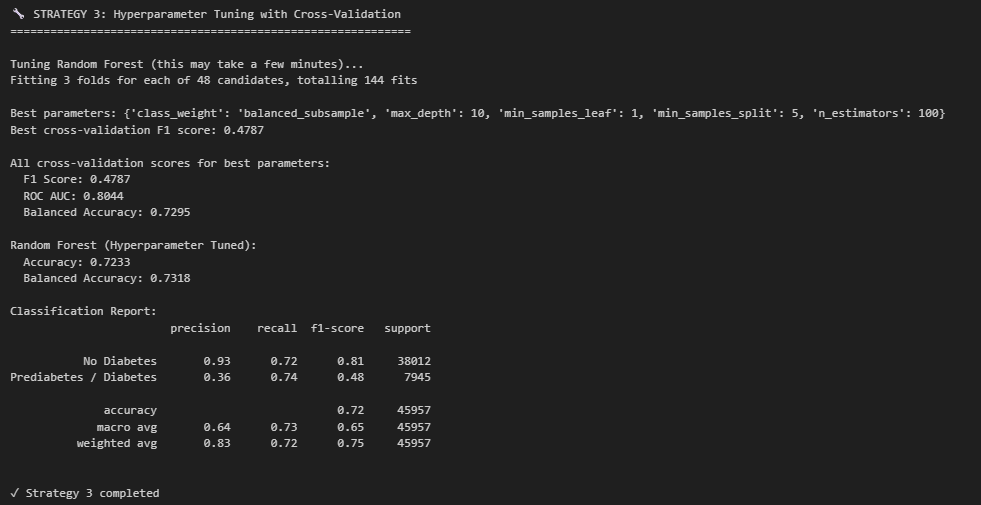
**Logistic regression Optuna**

****

**XGBoost Optuna**

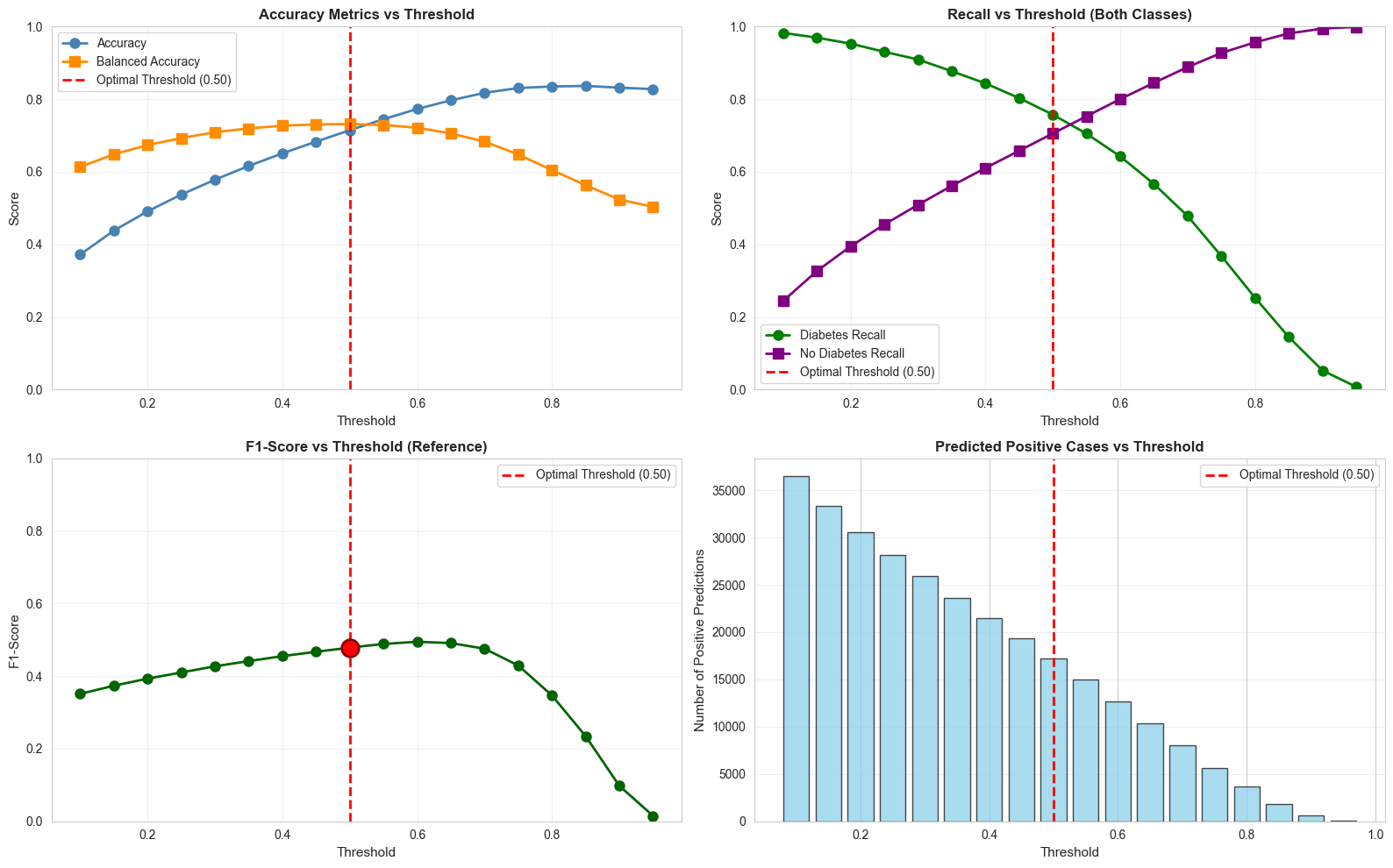
****

**Random forest grid search CV**

****

## Threshold Optimization

Threshold optimization was applied to the balanced XGBoost model to better control the trade-off between recall and precision. Instead of using the default 0.5 threshold, multiple thresholds were evaluated using predicted probabilities. The optimal threshold was selected by maximizing the minimum recall across both classes, ensuring balanced sensitivity while prioritizing the detection of diabetes cases. This approach aligns with the clinical objective of reducing false negatives, resulting in improved recall performance with acceptable impacts on other metrics.



## Calibration

Calibration was applied to the balanced XGBoost model to improve the reliability of predicted probabilities. Calibration quality was assessed using the **Brier score**, with lower values indicating better probability alignment.

Platt scaling and isotonic regression were evaluated using 5-fold cross-validation. Calibration curves were generated using quantile-based binning to ensure stable estimates under class imbalance. Both methods improved calibration compared to the original model, with the best approach selected based on the lowest Brier score. The calibrated model provides more trustworthy probabilities for risk-based decision-making.

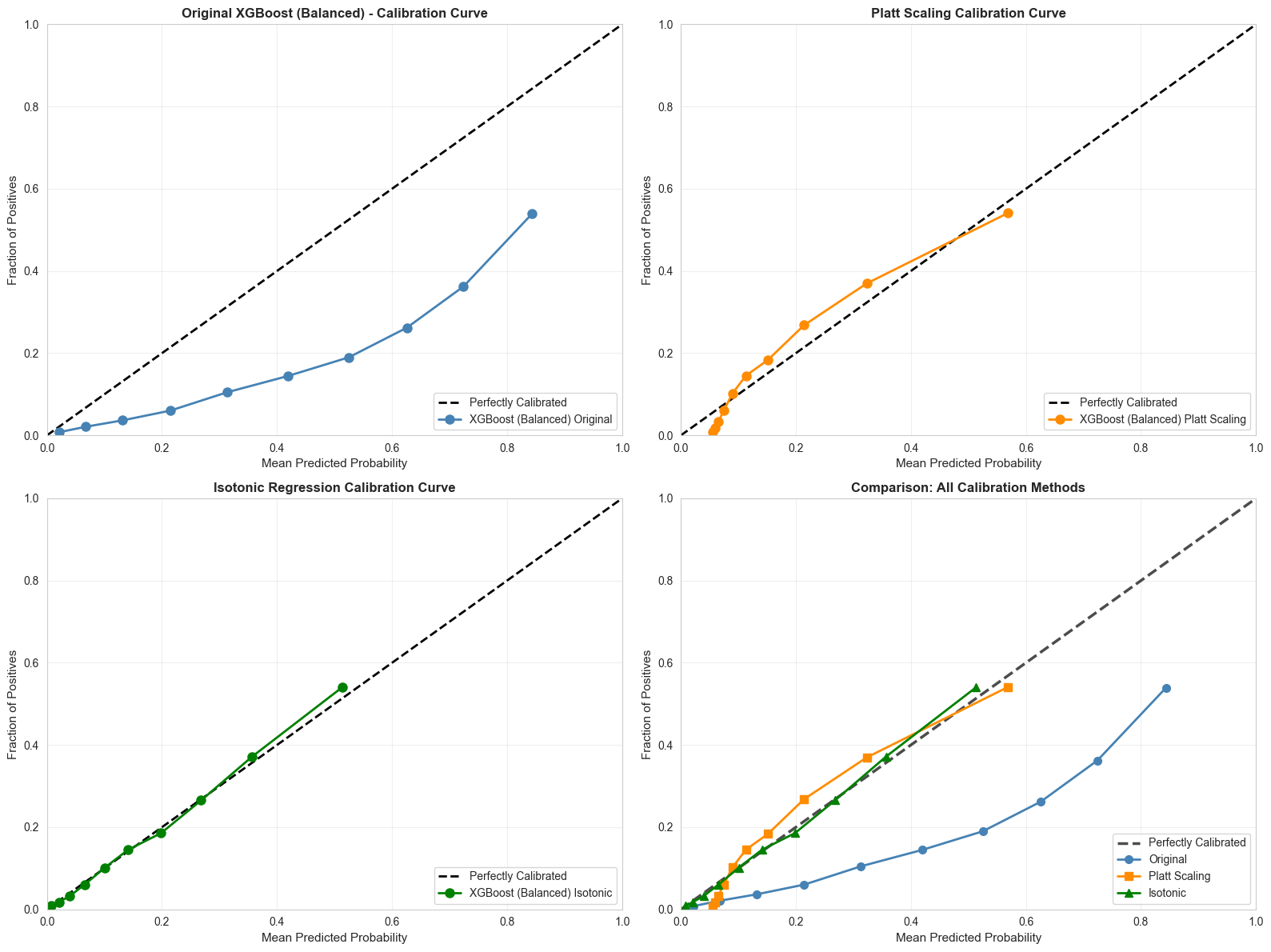
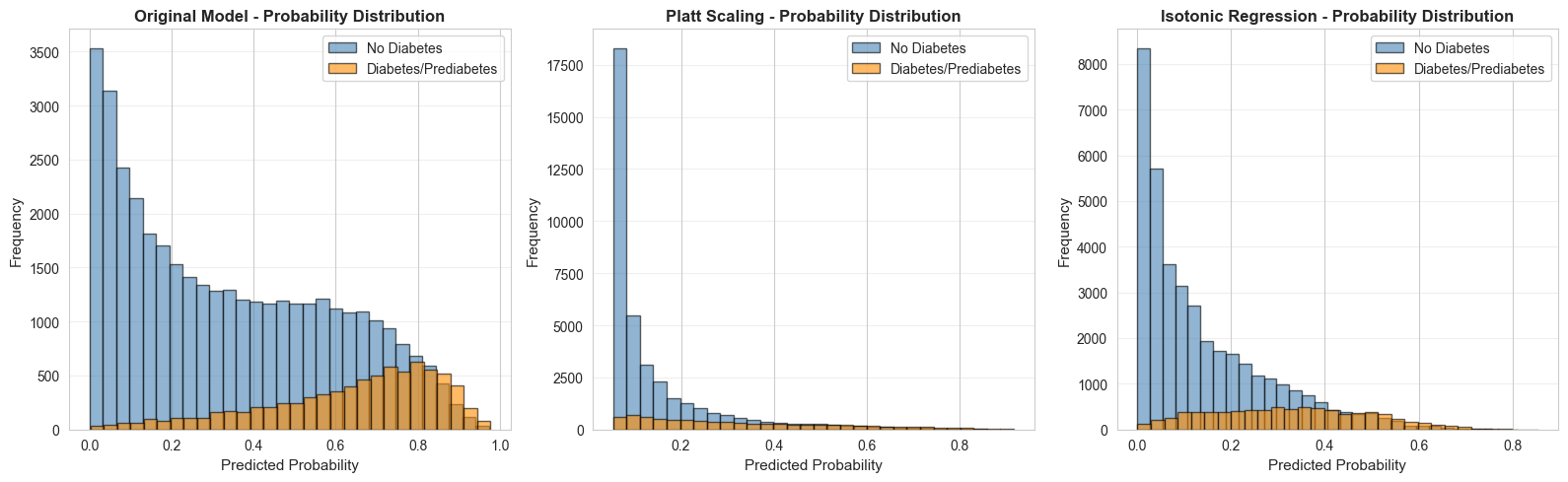


