Wydział Matematyki
i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA

Diabetes Classification Report

Mateusz Iwaniuk, Hubert Kowalski

Agenda

01 Project objectives and data sources

02 Data Exploration

03 Selecting the best model

04 Tuning the model to maximize accurate predictions

05 Interpretation of the results

Introduction

Financial Burden:

Diabetes imposes a staggering annual cost of nearly $400 billion on the U.S. economy, emphasizing the need for effective management strategies.

Lifestyle Intervention:

Lifestyle changes and medical treatments can help mitigate diabetes complications, underlining the importance of early diagnosis and intervention.

Awareness Gap:

Despite its prevalence, roughly 1 in 5 diabetics and 8 in 10 prediabetics are unaware of their condition, highlighting the critical need for increased awareness and accessibility to healthcare services.

Objectives

Use supervised machine learning methods to make accurate predictions about the risk of **type II diabetes** based on available **health factors**.

Make the model user-friendly and **easy to interpret** for hospitals and private clinics to assist in diagnosis of type II diabetes.

Emphasize the prediction of **probabilities** rather than solely determining whether the individual has diabetes or not.

General project info

The project was created as part of the Introduction to Machine Learning course at Warsaw University of Technology. Its aim was to gain experience with supervised machine learning methods.

This project was based on data related to diabetes. Its goal was to explore health and non-health factors influencing the risk of diabetes occurrence.

Data Sources

The data for this project was sourced from the Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC). Since its inception in 1984, the survey has collected responses from over 400,000 Americans annually, covering areas such as health-related risk behaviors, chronic health conditions, and the use of preventative services. This dataset comprises responses from 441,455 individuals and encompasses 330 features, which include both direct participant responses to survey questions and calculated variables based on individual participant data.

Specifically, for this project, a CSV dataset from the year 2015 was obtained from Kaggle, which was pre-cleaned and contained **253,680** rows and **22** columns.

Original dataset can be found at:
https://www.cdc.gov/brfss/annual_data/annual_data.htm

Exploratory Data Analysis

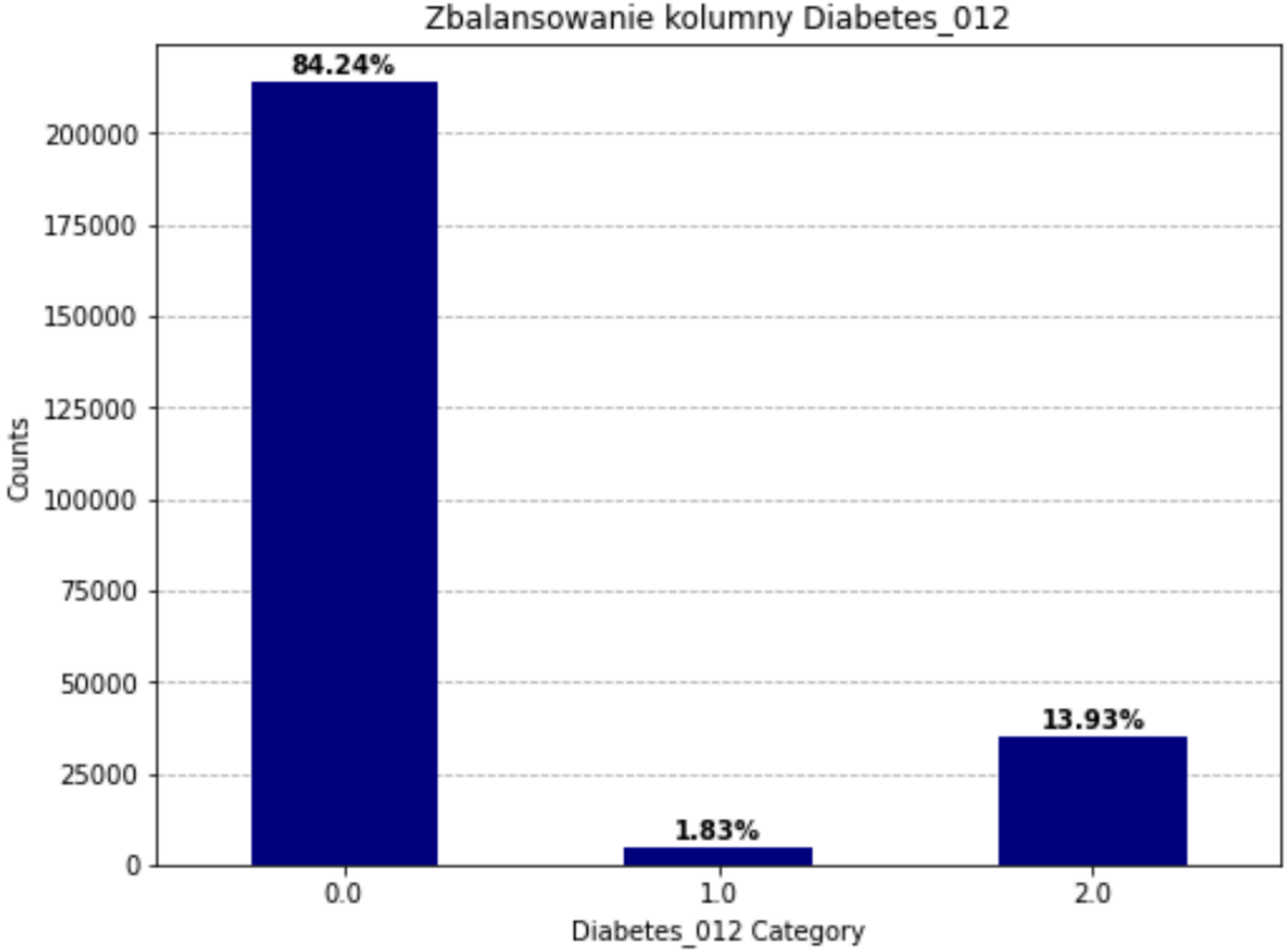
Our dataset primarily consists of categorical values spread across numerous columns, each meticulously curated to ensure ethical integrity and balance.

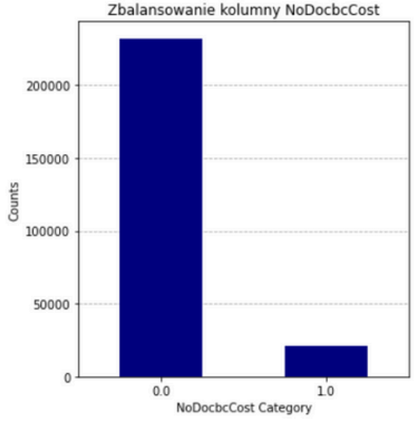
The most relevant features used in the dataset are:

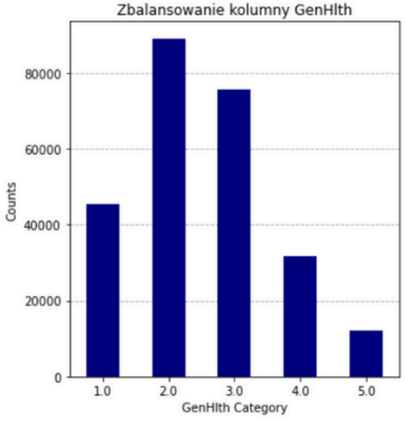
- **General Health**
- **High Blood Pressure**
- **High Cholesterol**
- **BMI**
- **Age**
- **Income**

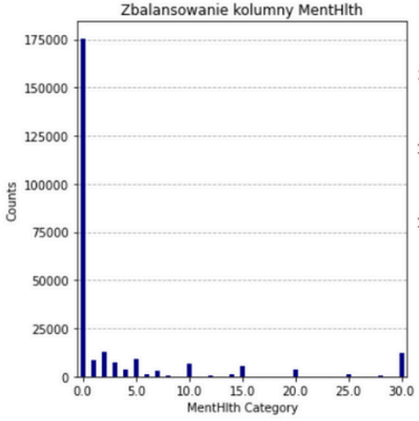
We meticulously handled the division of data to guarantee there are no missing values (NAs), supporting a robust and fair analytical foundation for our modeling process. The next slide presents the distribution of each of the most important features.

