Università di Bologna

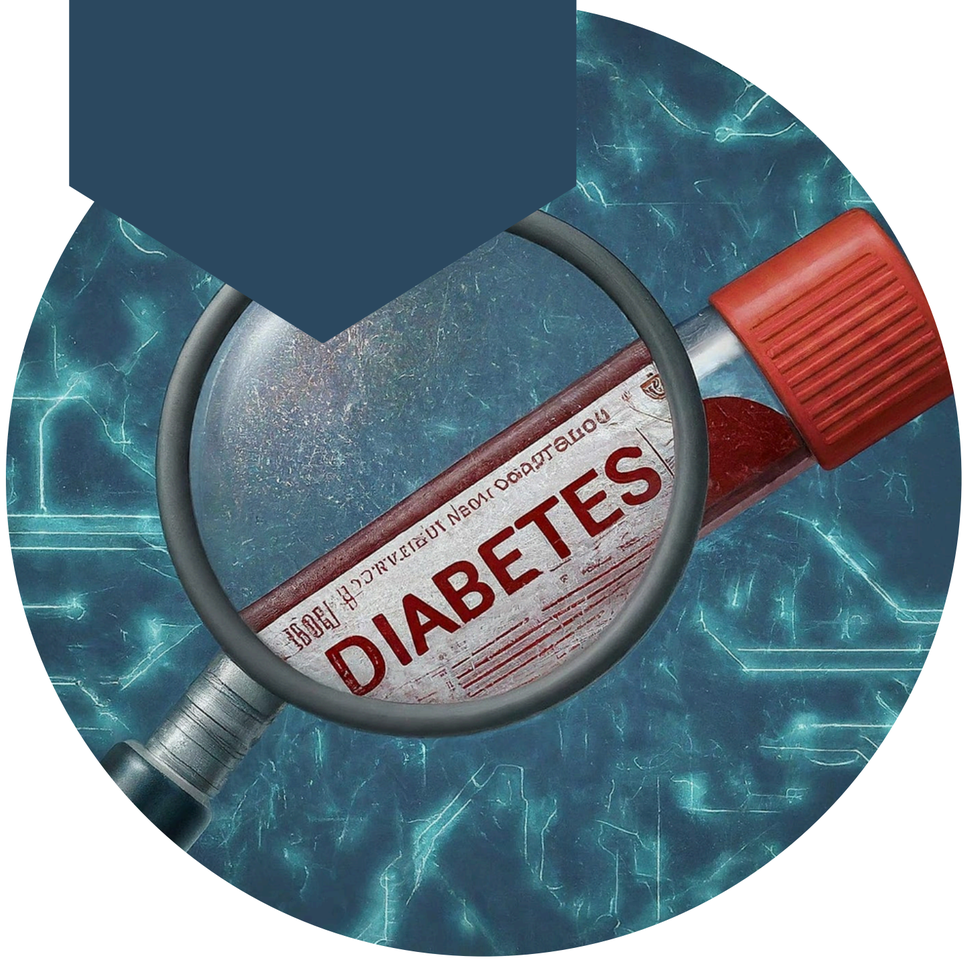
BIG DATA ANALYTICS

Exploratory Data Analysis and
Predictive Modeling for Diabetes Risk Classification

January 22, 2025

Project by José Ribeiro & Sofia Costa
1900130105 & 1900129396

Table of Contents



Problem Description

Problem Statement

Prevalence of diabetes is a major public health issue. This project analyzes factors influencing diabetes risk using the CDC Diabetes Health Indicators Dataset.

Objective

To identify key factors influencing diabetes prevalence and develop a predictive model for classification.