Figure 5 - XGBoost Calibration

**Comparing to the original model**

****

# Analysis, Results and Reflection

## Results comparison

**Report comparison among all implemented models**

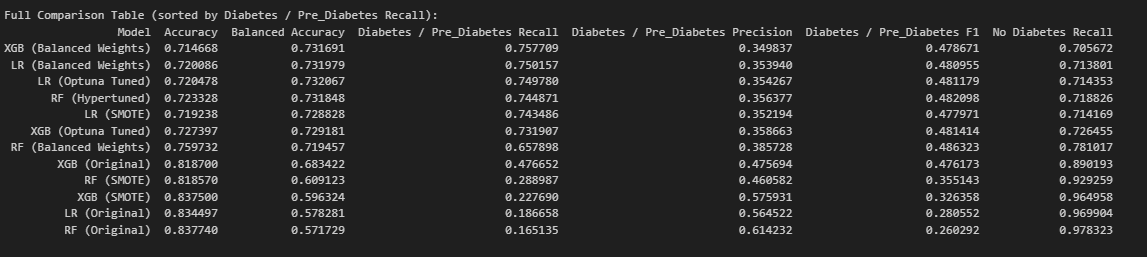


Figure 6 - Report comparison among all implemented models

**Top 3 models cross validation comparison**



Figure 7 - Top 3 models cross validation comparison

**Top 3 models ROC-AUC comparison**

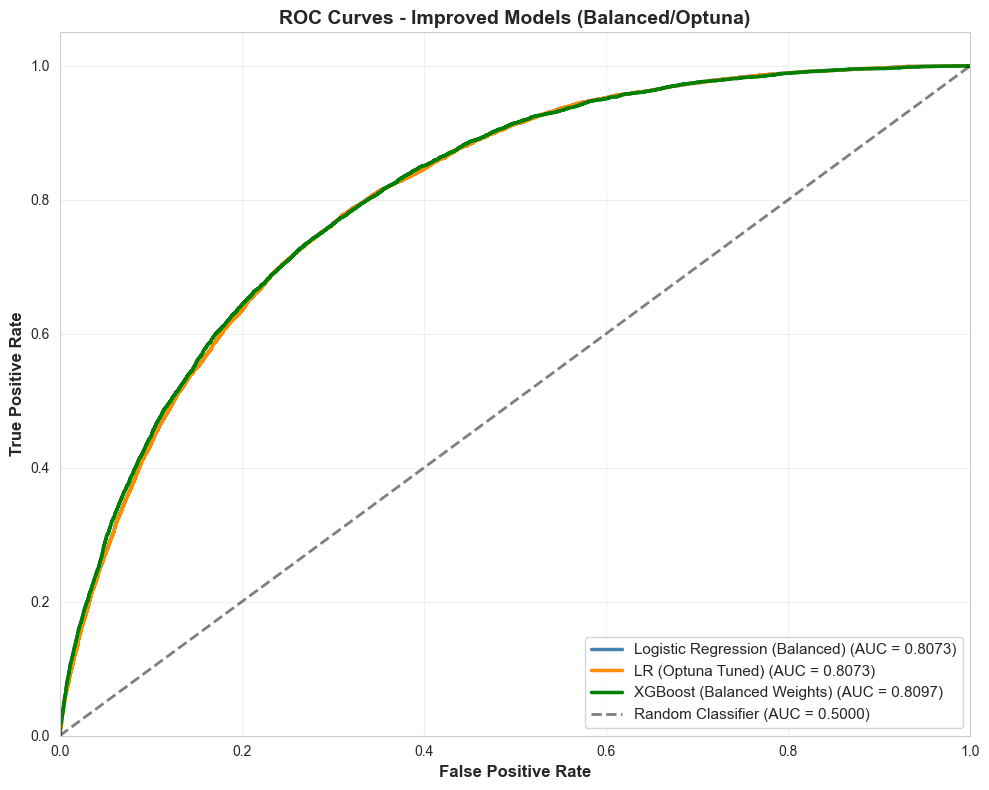
****

Figure 8 - Top 3 models ROC-AUC comparison

**All models metrics comparison**

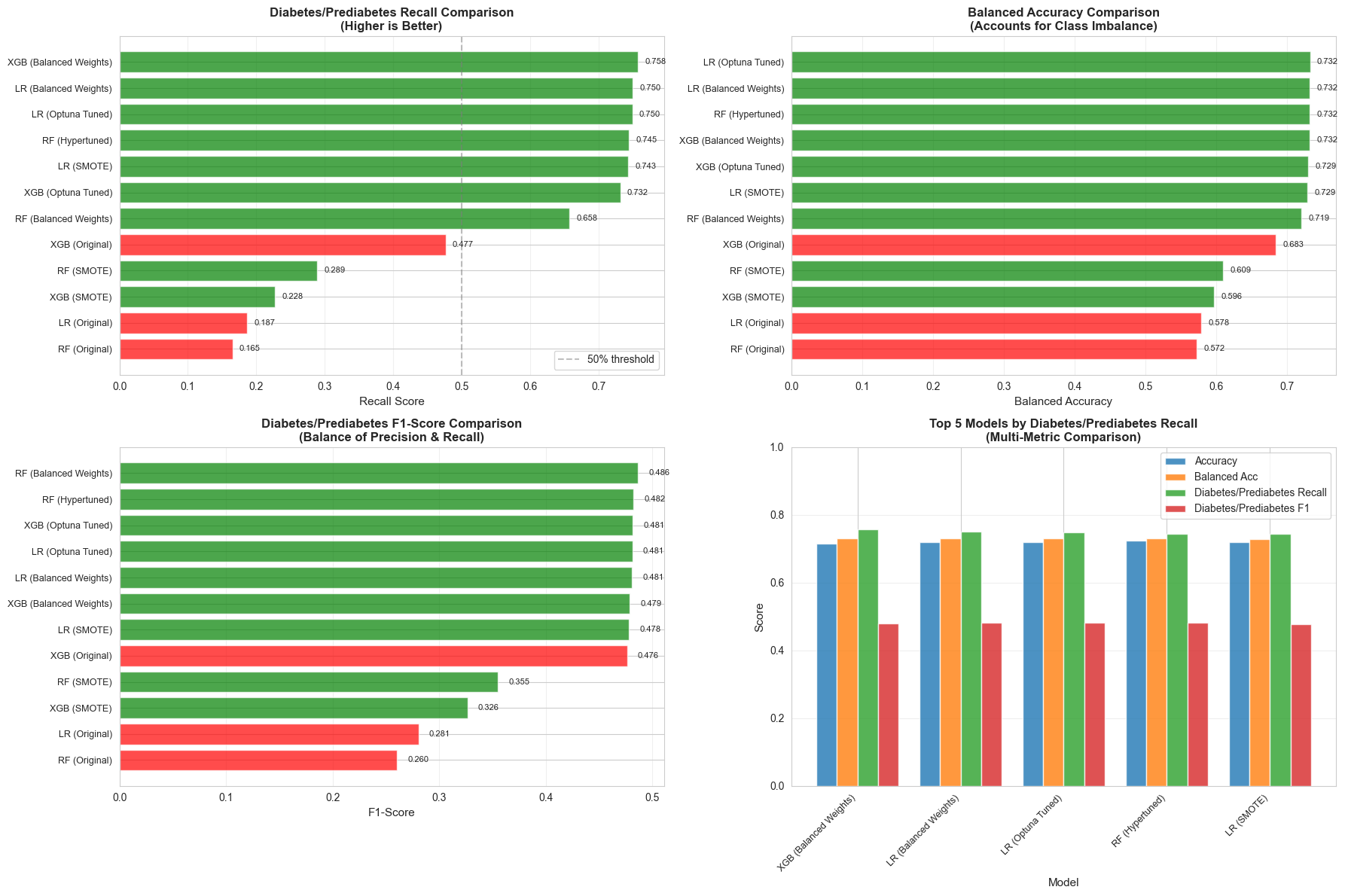
****

Figure 9 - All models metrics comparison

Based on the comparison results, the XGBoost model with balanced class weights achieved the highest recall for the diabetes/prediabetes class, while maintaining performance comparable to the other models across accuracy, F1-score, and balanced accuracy. Given the clinical context of the problem, prioritizing recall is appropriate, as failing to identify a diabetic individual (false negative) is more harmful than incorrectly flagging a non-diabetic individual (false positive). Therefore, the slight emphasis on recall represents a deliberate and justified trade-off, making the balanced XGBoost model the most suitable choice for this task.

## Explainability

Explainability was achieved using SHAP on the balanced XGBoost model. SHAP values identified the most influential features for predicting Diabetes/Prediabetes at both global (feature importance) and local (individual prediction) levels, improving transparency and interpretability of the model’s decisions, as the graphs shows:

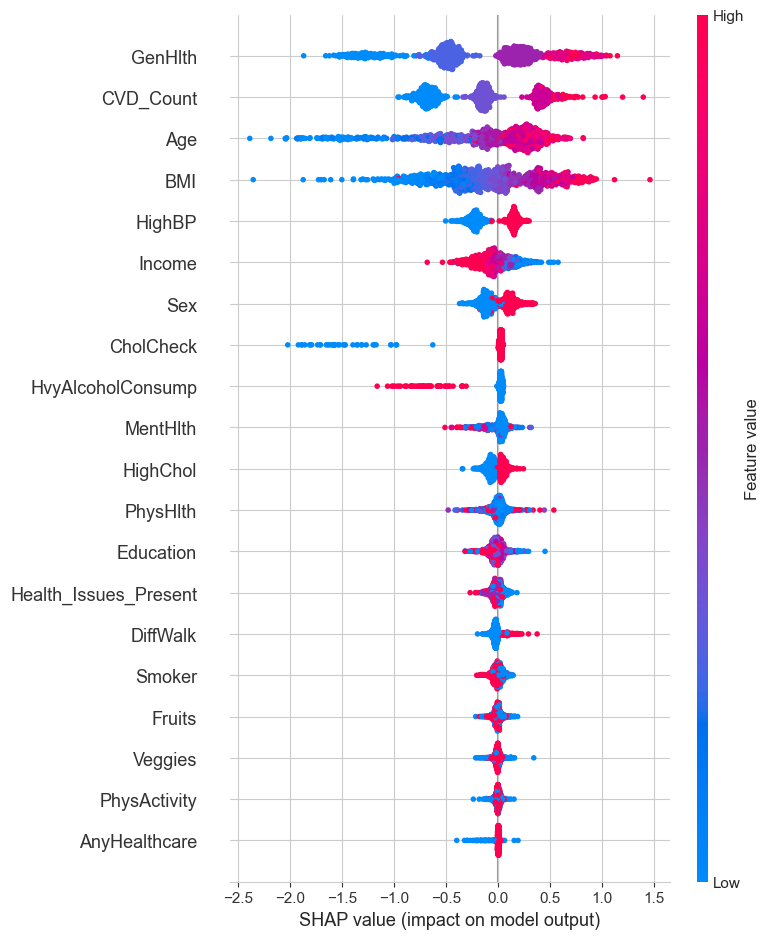
**SHAP value**

Figure 10 - SHAP value

**Features contribution in an individual result**

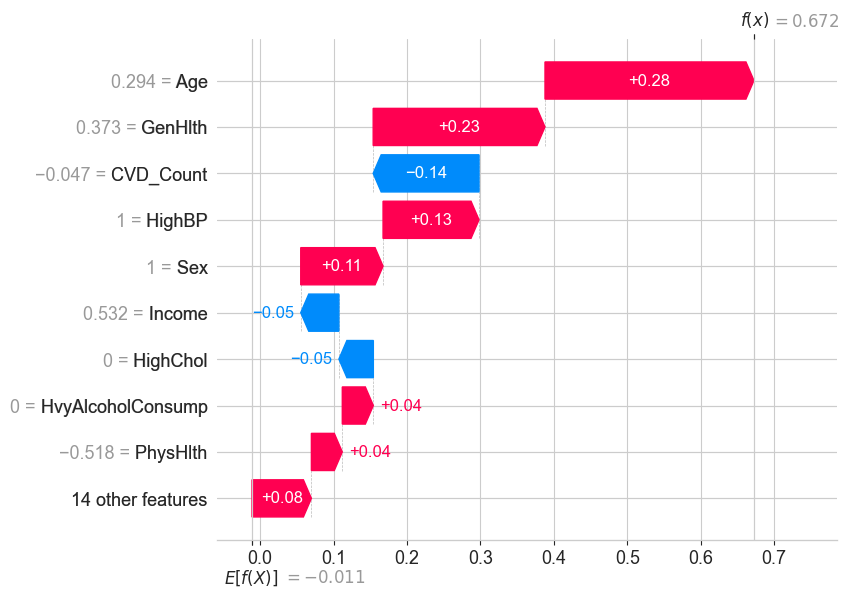
****

Figure 11 - local explainability

**Top 6 most important features**

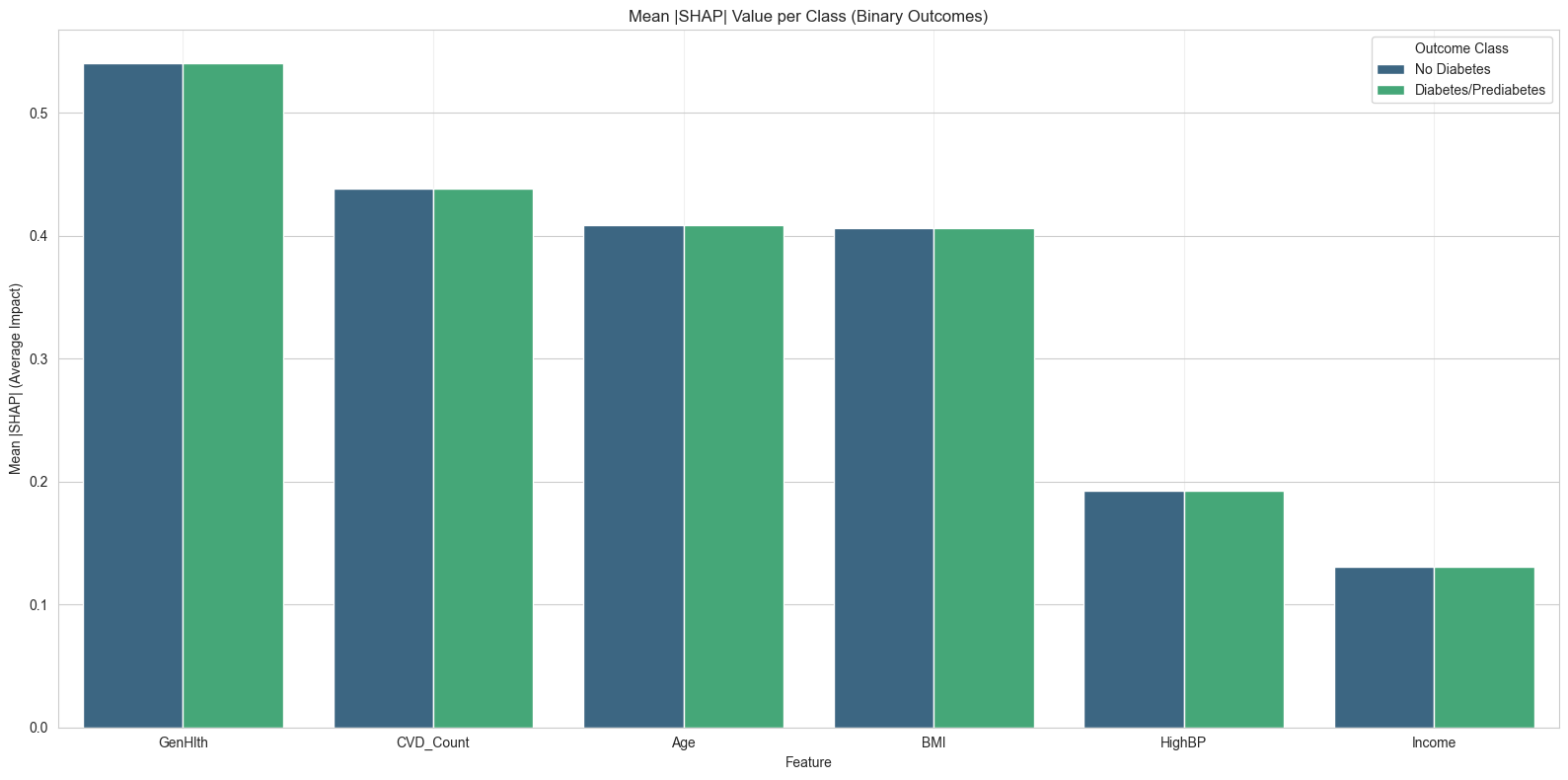
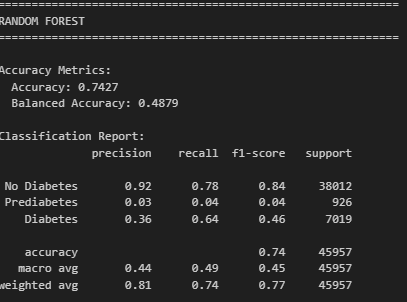
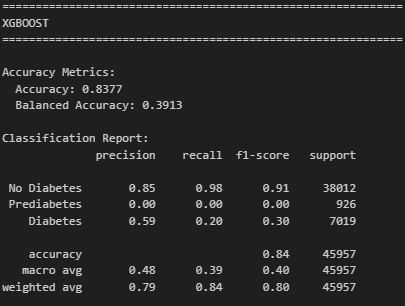
****

Figure 12 - top 6 important features SHAP

**NOTE:**

here are the results of the models showing three categories showing the effect on the pre-diabetes class:





In conclusion, we can see that the 3-class models fail to detect Prediabetes, with very low class wise performance and misleading high accuracy. Binary classification (Diabetes/Prediabetes vs No Diabetes) reduces imbalance, improves recall, and is more clinically and statistically meaningful, which was a considerable reason to move to binary classification

Unsupervised Learning

# Problem Definition and Rationale

## Problem Definition:

Identify natural groupings of individuals based on health and lifestyle indicators to uncover patterns associated with diabetes risk.

## Rationale for Unsupervised Learning:

Clustering is suitable as there is no pre-defined label for risk groups, allowing hidden structures in the data to emerge naturally.

## Practical Relevance:

Reveals high-risk population segments and lifestyle patterns, informing targeted public health interventions and preventive strategies.

## Analytical Value:

Enables exploration of population heterogeneity, identification of meaningful clusters, and insights that complement supervised predictive models.

# Unsupervised Learning Model Development

For the same preprocessing and analysis, we had for the data in the classification part we have implemented two main algorithms for the clustering part:

## K-Means

* **Determination of the optimal number of clusters using evaluation techniques such as the Elbow Method and Silhouette Score.**

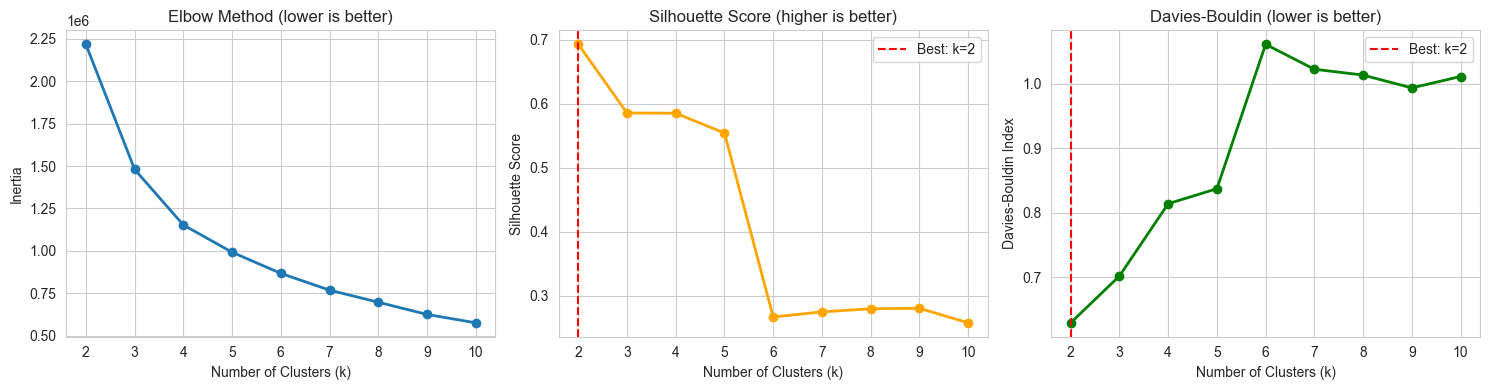
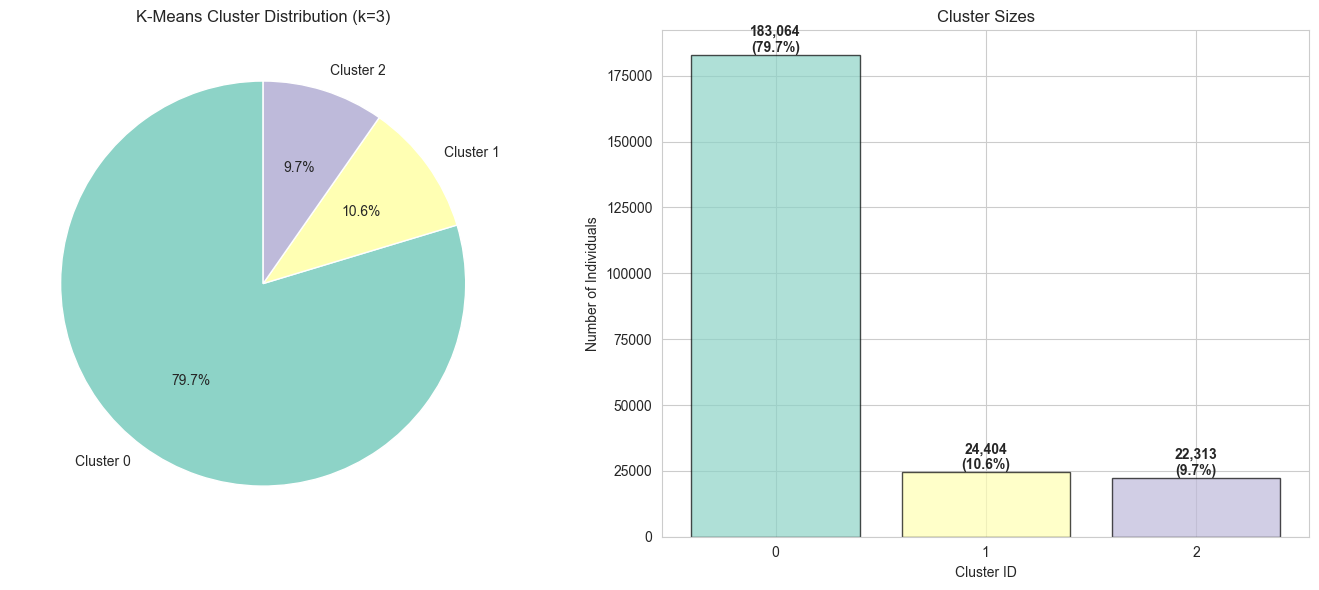
****

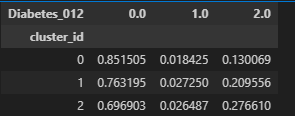
Figure 13 - elbow method

Selecting three clusters provided a clear and interpretable grouping based on the available health and lifestyle features. The clusters can be meaningfully described as: *low-risk group* with generally healthier behaviours and better health indicators, *moderate-risk group* showing mixed lifestyle patterns and early risk factors, and *high-risk group* characterised by poorer health status and indicators commonly associated with diabetes. This structure balances clustering quality with interpretability, making it suitable for understanding population risk profiles, besides to the high silhouette metric in comparison to other Ks which would be a strong statistical support.