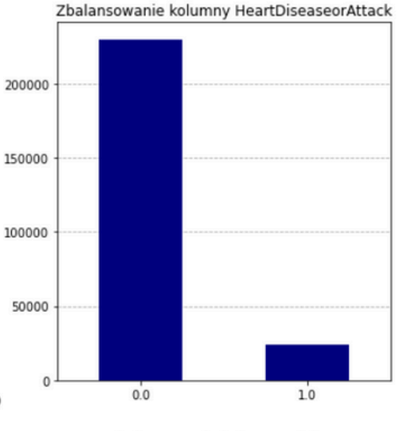
Exploratory Data Analysis



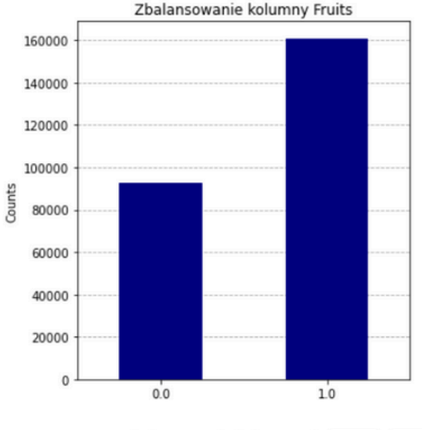

Zbalansowanie kolumny NoDocbcCost
Counts
NoDocbcCost Category

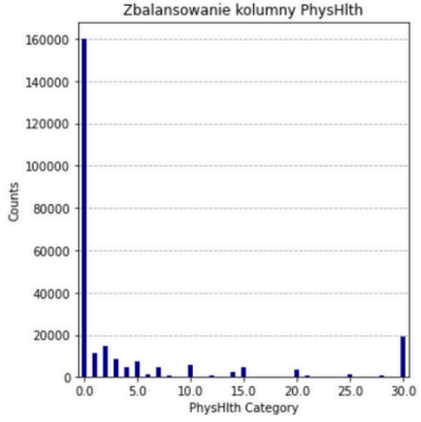

Zbalansowanie kolumny GenHlth
Counts
GenHlth Category

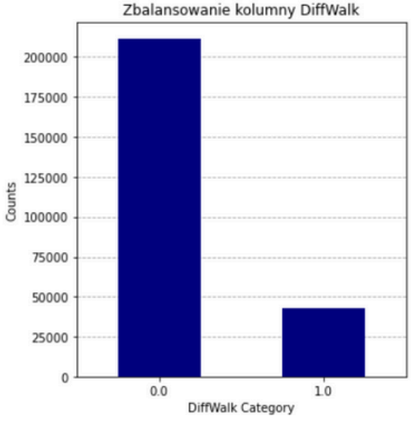

Zbalansowanie kolumny MentHlth
Counts
MentHlth Category

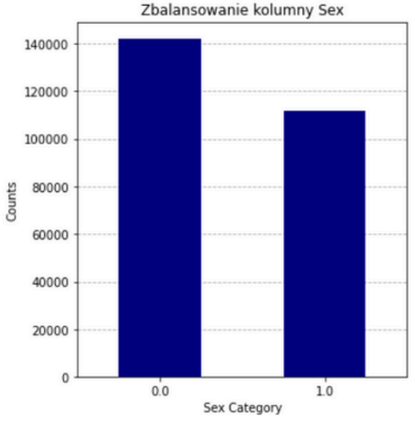

Zbalansowanie kolumny HeartDiseaseorAttack

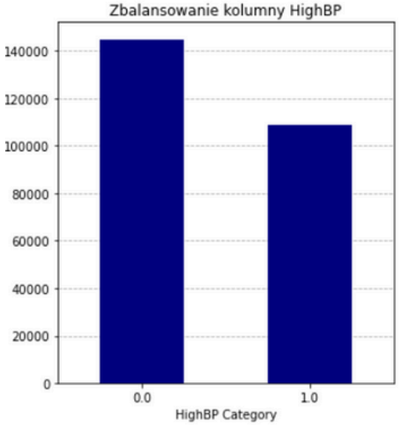

Zbalansowanie kolumny PhysActivity
Counts

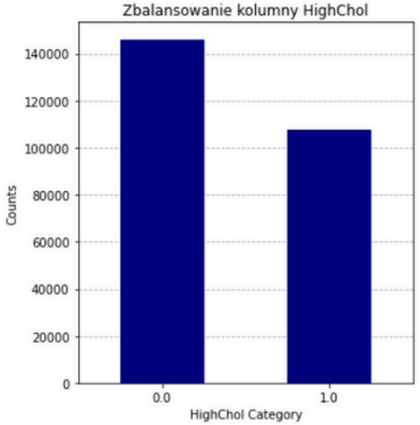

Zbalansowanie kolumny Fruits
Counts

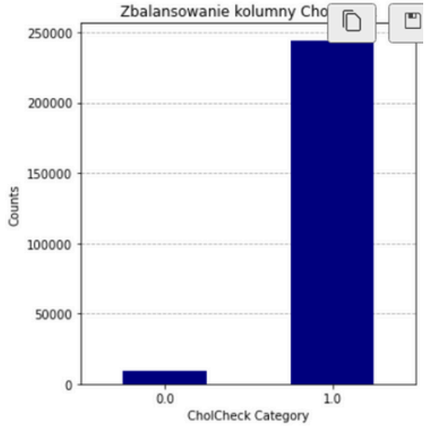

Zbalansowanie kolumny PhysHlth
Counts
PhysHlth Category

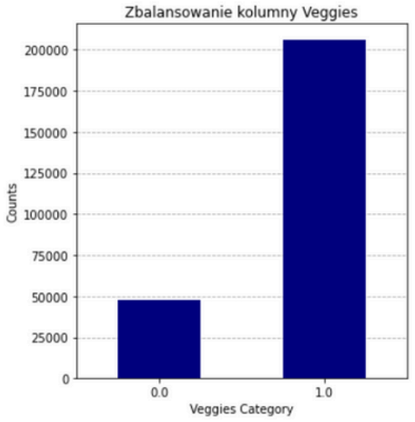

Zbalansowanie kolumny DiffWalk
Counts
DiffWalk Category

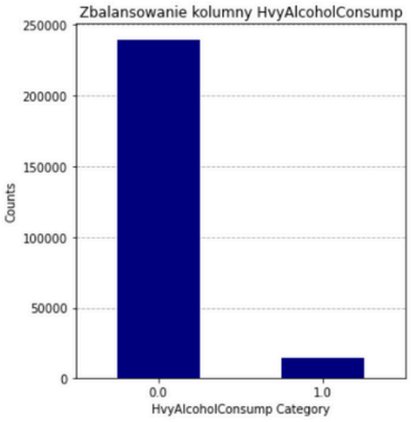

Zbalansowanie kolumny Sex
Counts
Sex Category

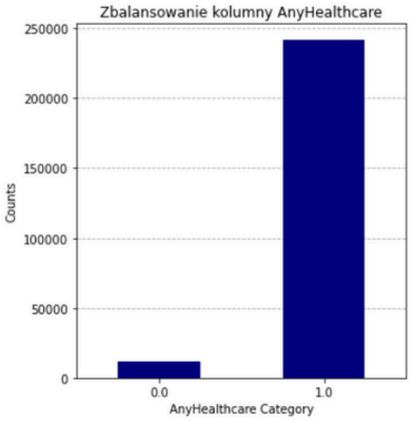

Zbalansowanie kolumny HighBP
HighBP Category

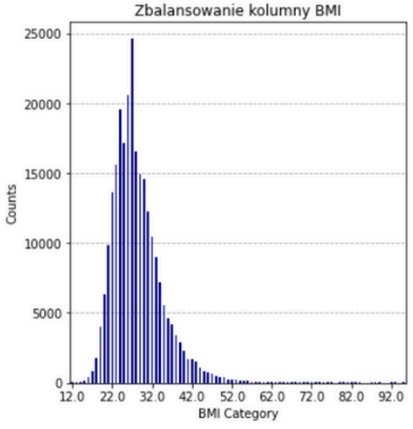

Zbalansowanie kolumny HighChol
Counts
HighChol Category

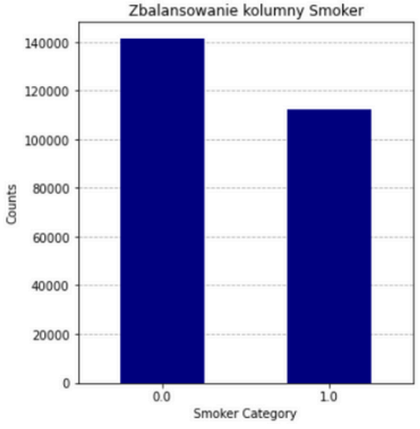

Counts
CholCheck Category

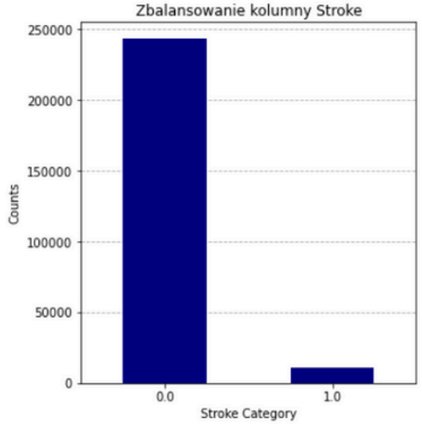

Zbalansowanie kolumny Veggies
Counts
Veggies Category


Zbalansowanie kolumny HvyAlcoholConsump
Counts
HvyAlcoholConsump Category

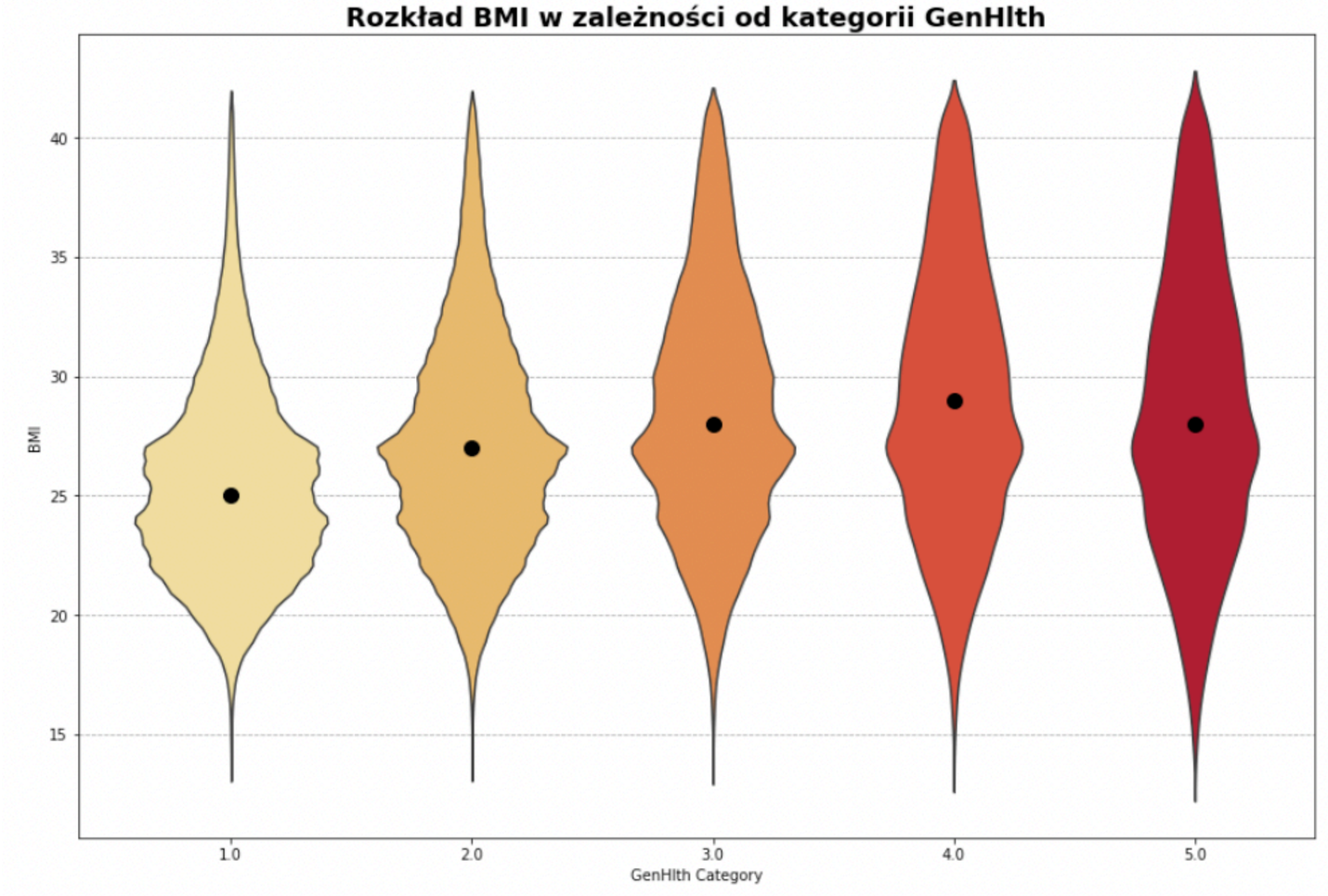

Zbalansowanie kolumny AnyHealthcare
Counts
AnyHealthcare Category


Zbalansowanie kolumny BMI
Counts
BMI Category


Zbalansowanie kolumny Smoker
Counts
Smoker Category


Zbalansowanie kolumny Stroke
Counts
Stroke Category

Exploratory Data Analysis



Prior to constructing the models, we conducted an analysis to identify correlations between various features. Plots, such as the one displayed on the slide, were utilized for this purpose.