Dataset

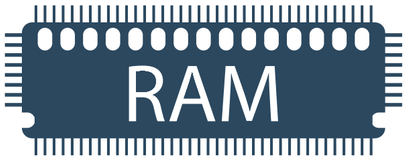
Kaggle's Diabetes Binary Health Indicators BRFSS 2015

Experimental Setting

Vagrant

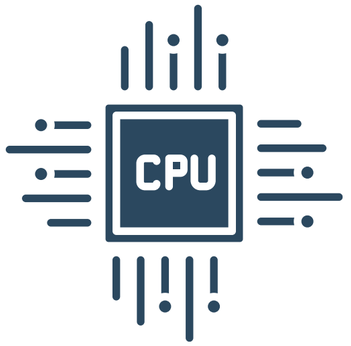
Number of VMs

1 VM
Ubuntu 22.04 LTS
(Jammy Jellyfish)



RAM Size

4GB



Processor

2 CPU Cores

Experimental Setting

Tools Used	VSCode, Jupyter Notebooks, GitHub, VirtualBox, Kaggle, ChatGPT
Software	Python 3.11.5
Libraries	Pandas, Scikit-learn, Seaborn, Matplotlib
Dataset Size	[229474 rows x 22 columns]
# of ML Algorithms Used	6

Dataset Description

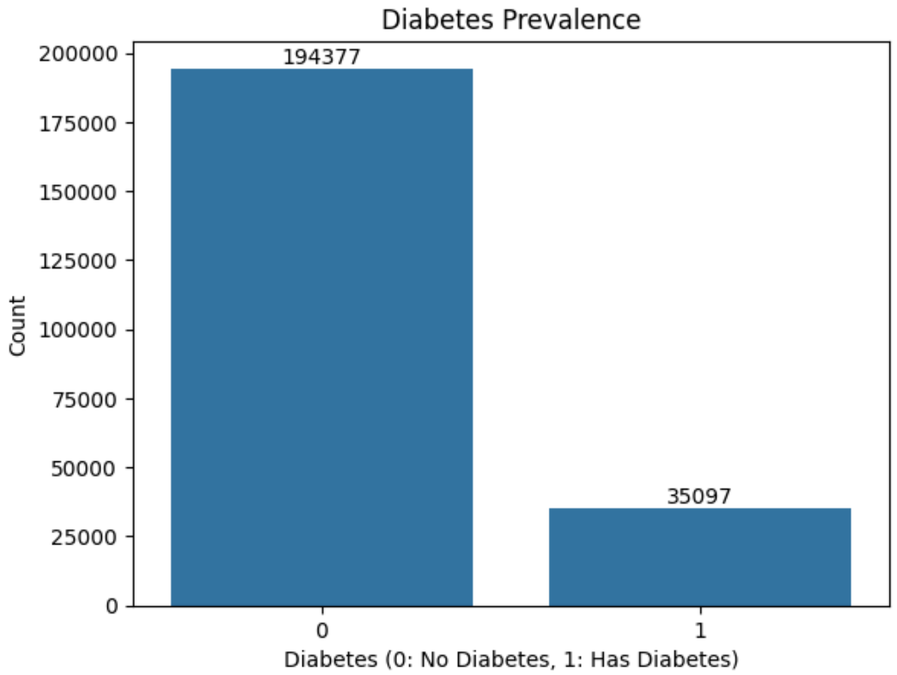
Feature	Description	Values
Diabetes	Diabetes diagnostic	0 = no 1 = yes
HighBP	High Blood Pressure	0 = no 1 = yes
HighChol	High Cholesterol	0 = no high cholesterol 1 = high cholesterol
CholCheck	Cholestrol check in 5 years	0 = no 1 = yes
BMI	Body Mass Index	{non_binary values}
Smoker	Smoked at least 100 cigarettes in your entire life	0 = no 1 = yes
Stroke	Ever had a stroke	0 = no 1 = yes
HeartDiseaseorAttack	Coronary Heart Disease (CHD) or Myocardial Infarction (MI)	0 = no 1 = yes
PhysActivity	Physical Activity in past 30 days	0 = no 1 = yes
Fruits	Consume Fruit 1 or more times per day	0 = no 1 = yes
Veggies	Consume Vegetables 1 or more times per day	0 = no 1 = yes
HvyAlcoholConsump	Adult Men >= 14 drinks per week Adult Women >= 7 drinks per week	0 = no 1 = yes