* **Model training using the selected number of clusters to assign everyone to a specific cluster.**

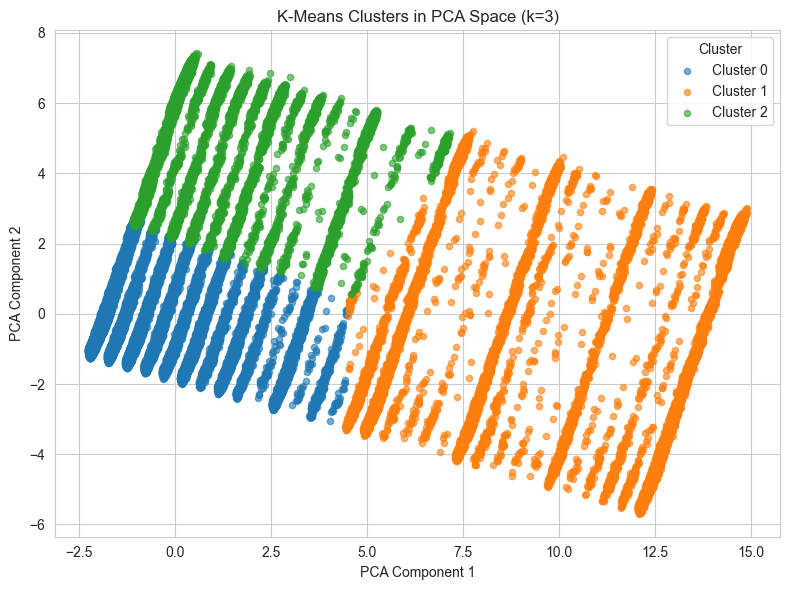
****

* **Analysis of cluster centroids to understand the distinguishing characteristics of each group.**

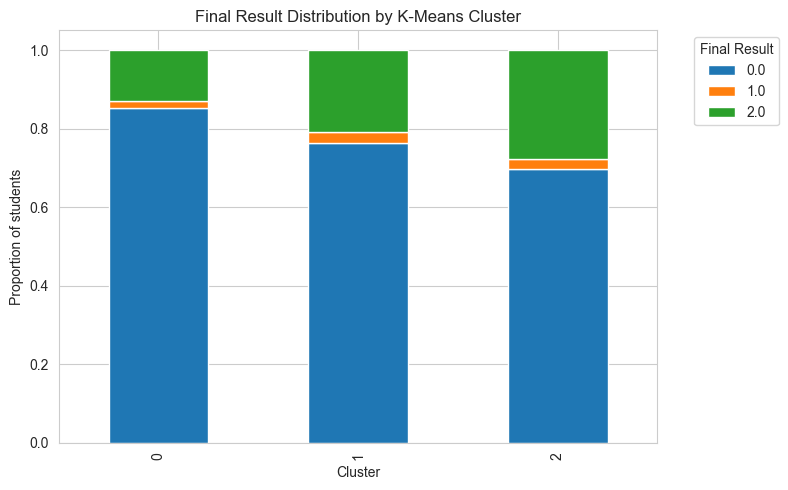
****

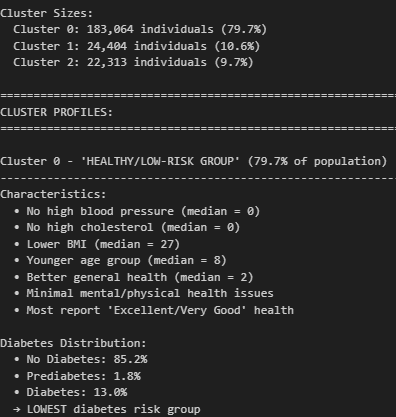
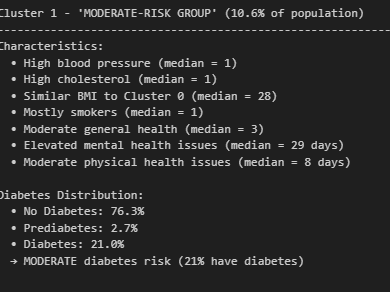
The table indicates a clear diabetes risk gradient across clusters. Cluster 0 is predominantly non-diabetic, representing a low-risk group. Cluster 1 shows an increased proportion of diabetic individuals, indicating moderate risk. Cluster 2 has the highest proportion of diabetes cases, representing a high-risk group. This supports the interpretation of the three clusters as low, medium, and high diabetes risk populations.

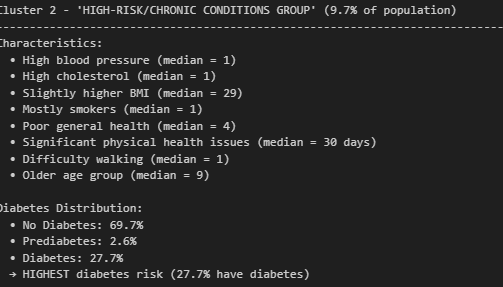
* **PCA visualization**

****

* **Interpretation of clusters in the context of diabetes risk and lifestyle patterns.**

****

****

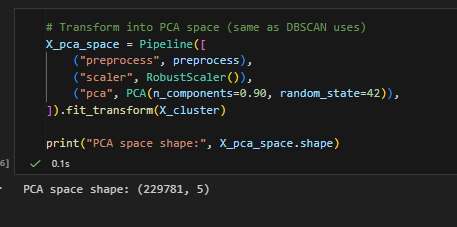
****

According to the figures:

* + Three clusters (k=3) reveal clear population segmentation by health status and diabetes risk.
  + Cluster 0 (≈80%) represents a healthy, low-risk group with good general health and the lowest diabetes prevalence (13%).
  + Cluster 1 (≈11%) shows moderate risk, characterised by hypertension, high cholesterol, smoking, and increased mental health issues (21% diabetic).
  + Cluster 2 (≈10%) is a high-risk group with poor physical health, chronic conditions, older age, and the highest diabetes prevalence (27.7%).

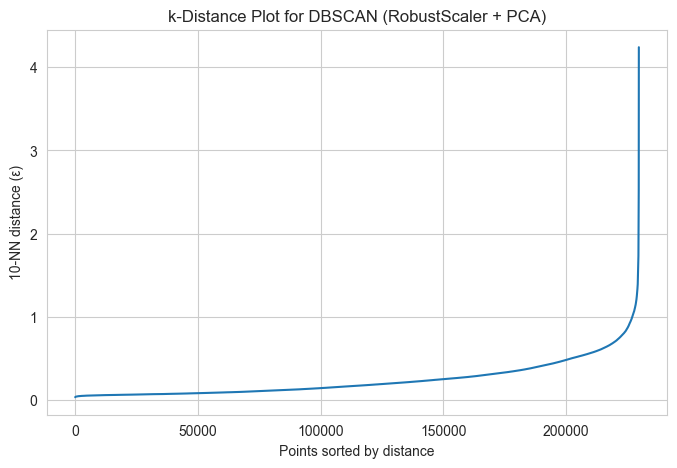
## DBSCAN

* **PCA transformation**

****

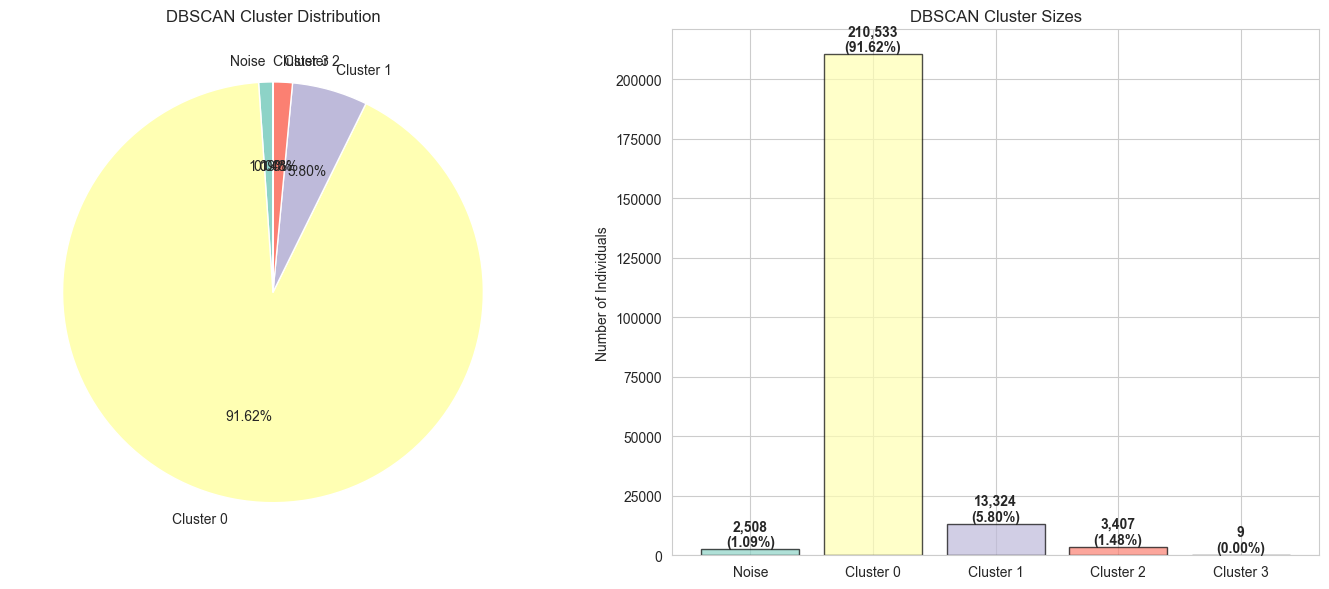
Specifying the number of PCA that would cover 90% of data’s total variance.

* **Selection of DBSCAN and parameter tuning (ε and min\_samples)**

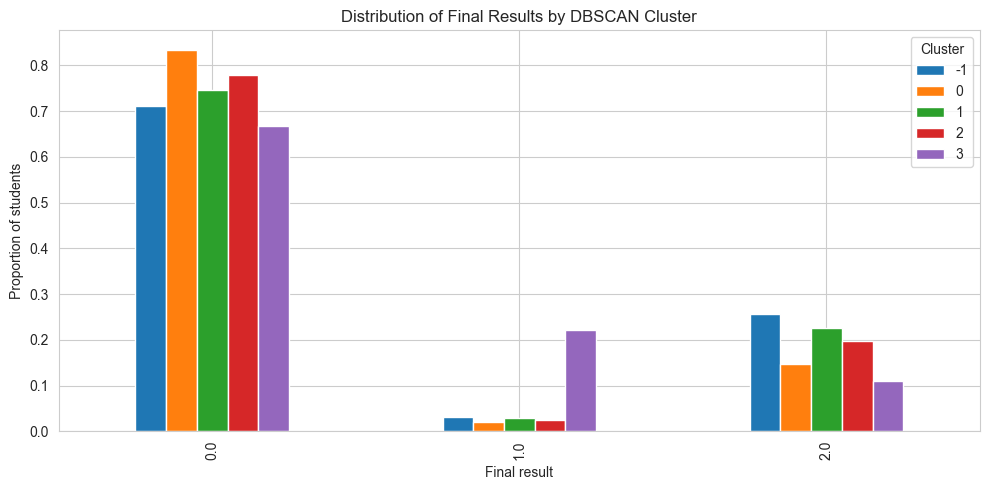


* **Model training and cluster assignment**

The trained model have created four clusters with a silhouette score of 0.59

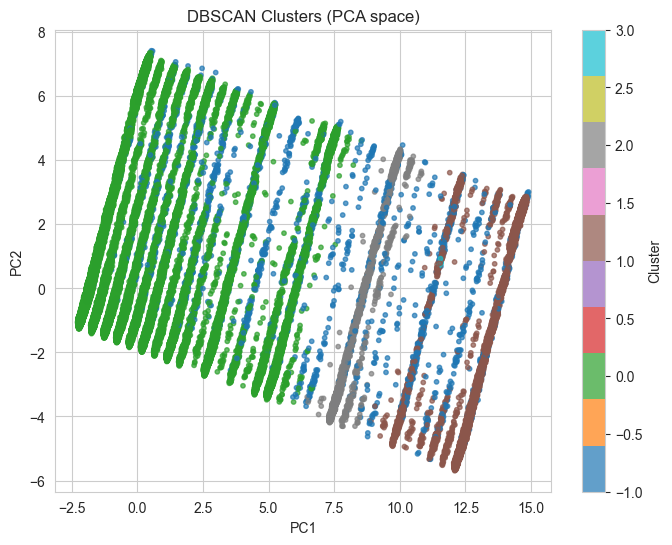
****

* **Analysis of discovered clusters and noise points**

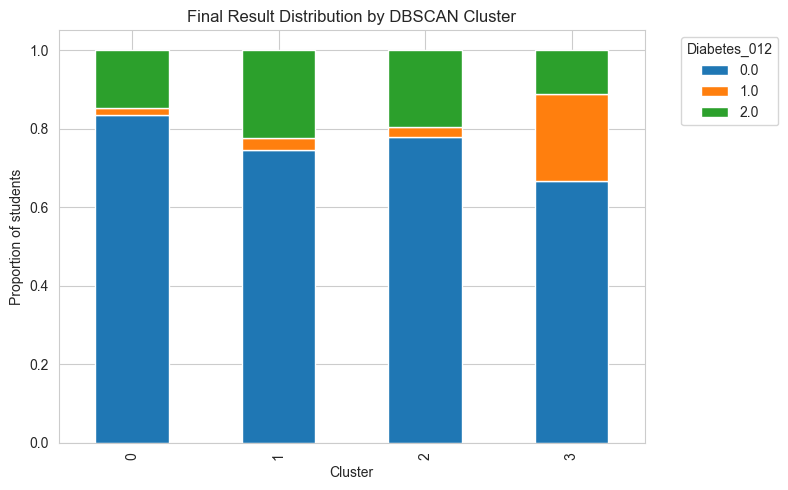
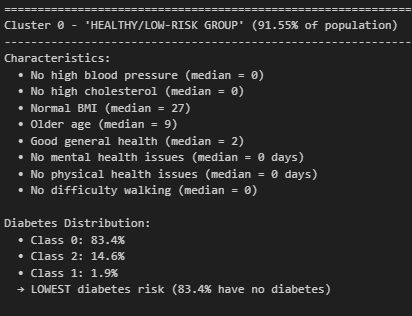
****