It is evident that there is a **positive** correlation between the values in the **General Health** column and **BMI**, on average. This suggests that individuals with poorer general health tend to have higher BMIs.

Exploratory Data Analysis

Subsequent analysis yielded the following conclusions regarding relationships between features in the dataset:

- **High cholesterol** and **high blood pressure** exhibit a positive correlation with diabetes.
- **Age** demonstrates a correlation with diabetes.
- **Income** shows a positive correlation with education.
- **Difficulties in walking** are associated with general health and physical health.
- **Mental health** and **physical health** tend to be interrelated.

Increase data quality

To optimize the efficiency of training the ideal model, we conducted preprocessing on the data. This enhanced the clarity and quality of the dataset.

Outliers

We removed values that likely resulted from incorrect responses given by survey respondents because they did not align with the average values of a specific feature.

Feature engineering

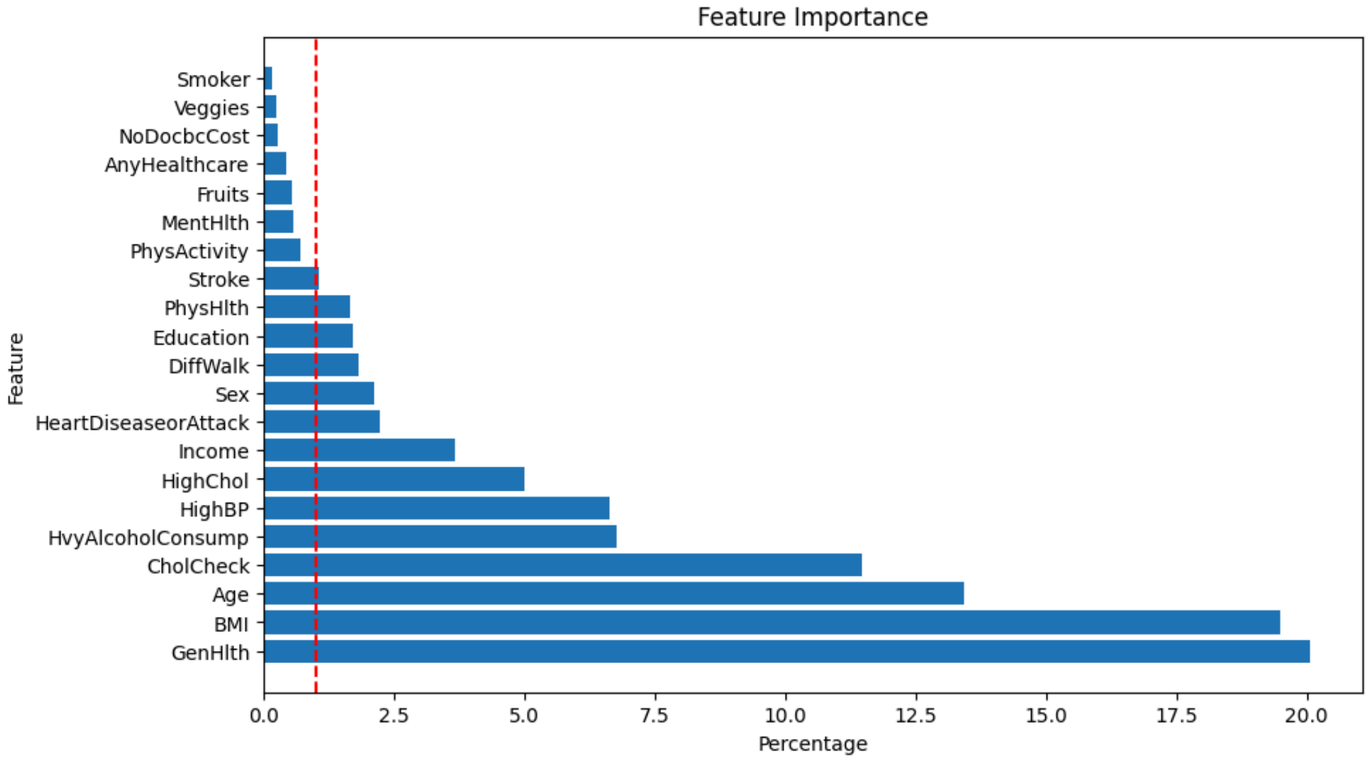
We leveraged existing features, even those with inherent faults, to engineer new ones with a higher potential to enhance model accuracy.

Scaling

We standardized all features to have values within the range [0,1], ensuring that each feature has an equal opportunity to influence the final predictions of the model.

Selecting the most promising features

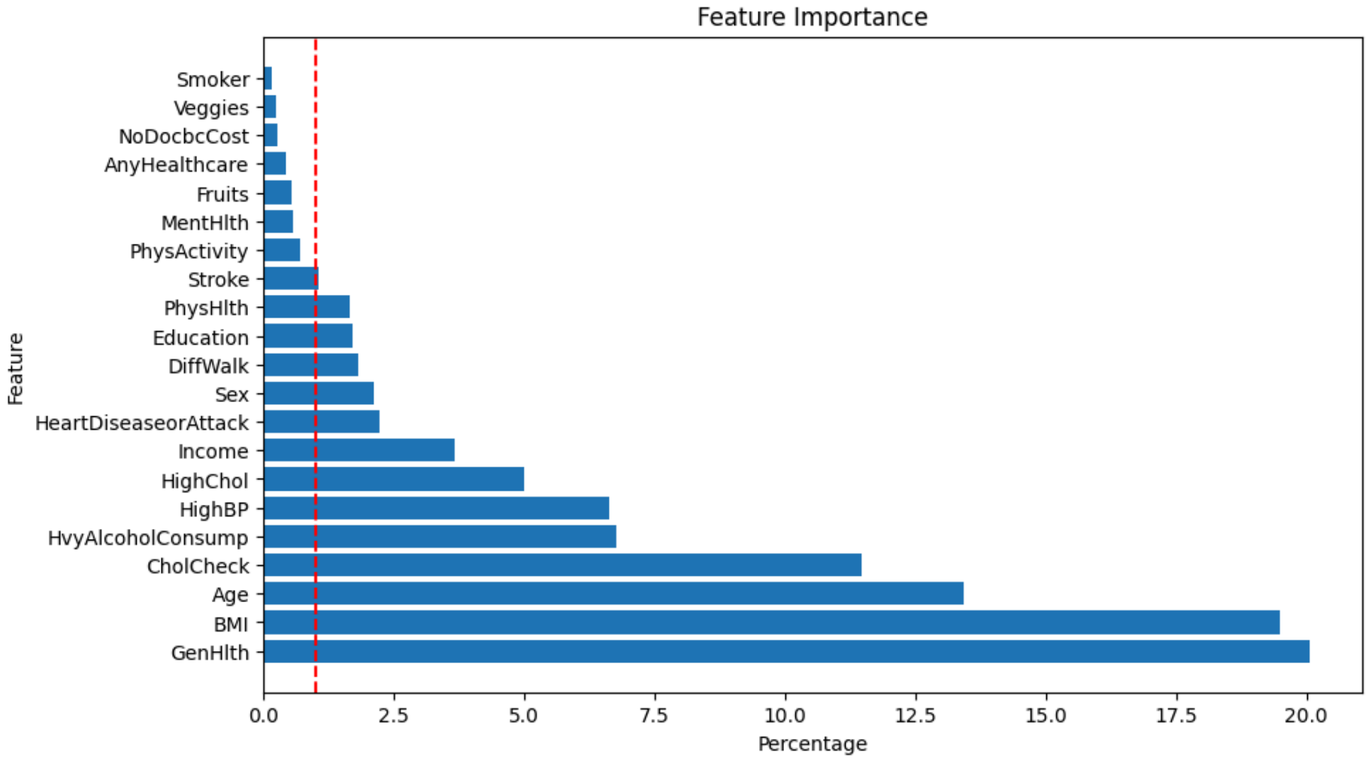
During the data preprocessing phase, we ensured that all feature coefficients were positive, setting a solid foundation for further analysis.