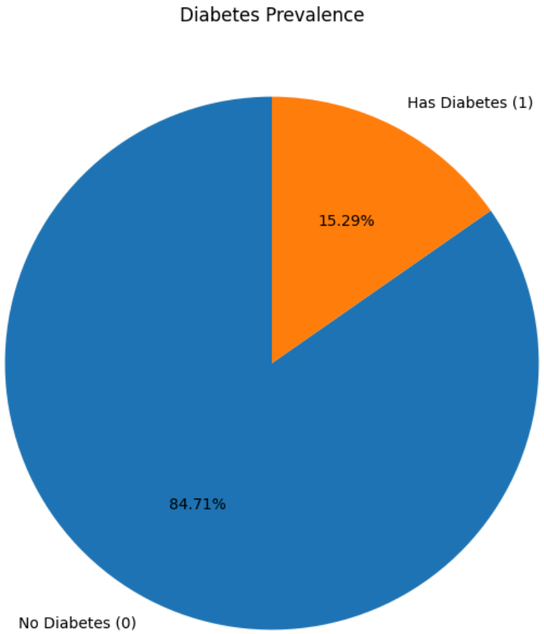
Dataset Description

Feature	Description	Values
AnyHealthcare	Any kind of health care coverage (including health insurance, prepaid plans such as HMO, etc.)	0 = no 1 = yes
NoDocbcCost	Need to see a doctor in the past 12 months but couldn't because of the cost	0 = no 1 = yes
GenHlth	General Health Status (1-5 levels)	1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
MentHlth	Days of poor mental health (scale 1-30 days)	1-30
PhysHlth	Physical illness or injury days in past 30 days	1-30
DiffWalk	Serious difficulty walking or climbing stairs	0 = no 1 = yes
Sex	Gender	0 = female 1 = male
Age	13-level age category	1 = 18-24, 9 = 60-64, 13 = 80 or older
Education	Education level	scale 1-6: 1 = Never attended school or only kindergarten 2 = elementary etc.
Income	Income scale	scale 1-8: 1 = less than 10,000 ; 5 = less than10,000 ; 5 = less than 35,000 ; 8 = $75,000 or more

Exploratory Data Analysis (EDA)

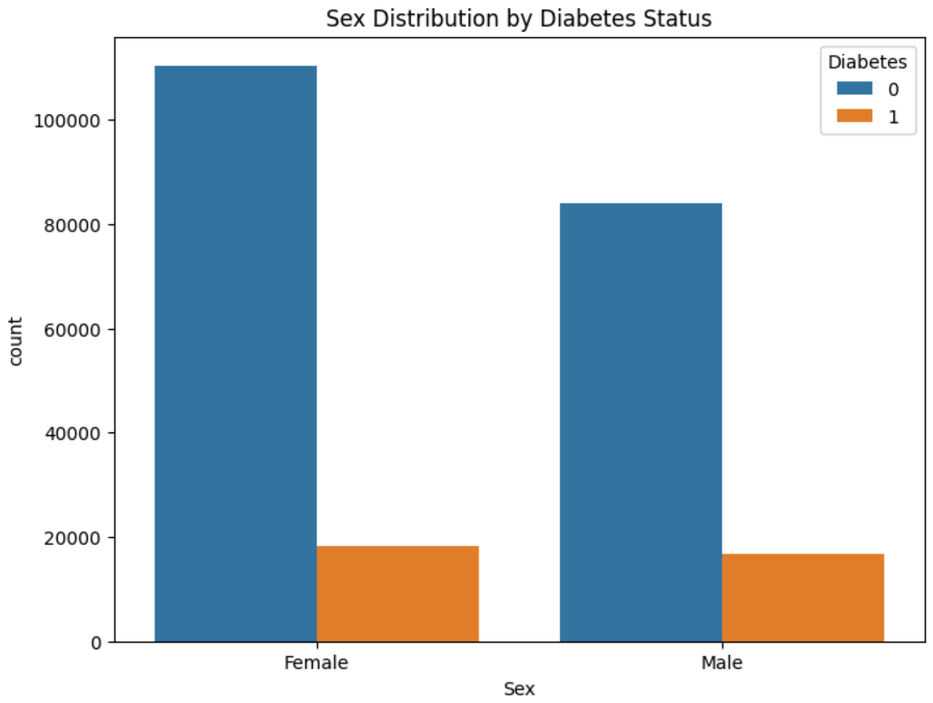
This dataset consists of a total of **229474** entries (after cleaning) with **35097** instances of the positive class (Diabetes), which consists of a **15.29%** ratio, making this dataset **imbalanced.**