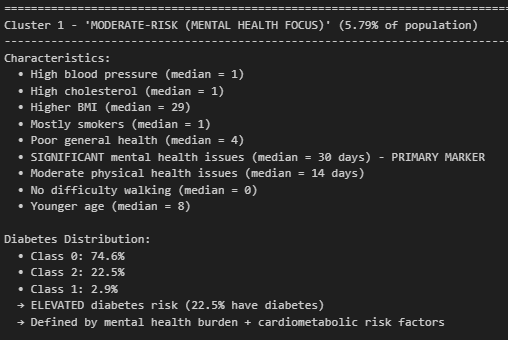
Where 0 = no diabetes, 1 = pre-diabetes, and 2 = diabetes

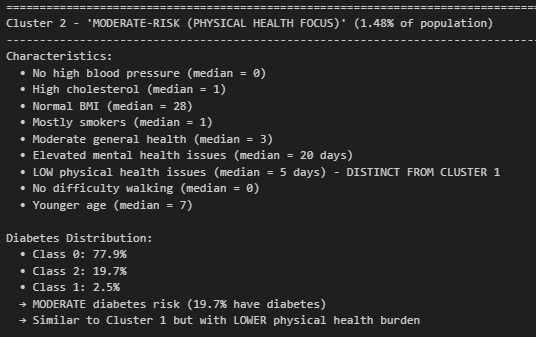
* **PCA visualization**

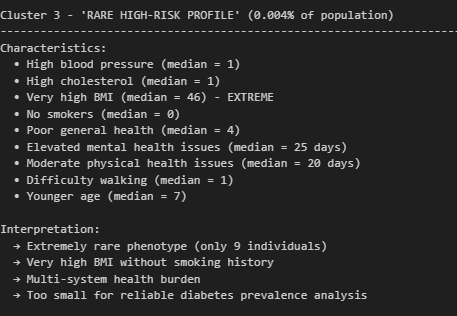


* **Interpretation in the context of diabetes risk and population health**









* DBSCAN identified four clusters and a small noise group using ε = 0.8 and min\_samples = 10.
* Cluster 0 dominates, indicating a largely healthy population.
* Clusters differ mainly by health dimensions, separating mental-health-driven risk from more balanced profiles.
* Noise points capture outliers with extreme BMI values.
* A rare cluster highlights obese non-smokers with multiple chronic conditions, revealing atypical high-risk profiles.

Association Rules Mining

# Problem Definition and Rationale

## Problem Definition:

Identify frequent combinations of health and lifestyle factors associated with diabetes risk.

## Rationale for ARM:

Association rule mining uncovers hidden relationships in the data without relying on predefined labels.

## Practical Relevance:

Reveals co-occurring risk factors that can inform targeted public health interventions.

## Analytical Value:

Provides interpretable insights using support, confidence, and lift, complementing clustering and classification results

# Model Development

Association rule mining was applied to discover frequent item sets and generate rules using standard evaluation metrics.

## Feature engineering for ARM

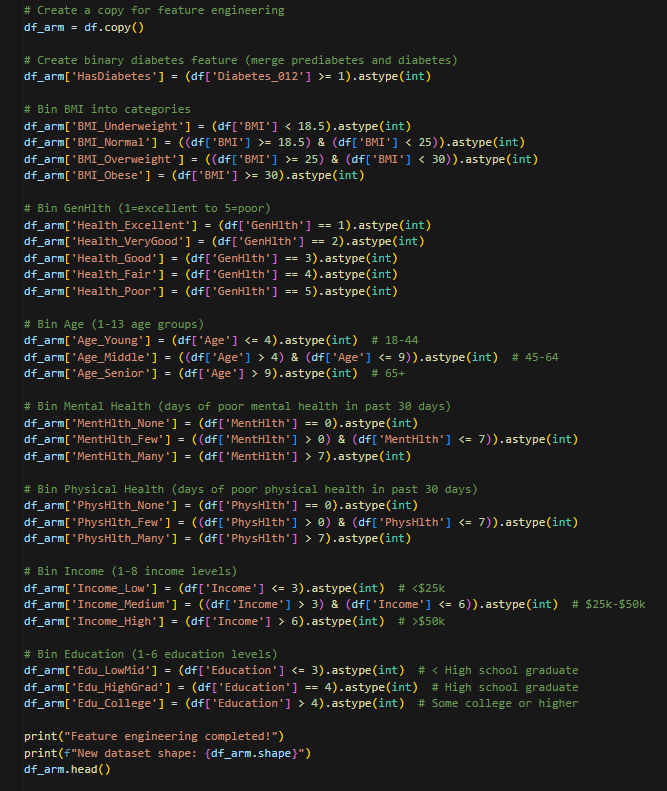
For association rule mining, we need binary features. The dataset contains:

· Binary features (already 0/1): HighBP, HighChol, Smoker, Stroke, etc.

. Categorical features: GenHIth, Age, Education, Income (need binning)

. Continuous features: BMI, MentHIth, PhysHIth (need binning)

So, we created meaningful categories and labelled them for interpretability.

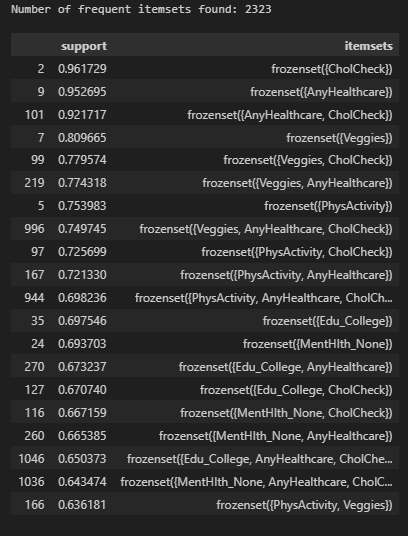


Transaction Construction

* Represented each individual as a transaction containing all applicable binary health and lifestyle items.
* Removed infrequent items to reduce sparsity and improve computational efficiency.

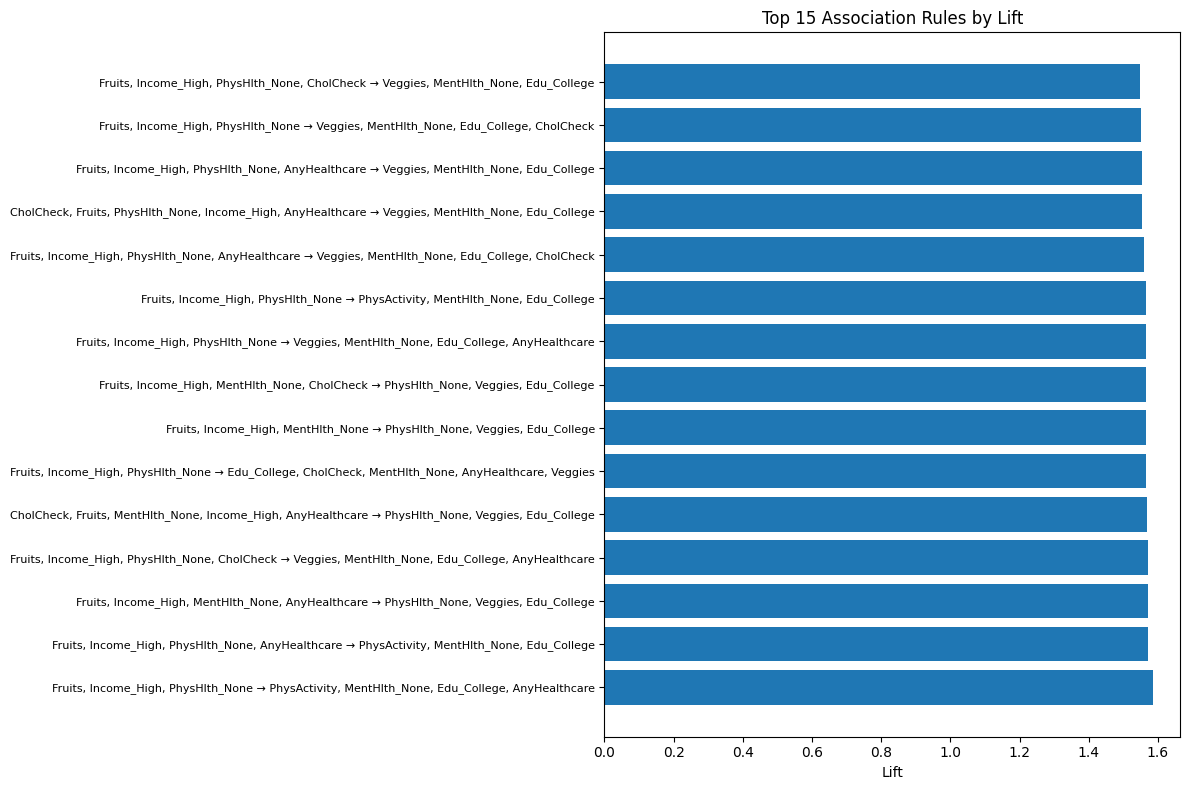
## Training Apriori

* Applied the Apriori algorithm to identify frequent item sets based on a minimum support threshold.
* Generated association rules from frequent item sets using confidence and lift criteria.

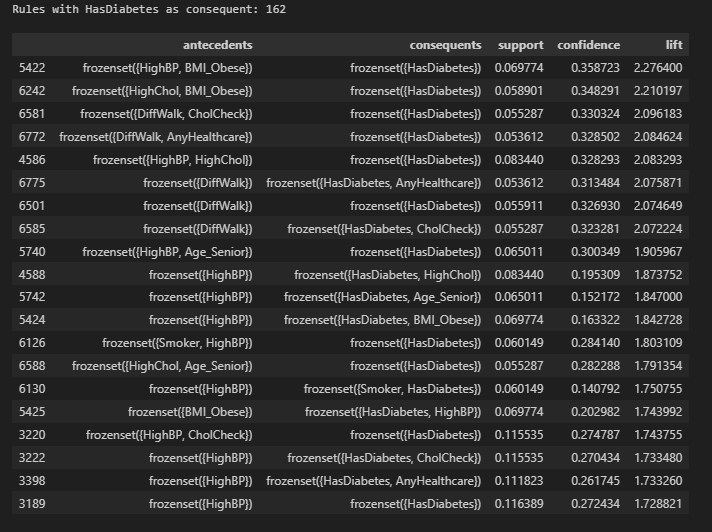


## Rule Evaluation and Interpretation

* Evaluated rules using support, confidence, and lift to assess relevance and strength.
* Interpreted high-lift rules to identify non-random associations between lifestyle factors, health conditions, and diabetes risk.



By filtering rules that has diabetes as consequent, we can see a string relation between diabetes and high blood pressure in the first level which aligns with our results in bot supervised and unsupervised models.



Deployment and Business Alignment

* Integrated Analytical Deployment:  
  The combined outputs of supervised classification, clustering, and association rule mining provide a comprehensive decision-support framework. Classification models deliver individual level diabetes risk prediction, while clustering and association rules offer population-level insights and pattern discovery.
* Supervised Model Deployment:  
  The optimised XGBoost, Random Forest, and Logistic Regression models can be deployed to estimate diabetes risk probabilities, supporting early detection and prioritisation of individuals for preventive screening and intervention.
* Unsupervised Model Deployment:  
  K-Means and DBSCAN clustering results enable segmentation of the population into distinct health risk profiles, supporting targeted public health campaigns, efficient resource allocation, and focused interventions for high risk and vulnerable groups.
* Knowledge Discovery and Policy Support:  
  Association rule mining provides interpretable co-occurrence patterns between lifestyle behaviours, health conditions, and diabetes, reinforcing insights from supervised and unsupervised models and informing evidence-based healthcare policy design.
* Operational Integration and Monitoring:  
  The models can be integrated into periodic survey analysis pipelines or public health dashboards, with regular retraining and validation to maintain performance as population health trends evolve.
* Ethical, Fairness, and Sustainability Considerations:  
  Deployment must address data bias, class imbalance, and transparency to ensure equitable decision making and responsible use of AI in public health contexts.

References

* Wirth, R. and Hipp, J., 2000, April. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).
* Schröer, C., Kruse, F. and Gómez, J.M., 2021. A systematic literature review on applying CRISP-DM process model. *Procedia computer science*, *181*, pp.526-534.
* Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.
* Chen, T., 2016. XGBoost: A Scalable Tree Boosting System. *Cornell University*.
* Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.
* Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996, August. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
* Agrawal, R., Imieliński, T. and Swami, A., 1993, June. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).
* Batista, G.E., Prati, R.C. and Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), pp.20-29.
* Schubert, E., Sander, J., Ester, M., Kriegel, H.P. and Xu, X., 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), pp.1-21.
* Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), pp.651-666.
* Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, pp.53-65.