Utilizing L2 regularization, we effectively identified the most critical features

To confirm the significance of these features, we applied Recursive Feature Elimination (RFE), which removed the least important features, solidifying our confidence in the selected feature set.

Selecting the most promising features



Based on the plot we selected every feature with ‘Percentage’ value of over 1%. This resulted in the following set:

- **GenHlth**
- **BMI**
- **Age**
- **CholCheck**
- **…**
- **PhysHlth**
- **Stroke**

Model construction and selection

We conducted a search using ROC-AUC scoring to meticulously select the best classification algorithms, ensuring optimal performance in our predictive model.

The ROC-AUC metric guarantees that the chosen model excels in predicting the **probability** of diabetes occurrence in individual patients.

The table shows measurements for the performance of each model on train set and test set.

Consistency between the scores in the Train and Test columns ensures the reliability of the models, indicating they will perform equally well in real-world scenarios.

Model	Roc_auc Train	Roc_auc Test
gboost	0.831451	0.830284
adaBoost	0.827609	0.826557
xgb	0.824423	0.825466
Logistic Regression	0.823275	0.821773
Random Forest Classifier	0.796776	0.791453
bayes	0.787853	0.784966
KNN	0.720528	0.714200

Model construction and selection

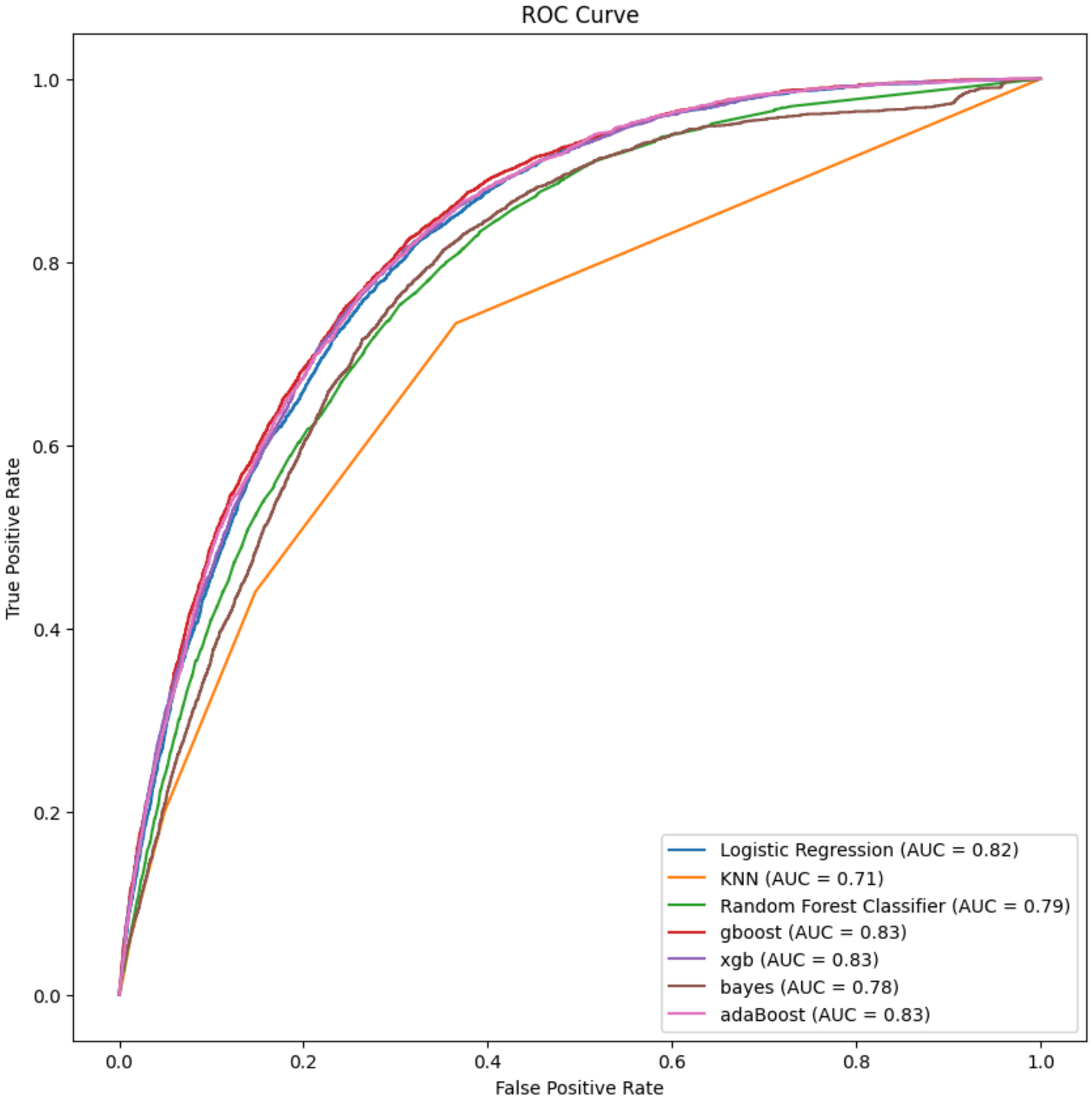
The ROC-AUC metric pertains to a specific plot known as ROC. To enhance the visualization of differences between the models presented in the table, the subsequent slide displays the ROC curves for all models.

The **interpretation** is that the greater the **area** under the curve, the more proficient the model is in **predicting probabilities** of diabetes

Also, the validation was performed by an independent team to ensure reproducibility of the results. The validation was performed for the LogisticRegression model.

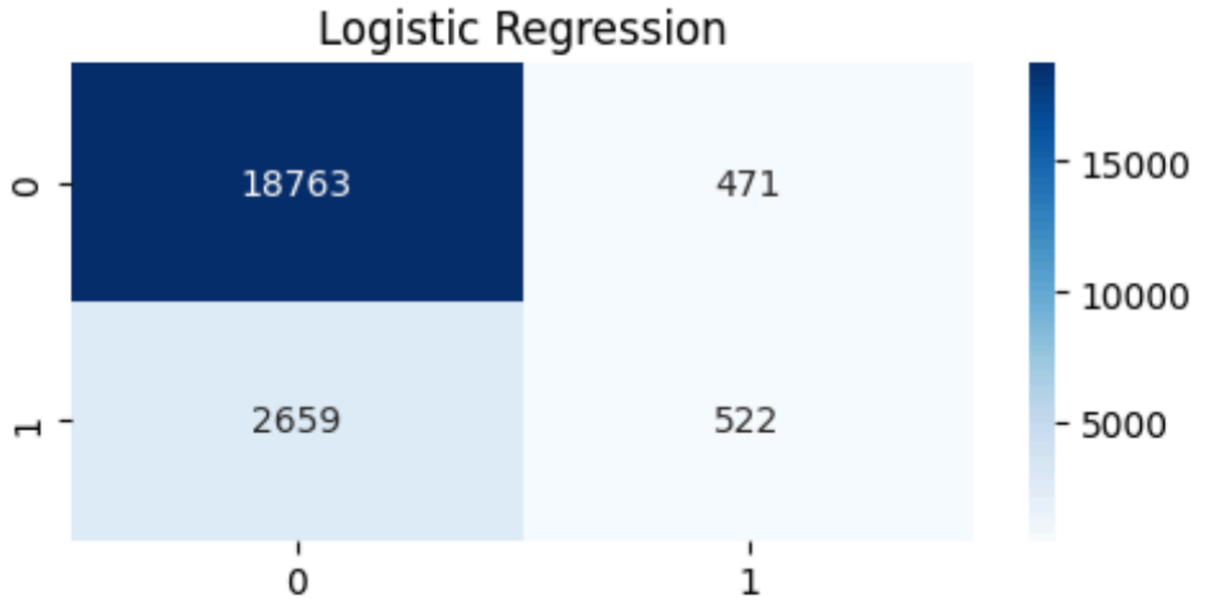
Validation roc_auc: 0.8431

Model	Roc_auc Train	Roc_auc Test
gboost	0.831451	0.830284
adaBoost	0.827609	0.826557
xgb	0.824423	0.825466
Logistic Regression	0.823275	0.821773
Random Forest Classifier	0.796776	0.791453
bayes	0.787853	0.784966
KNN	0.720528	0.714200


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.82)
KNN (AUC = 0.71)
Random Forest Classifier (AUC = 0.79)
gboost (AUC = 0.83)
xgb (AUC = 0.83)
bayes (AUC = 0.78)
adaBoost (AUC = 0.83)