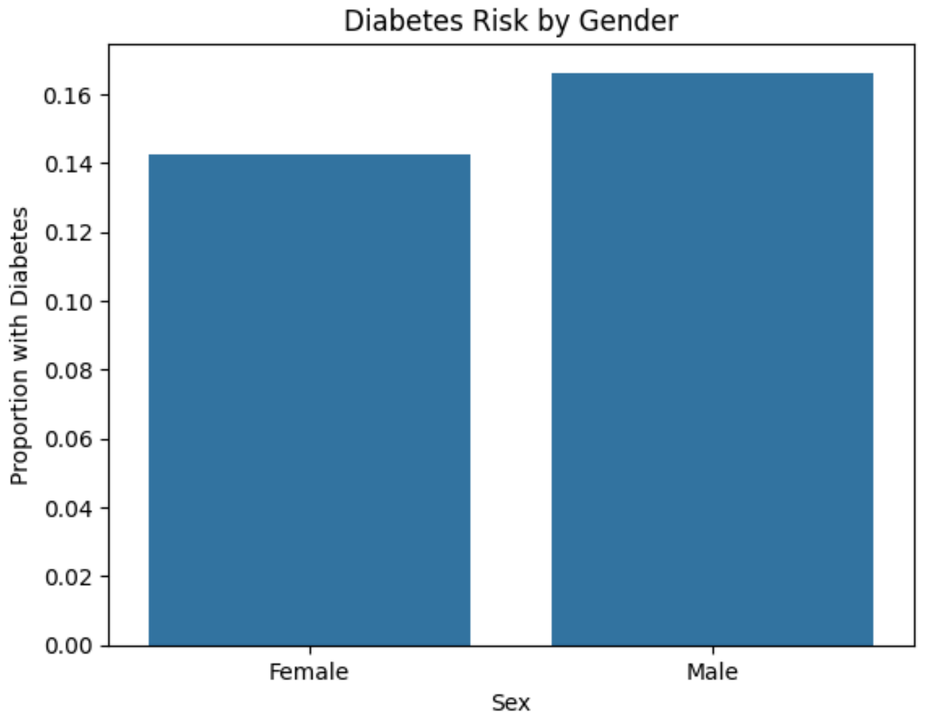




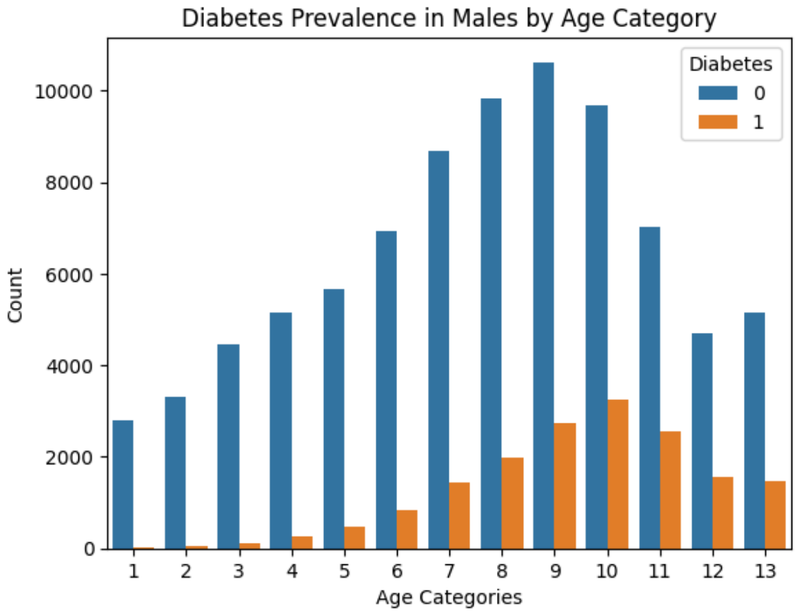
Exploratory Data Analysis (EDA)

There are proportionaly more men with Diabetes than women in the dataset, despite there being less men in the dataset.