Appendix

# Supervised learning

**Feature Distribution Analysis**

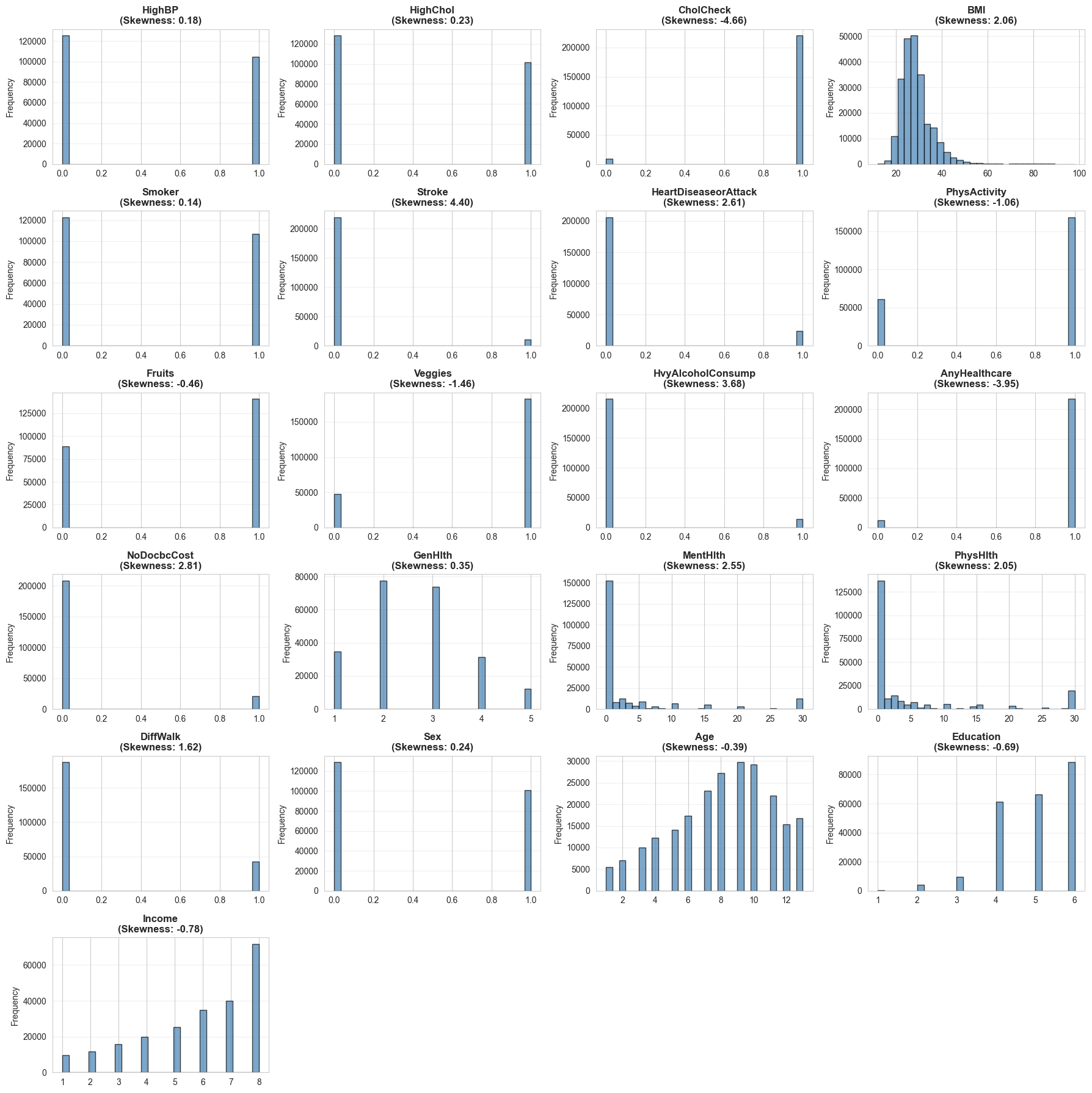


Figure 14-Feature Distribution Analysis

**Feature Engineering 1**

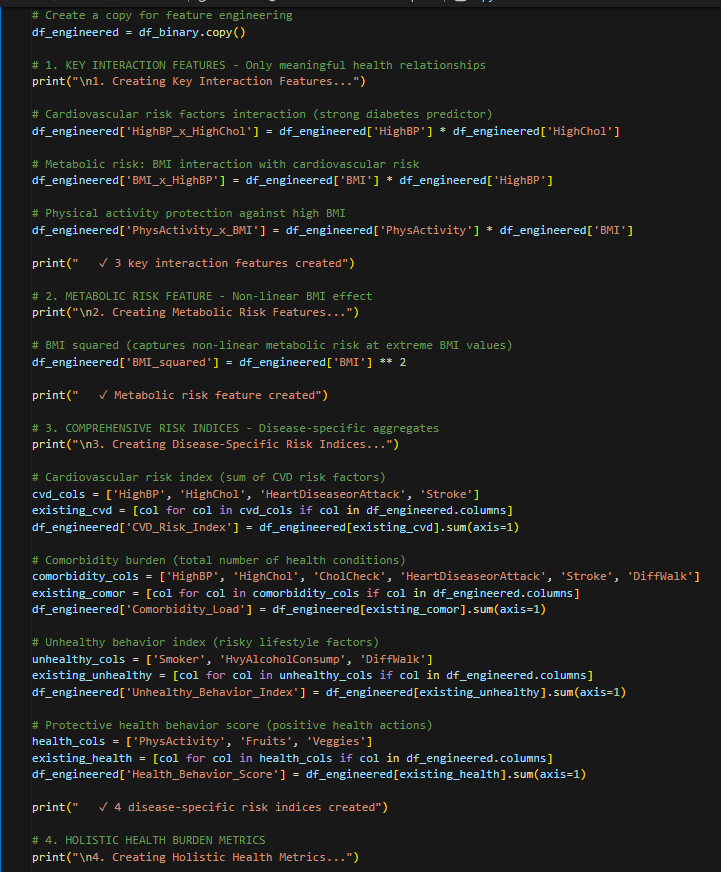


Figure 15 - new features

**Data splitting transforming & scaling**

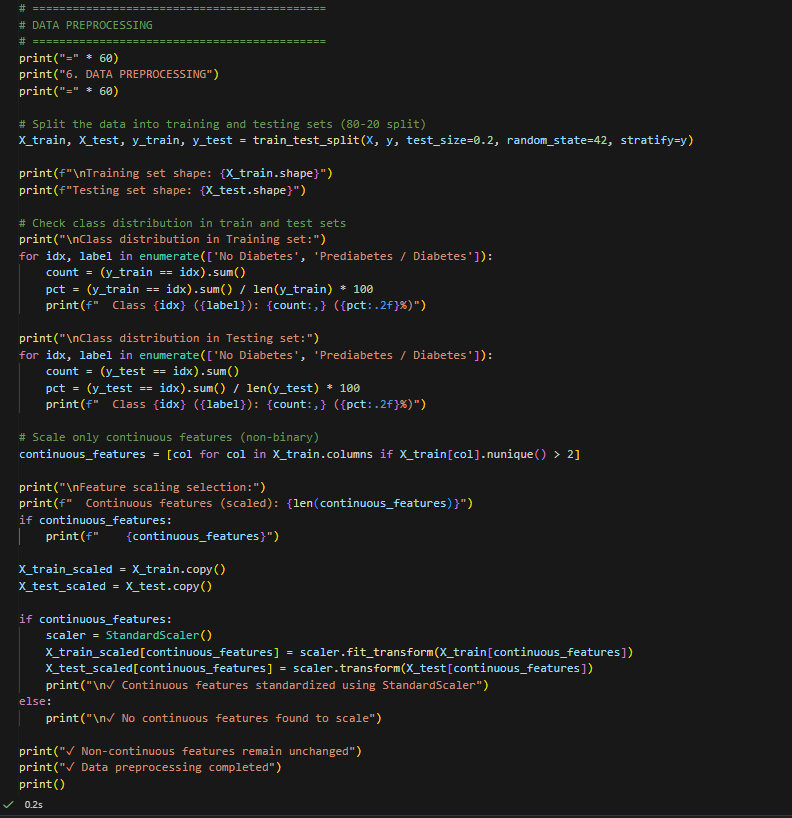
****

Figure 16 - Data splitting transforming & scaling

**Baseline models**

****

Figure 17 - Baseline models code

**Cross validation for baseline models**

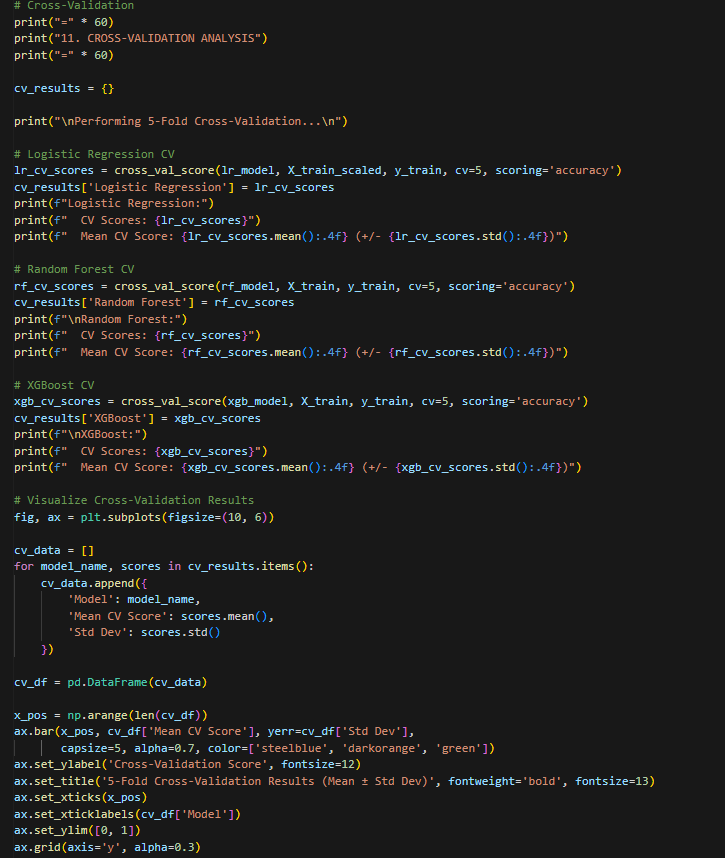
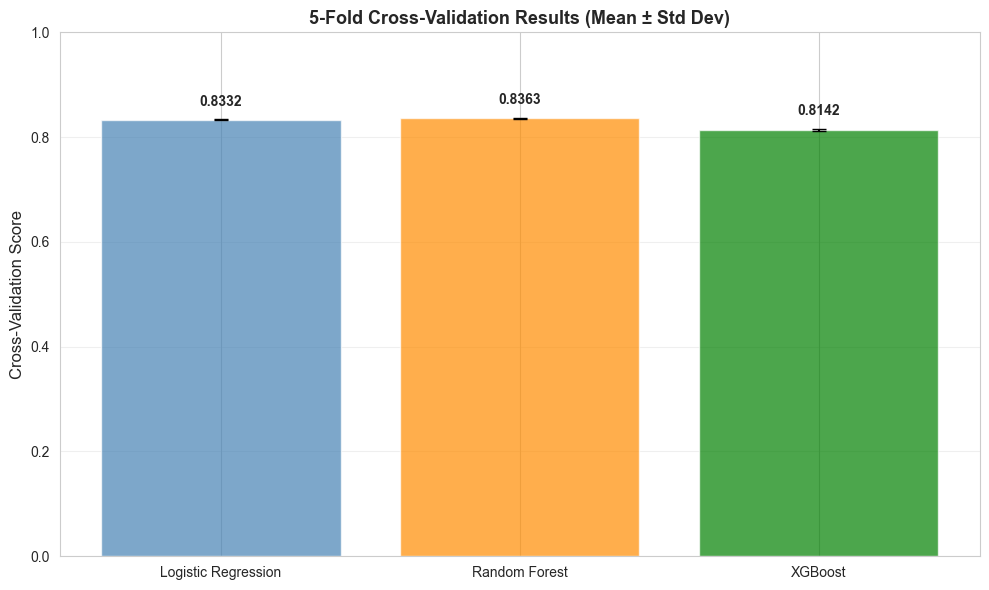
****