Confusion matrix

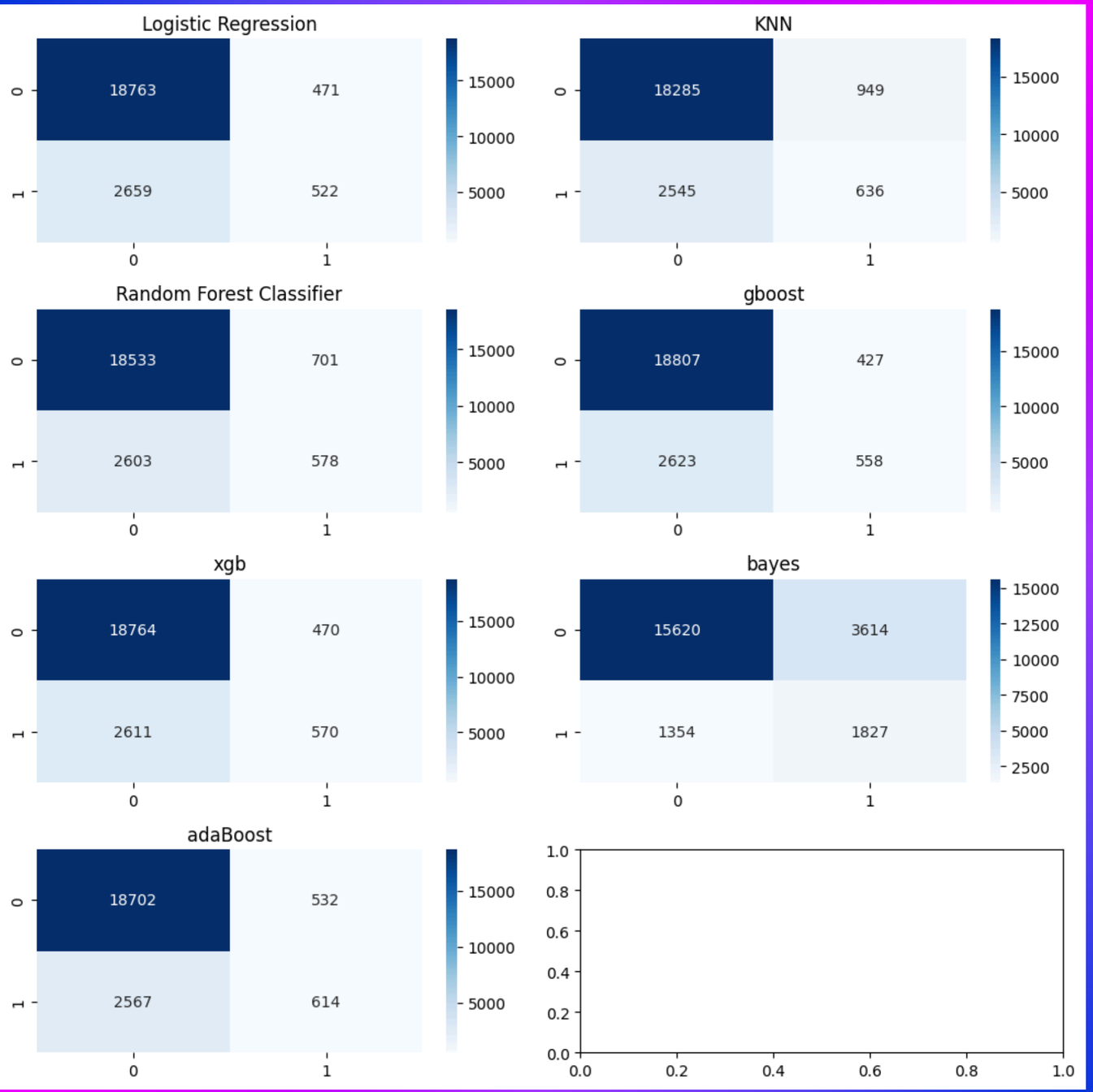
Reality: Do not have diabetes
Predicted: Do not have diabetes



Reality: Do not have diabetes
Predicted: Do have diabetes

Reality: Do have diabetes
Predicted: Do not have diabetes

Reality: Do have diabetes
Predicted: Do have diabetes

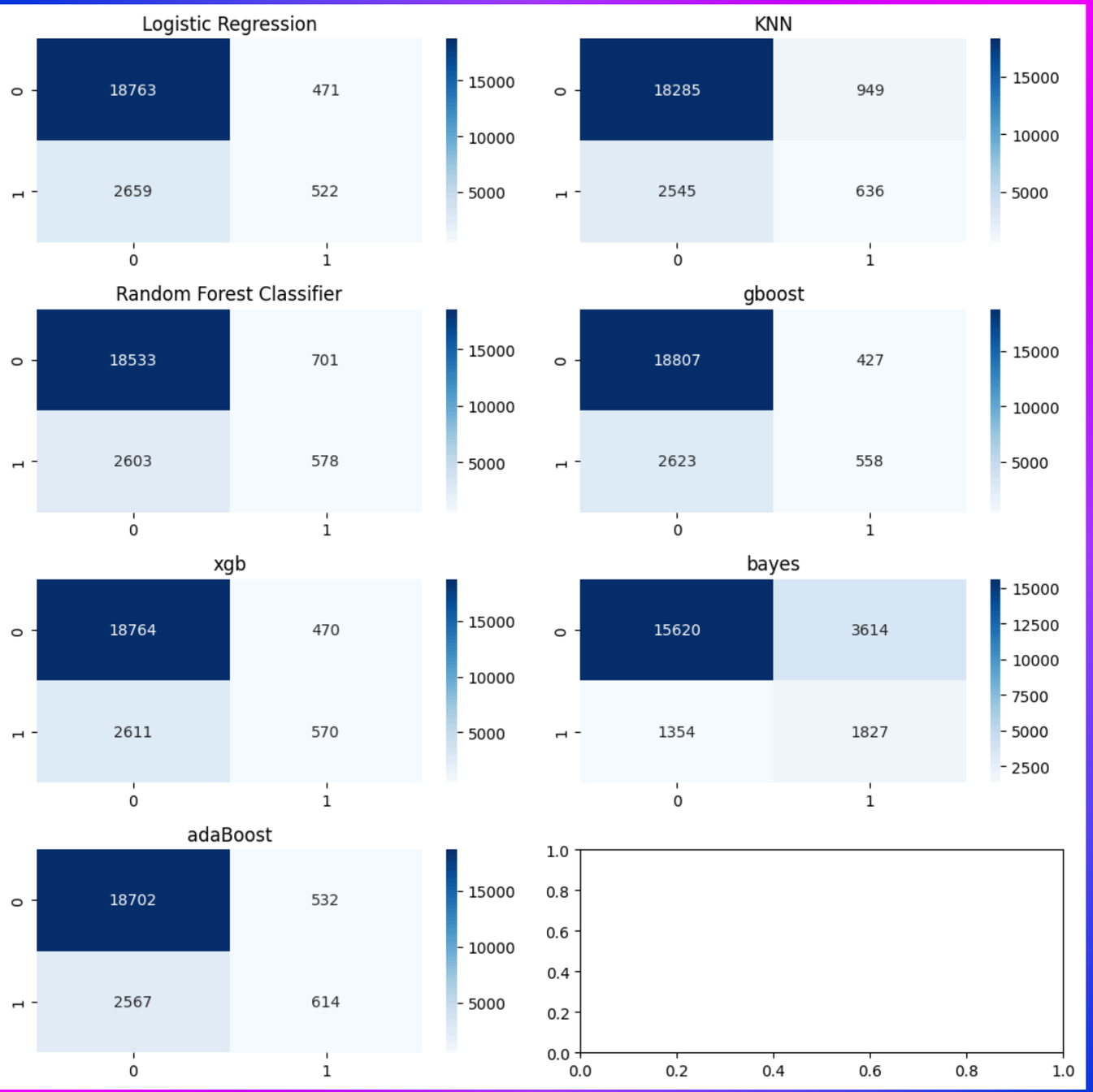


Check model performance

Logistic Regression and gboost

These two models were ultimately chosen for further development. Their confusion matrices exhibit striking similarity, notably with a relatively low percentage of healthy patients falsely classified as having diabetes.

precision = 522/933 = **56%**



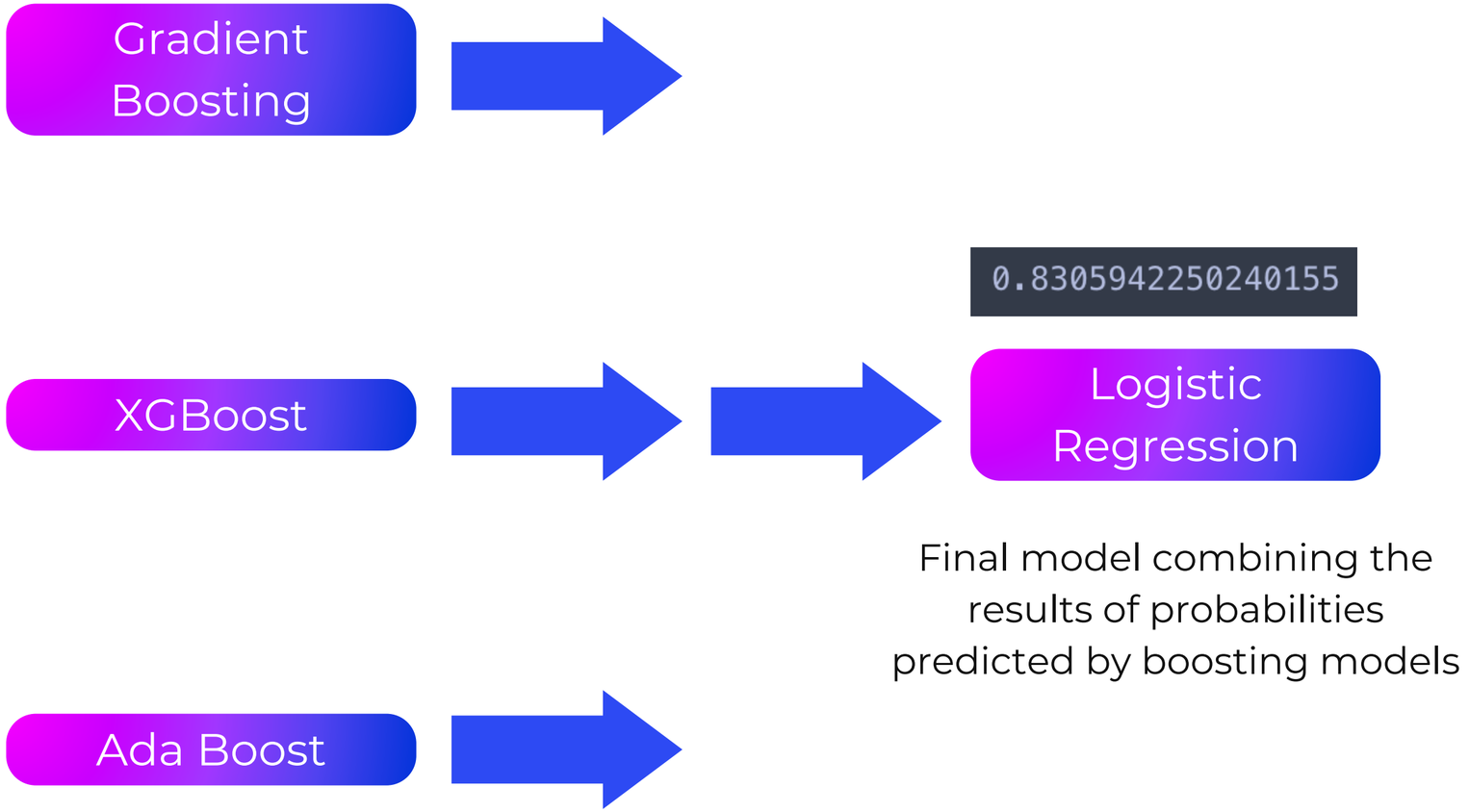
Check model performance

Bayes

While this model was not chosen for further development, it's noteworthy that it achieved a relatively high accuracy in identifying individuals with diabetes:

recall = 1827/ 3181 = **57 %**

Stacking - attempt to combine best models



Hyperparametrization

To ensure the optimal performance of the final models, we conducted a search for the best parameters for the two selected models from the previous stage. While the differences were minimal, we carefully selected the best parameters, which are outlined below.