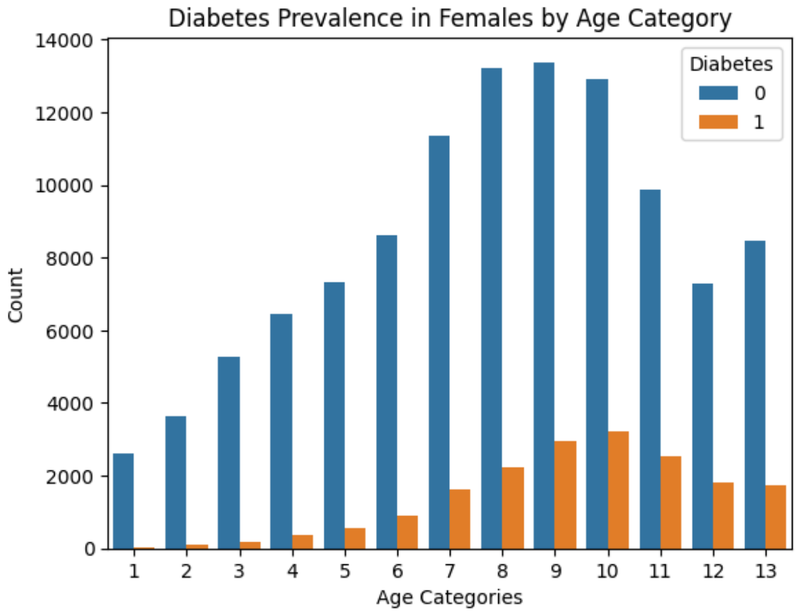




Exploratory Data Analysis (EDA)



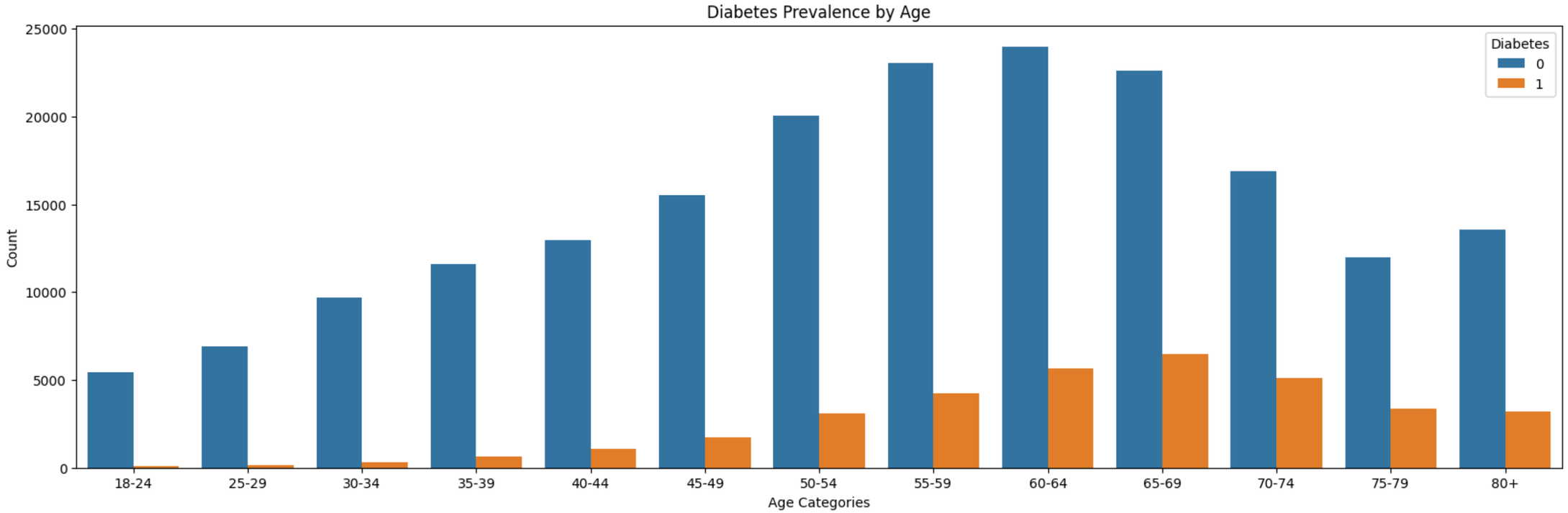
Age category in **men** with the highest number of people with diabetes: 10 (65-69 years old) with a total of 3263 people.