Figure 18 - Cross validation for baseline models****

**ROC-AUC for base models**

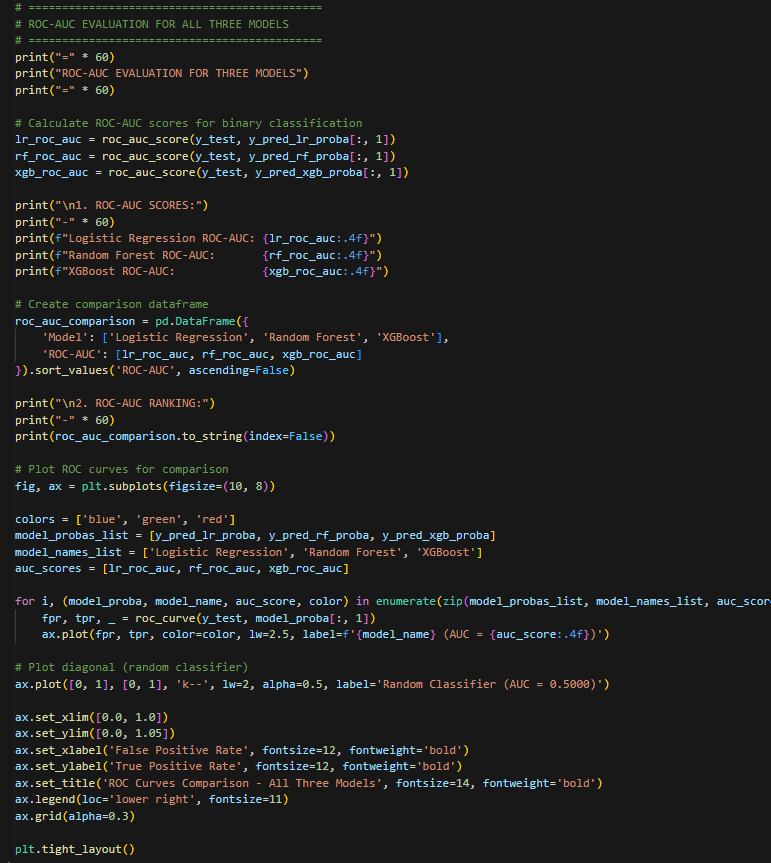
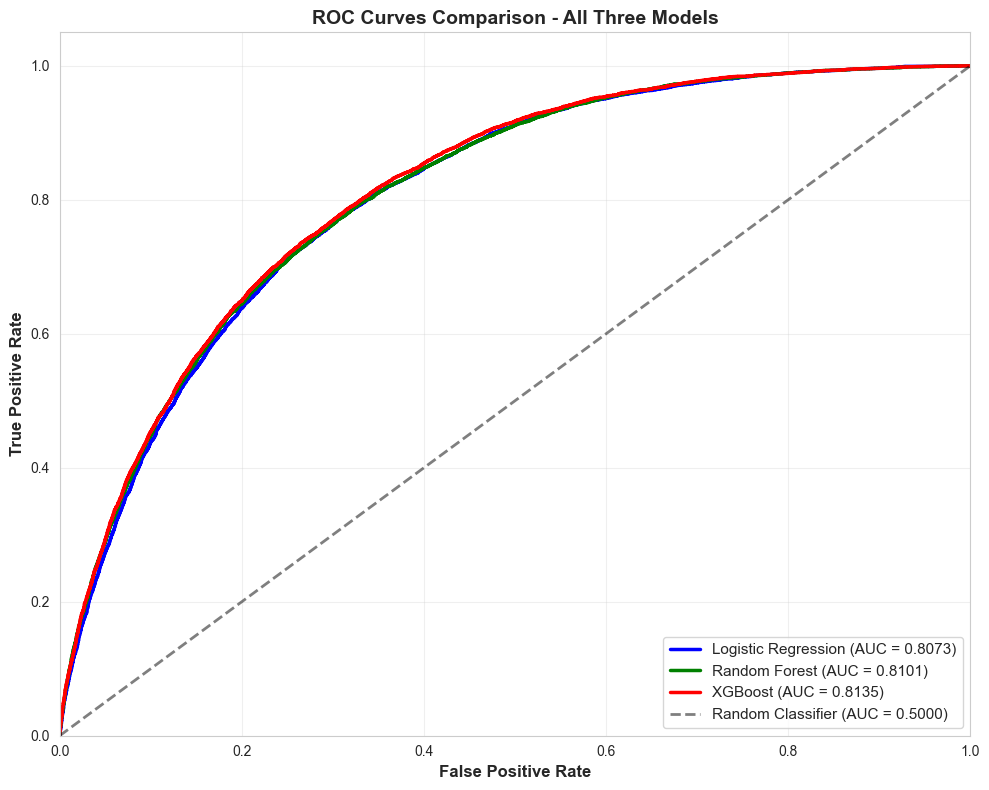