gBoost

Best parameters:
{'subsample': 1, 'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 3, 'learning_rate': 0.1}

Best ROC/AUC score:
0.8319302780420162

logReg

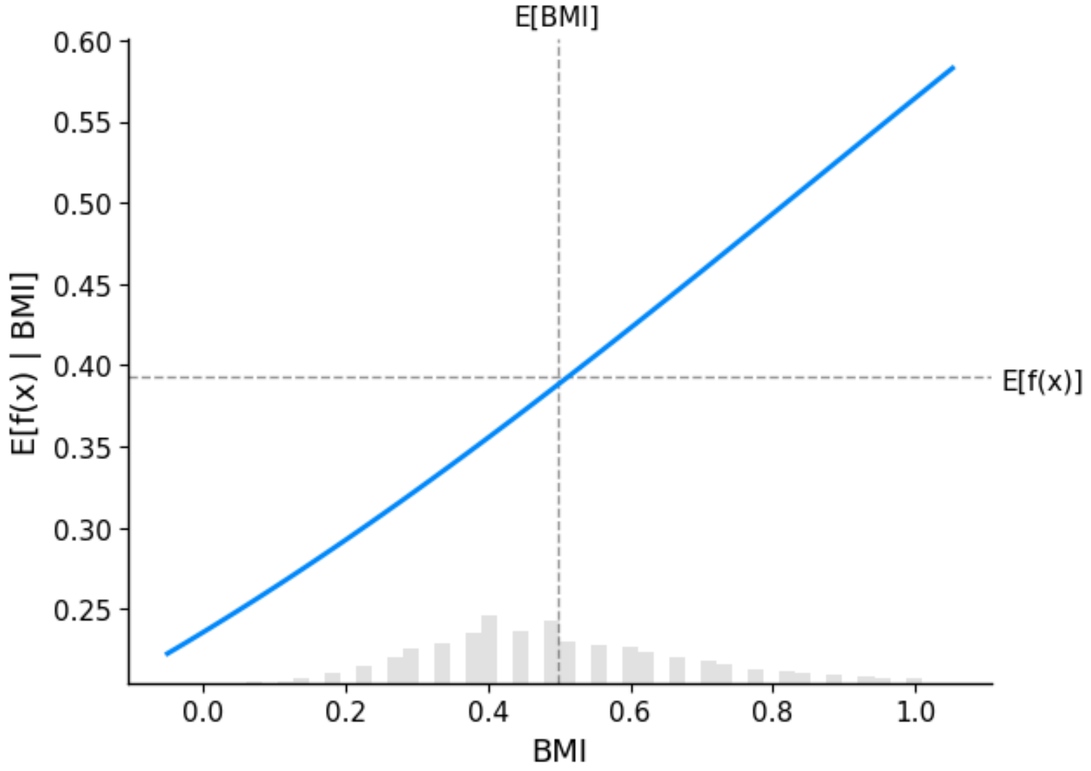
Best parameters:
{'solver': 'liblinear', 'penalty': 'l2', 'l1_ratio': 0.5, 'fit_intercept': True, 'class_weight': 'balanced', 'C': 0.1}