Age category in **women** ith the highest number of people with diabetes: 10 (65-69 years old) with a total of 3220 people.

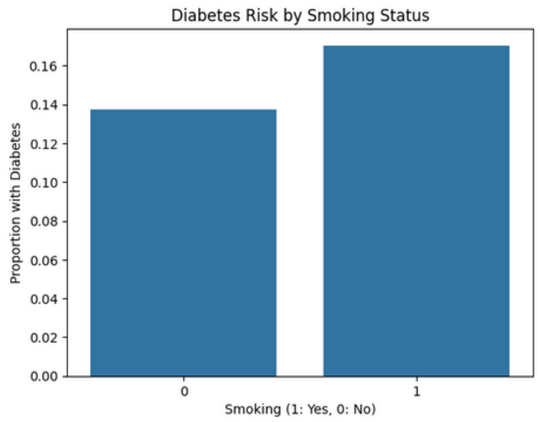
Exploratory Data Analysis (EDA)

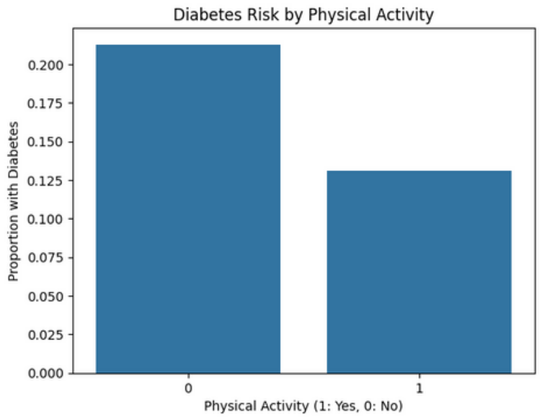
- The Age category with the highest number of people with diabetes is **10 (65-69 years old)** with a total of **6483** people, although the age group with the highest percentage of people with diabetes is **11 (70-74 years old)** with **23%,** while the lowest percentage age group is the 1st one **(18-24)**, with just **1.42%.**

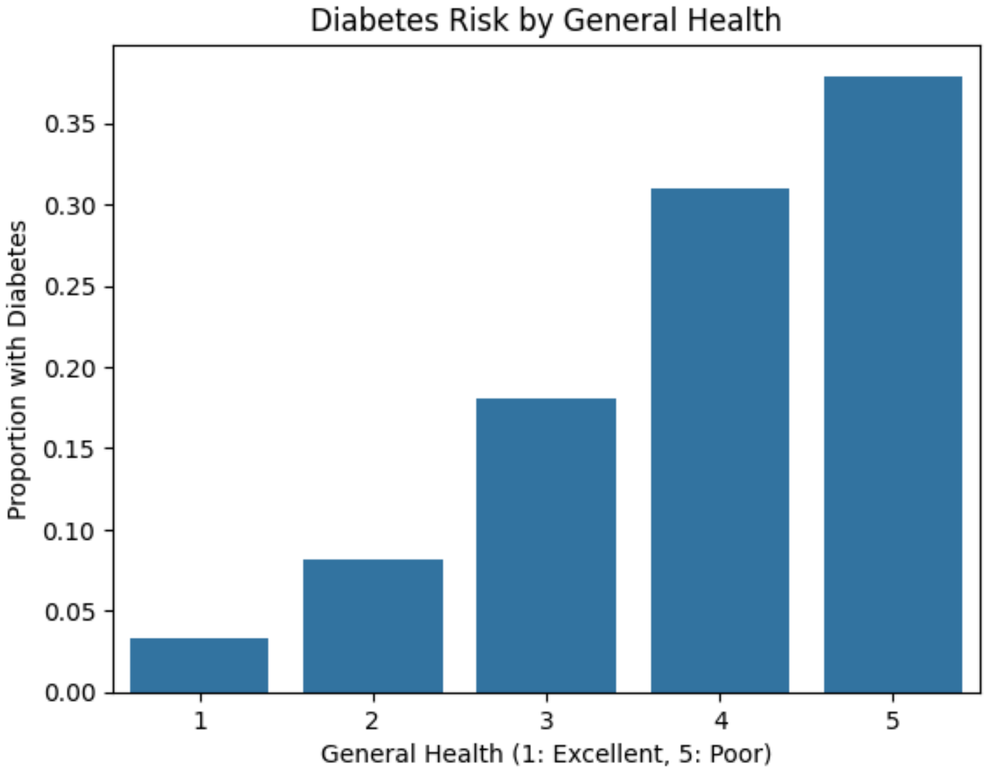


Exploratory Data Analysis (EDA)