Figure 19 - ROC-AUC for base models



**Random forest grid search code**

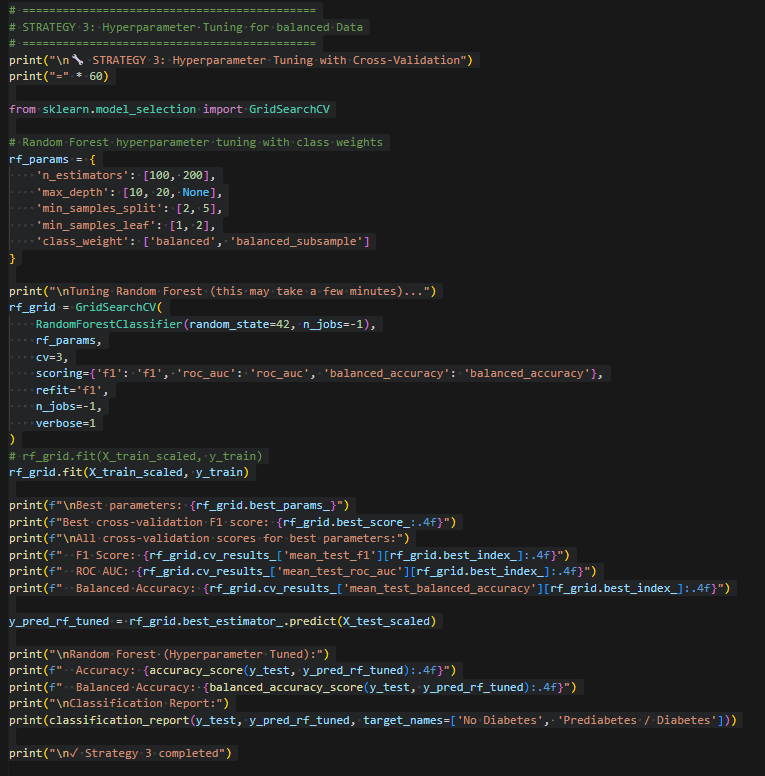
****

Figure 20 - random forest grid search CV

**XGBoost Optuna code**

****

Figure 21 - XGBoost optuna

**Logistic regression Optuna**

****

Figure 22 - Logistic regression optuna

**RFE code**

****

Figure 23 - RFE code

**Calibration code**

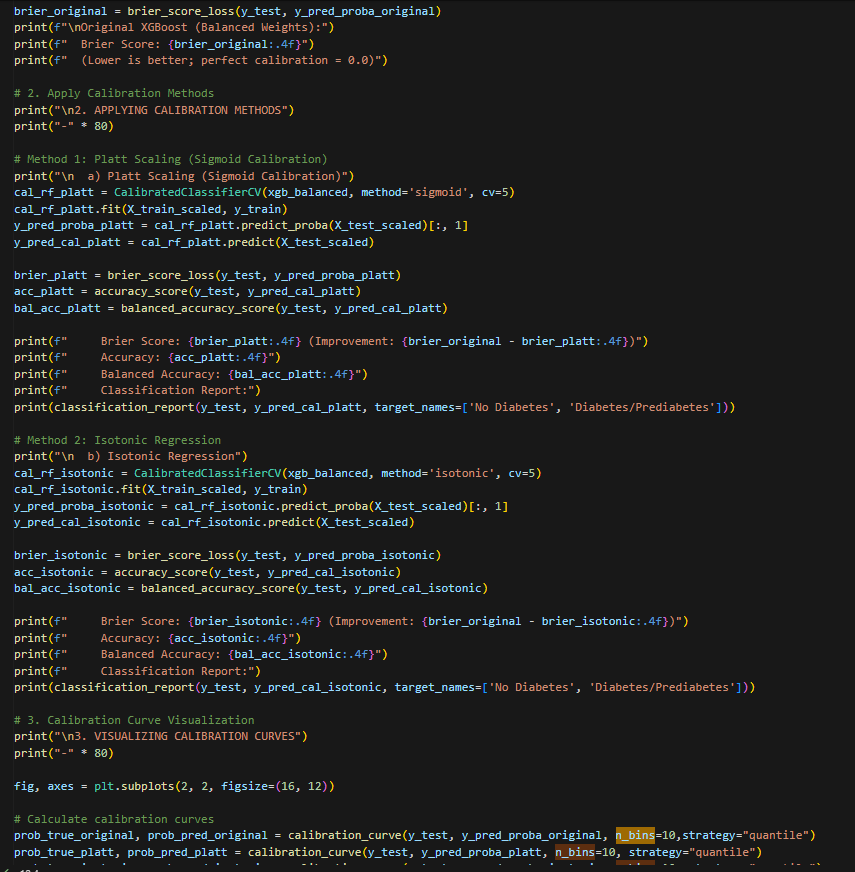
****

Figure 24 - Calibration code

**Threshold optimization**

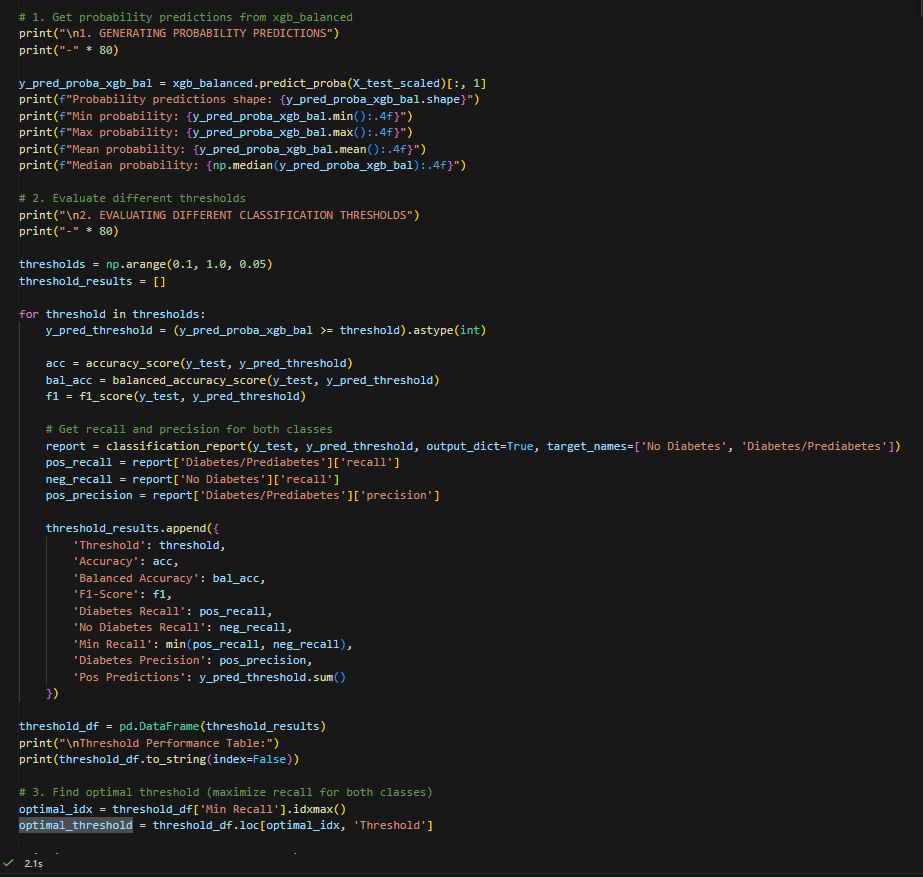


Figure 25 - Threshold optimization

**Explainability**

****

Figure 26 – Explainability

# Unsupervised learning

**Elbow method with silhouette score – Kmean**

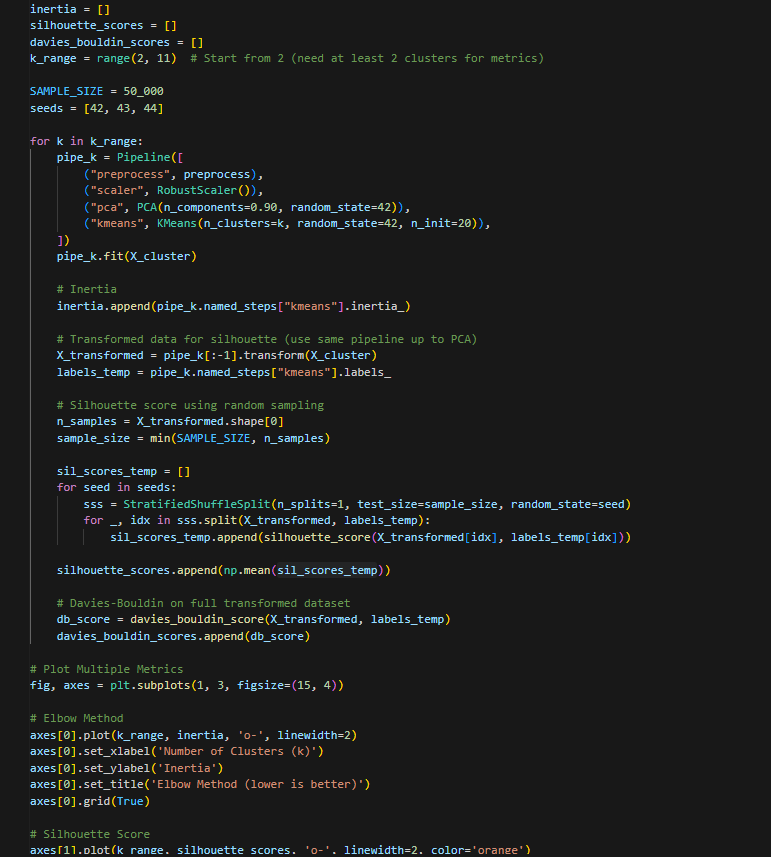


Figure 27 - Elbow method with silhouette score – Kmean

**Kmean training and visualizing code**

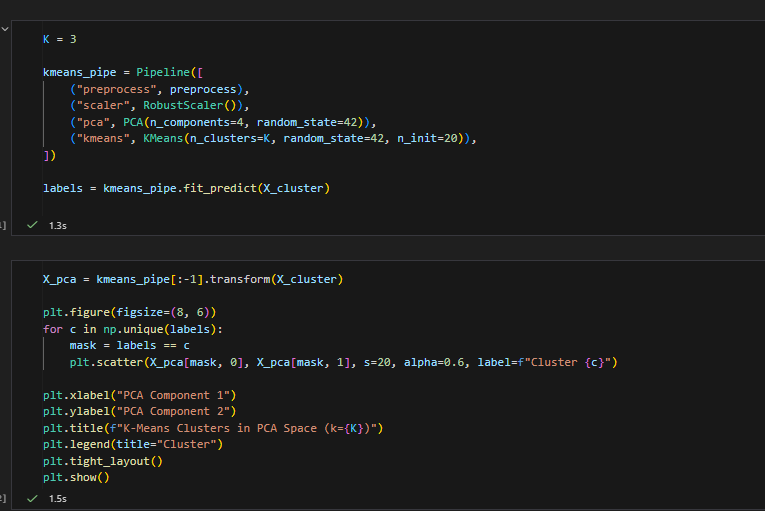
****

Figure 28 - Kmean training and visualizing code

**Kmean post analysis**



Figure 29 - Kmean post analysis

# DBSCAN

**PCA and hyperparameter tuning**

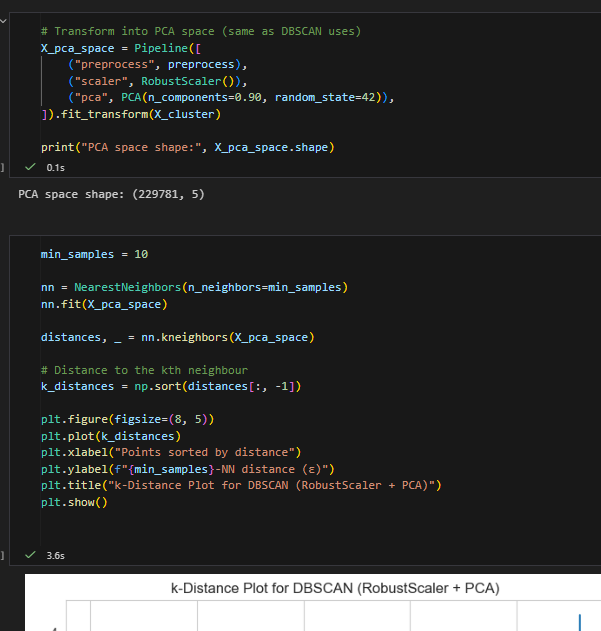


Figure 30 - PCA and hyperparameter tuning

**Pipeline and plotting**

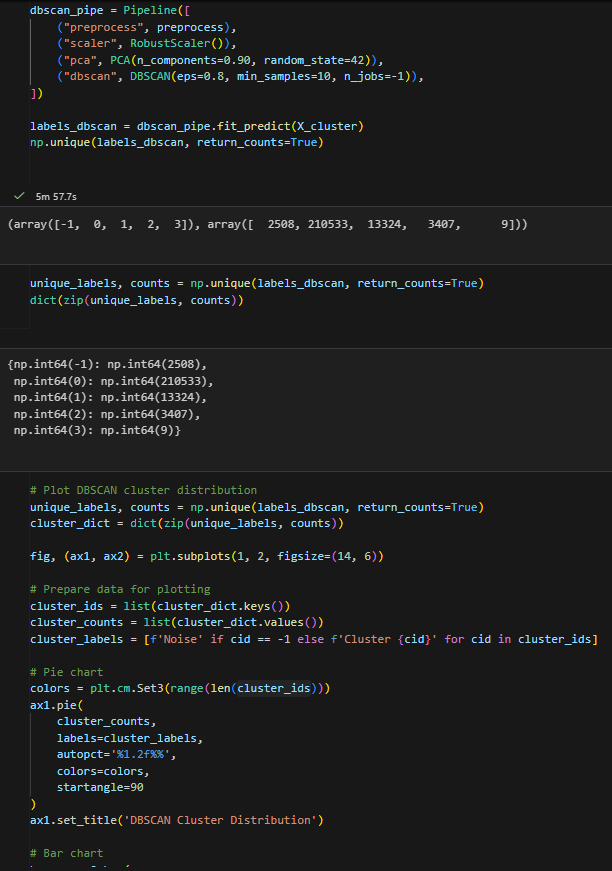


Figure 31 - Pipeline and plotting

**Evaluation**

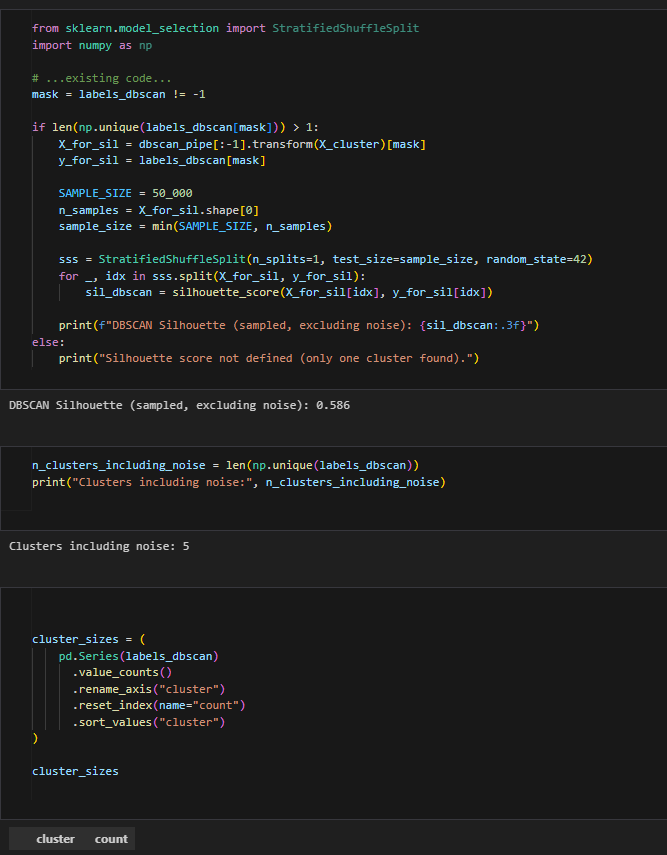
****

Figure 32 – Evaluation

**Post analysis**

****

Figure 33 - DBSCAN Post analysis

# Association rules mining

**Feature engineering ARM**

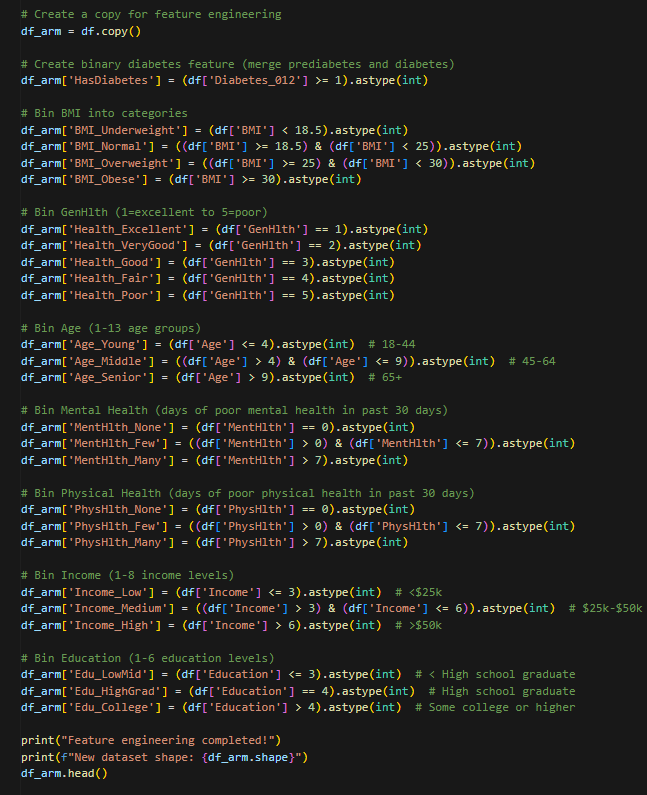
****

Figure 34 - Feature engineering ARM

**Train-test split with Apriori**

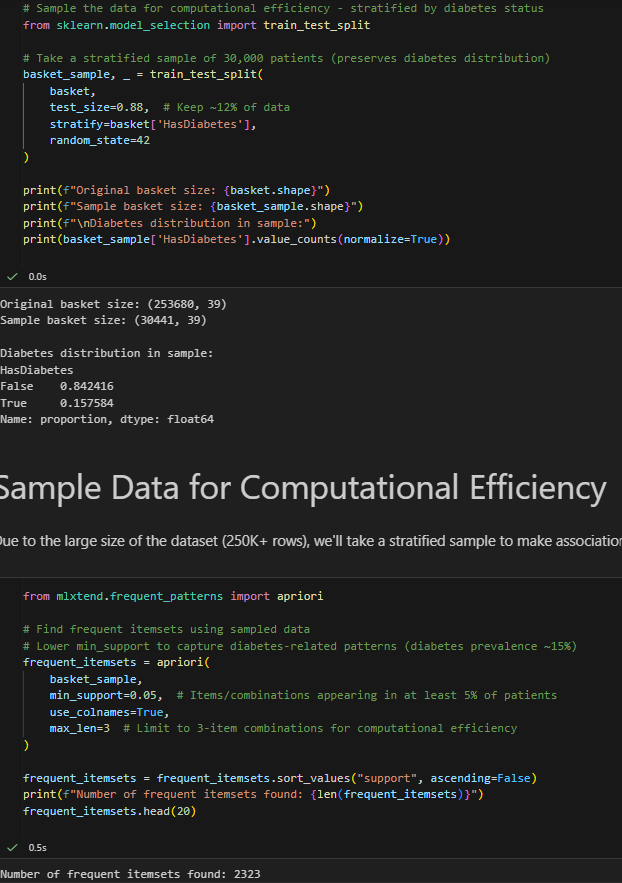


Figure 35 - Train-test split with Apriori

**Filtering rules**

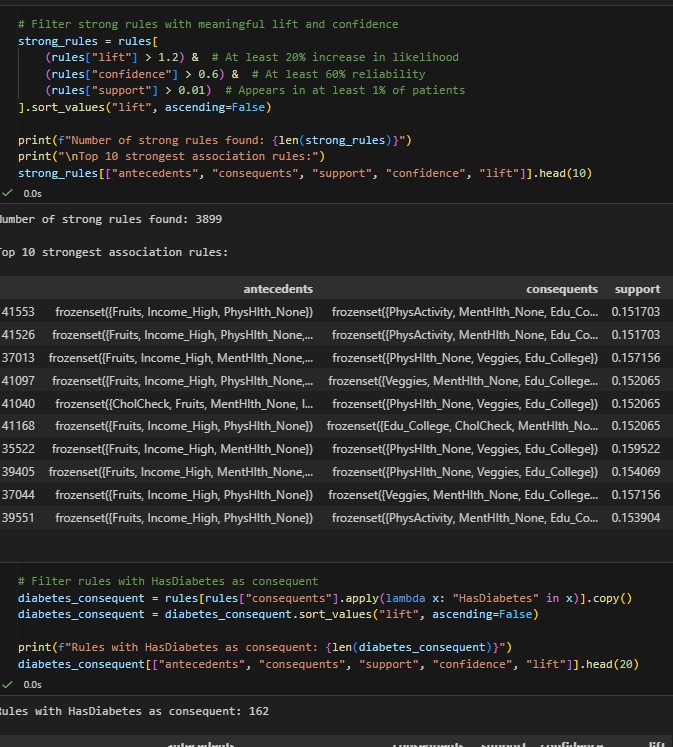
****

Figure 36 - filtering rules