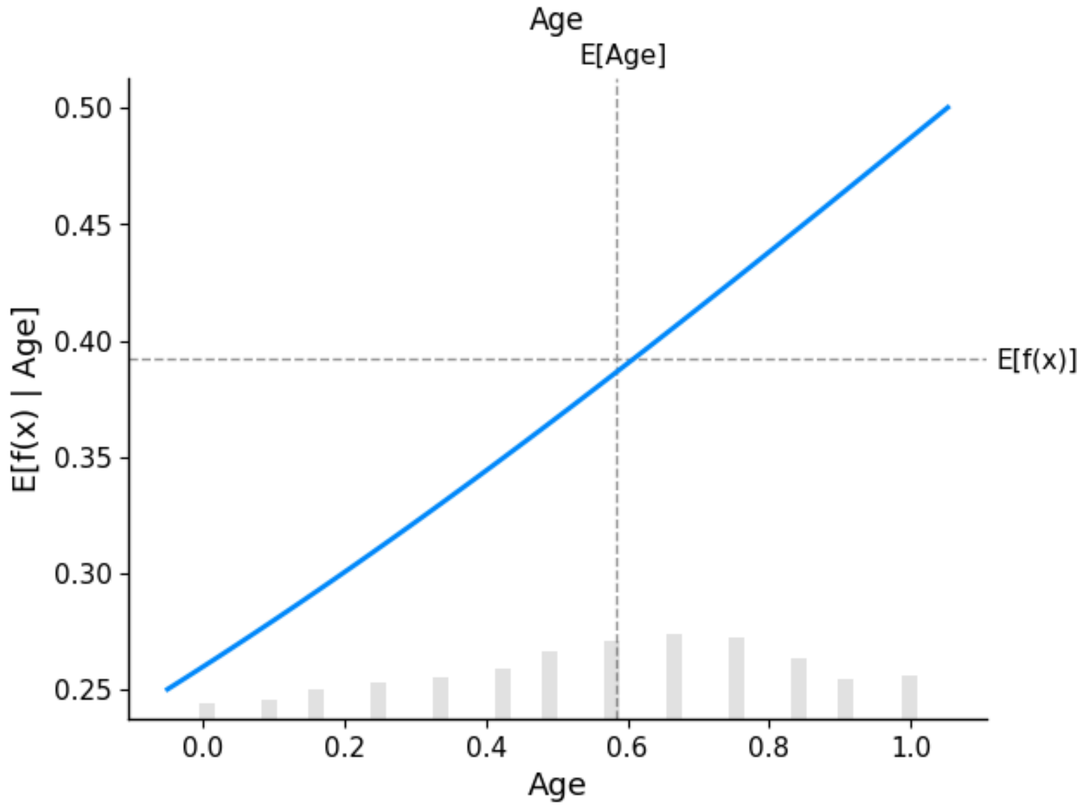
Best score: ROC/AUC
0.8235209328478534

SHAP Interpretation

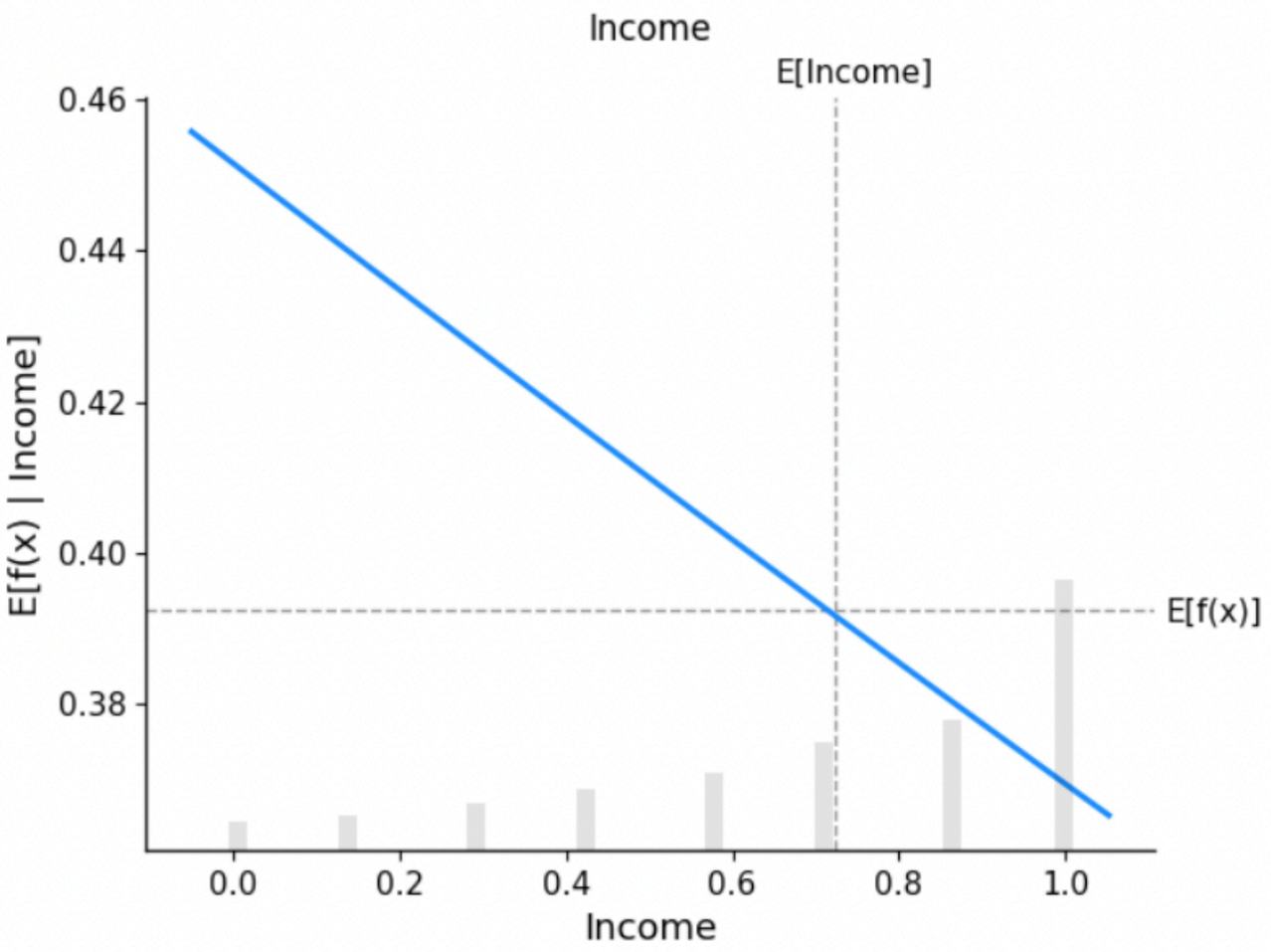
Feature impact

The partial dependence plots shown here illustrate how the predicted probability of diabetes changes as the feature of interest varies, while keeping other features fixed. This provides insight into the impact of the selected feature on final model's, thus making it easier to interpret the results.





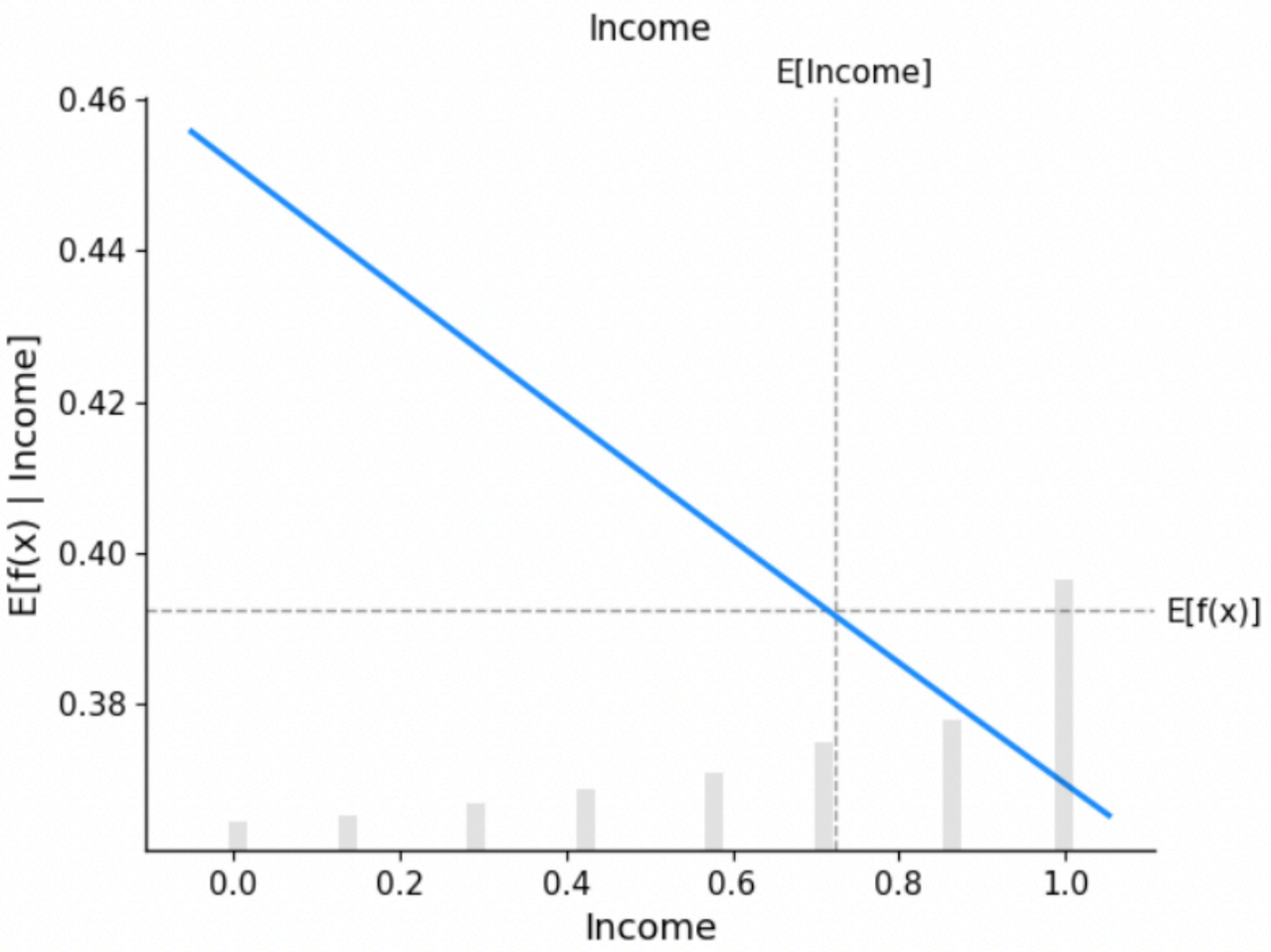
Feature impact



Key points to observe on the plot include:

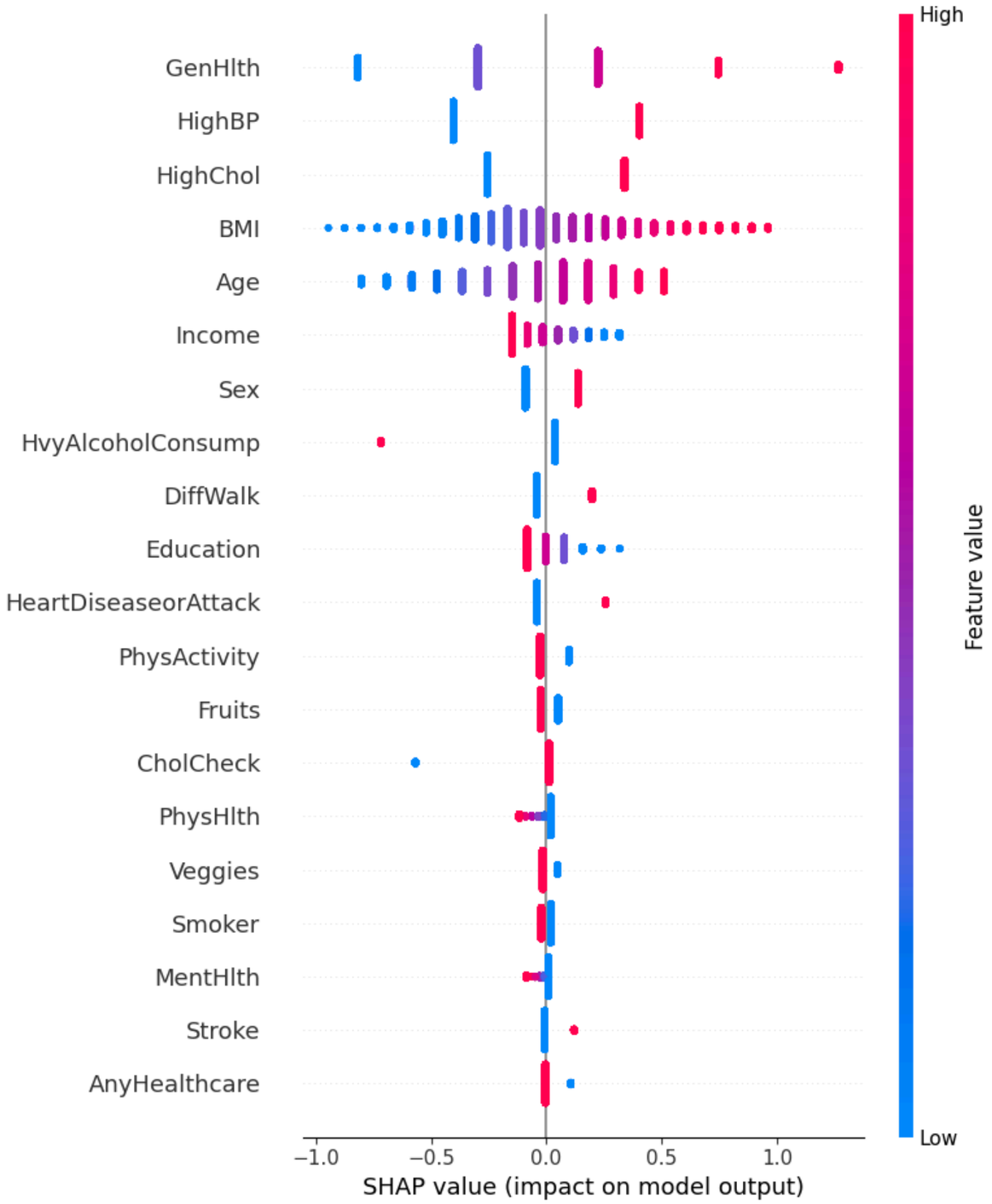
Direction of Influence: The direction of the curve indicates whether increasing the feature value generally leads to higher or lower predictions from the model. A rising curve suggests a positive relationship between the feature and the diabetes probability, while a declining curve indicates a negative relationship. For example higher Income indicates lower risk of diabetes.

Feature impact



Magnitude of Influence: A steeper slope suggests that changes in the feature value have a l**arger impact** on the model predictions.

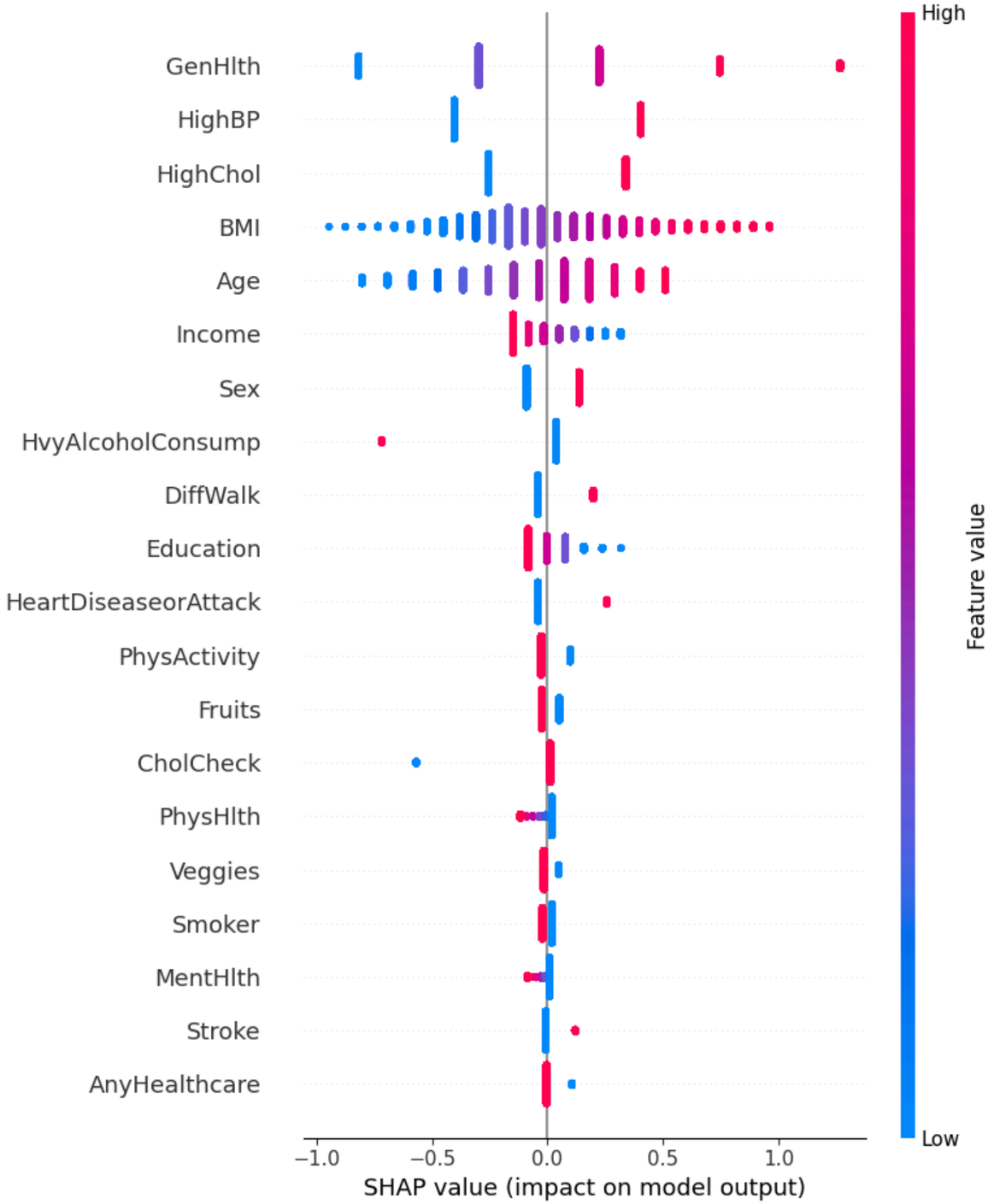
Feature Importance: The height of the curve at different points along the x-axis represents the average or median model prediction for each value of the feature. Higher values indicate greater importance of the feature in influencing the model predictions, while lower values suggest lesser importance. This leads to conclusion that **Income** has less impact on the model than **BMI or Age**.



Summary plot

SHAP summary plot gives insights into which features are driving the model predictions and how they influence individual predictions. This helps in understanding the model's decision-making process.

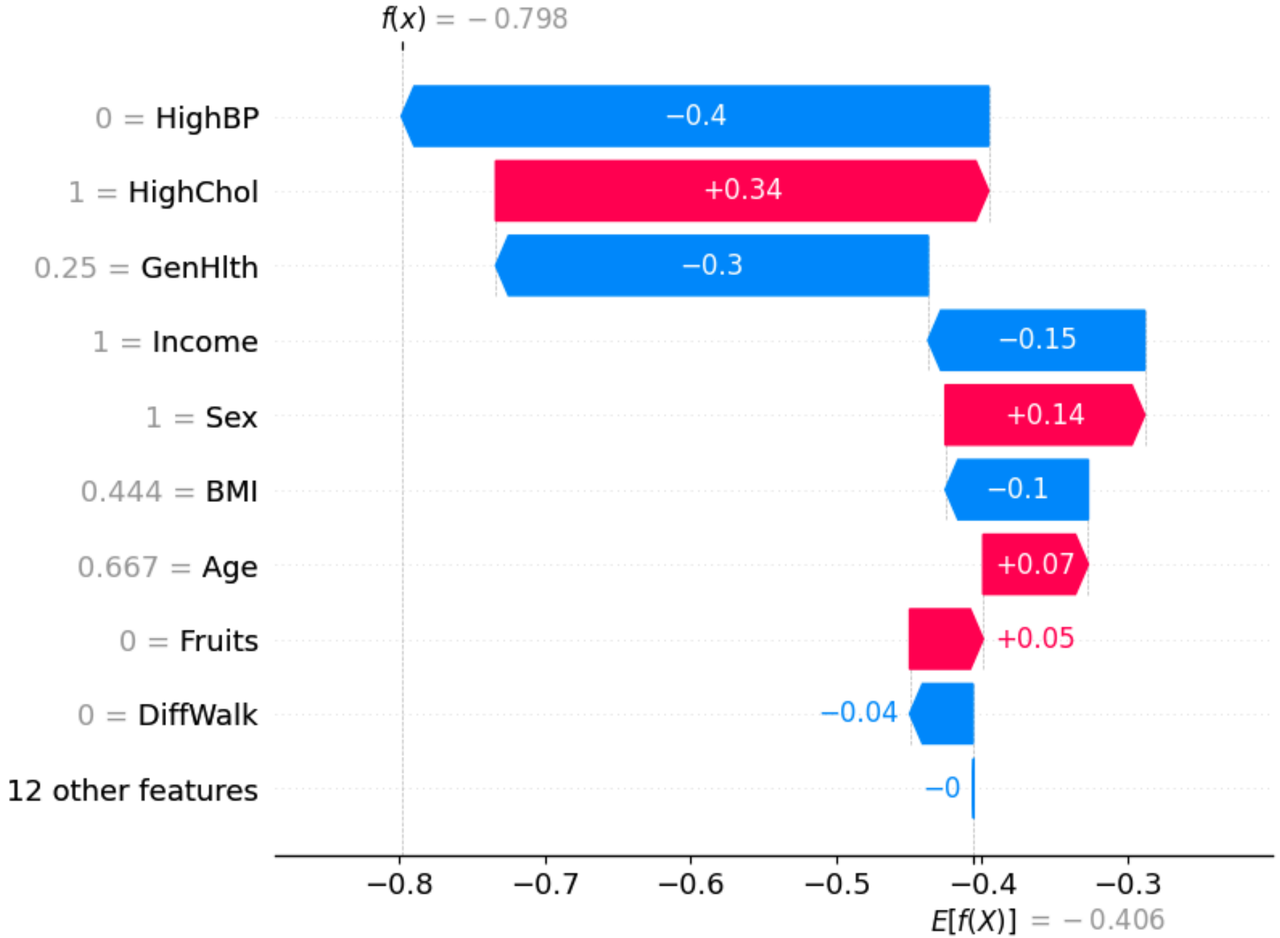
The features are listed on the y-axis, ranked in descending order based on their **importance**. The color of each feature bar represents the magnitude and direction of its impact on predictions. Features associated with positive SHAP values contribute to increasing the prediction, while those with negative SHAP values contribute to decreasing the prediction.



Summary plot

- **General Health:** Demonstrates a strong impact on predictions, with extreme values exerting the most influence on the model.
- **High Blood Pressure:** High blood pressure significantly increases the risk of diabetes, whereas low blood pressure substantially reduces this risk.

Explaining model’s predicitons



The waterfall plot enables the medic to examine which **specific features** influenced the probability of diabetes diagnosis, and which did not. Here, we observe that the individual's low blood pressure contributed to a decrease in the probability of diabetes diagnosis.

Final takeaways

Feature importances

The most significant predictors for diabetes include general health, age, BMI, blood pressure, cholesterol level, and notably, income."

Number of features

The number of available features for the model was relatively small. We noticed that reducing the number of features improved the recall score, but it led to a decrease in the ROC-AUC score, which was the primary metric used in the project.

Scoring methods

Assessing whether the final models are favorable or unfavorable depends on the chosen metric. Given our models' high ROC-AUC scores, they fulfill our objective of creating models capable of accurately predicting probabilities.

Further improvements

There is potential to enhance model performance by employing deep learning techniques, such as deep neural networks. Although complex, these methods have the capability to yield superior results.

Resources and technologies used during the project

scikit-learn.org/stable/tutorial/basic/tutorial.html

Lecture materials from Warsaw University of Technology

https://www.kaggle.com/code/chanchal24/diabetes-dataset-eda-prediction

https://www.cdc.gov/brfss/annual_data/annual_data.htm

Python, NumPy, pandas, Matplotlib, scikit-learn, XGBoost, AdaBoost

Thank you for reading the report

Project repository

https://github.com/Iwaniukooo11/ml-diabetes-prediction