Doing a basic analysis of 2 of the most associated risk factors of diabetes, does not show a clear discrepancy (<0.1 ≠); contrary to this, the **General Health** indicator showed a clearer ≠ between the best (1) and worst possible mark, 5, by more than **30%.**



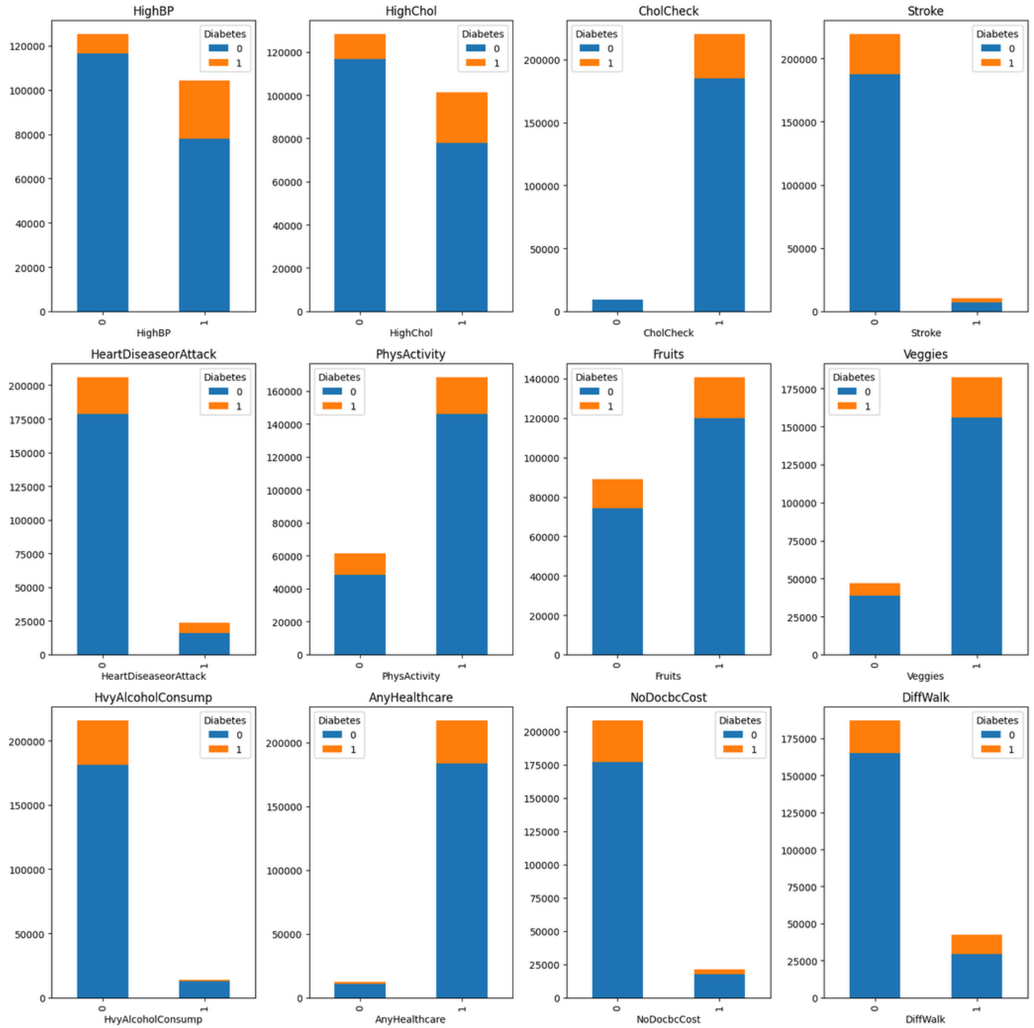




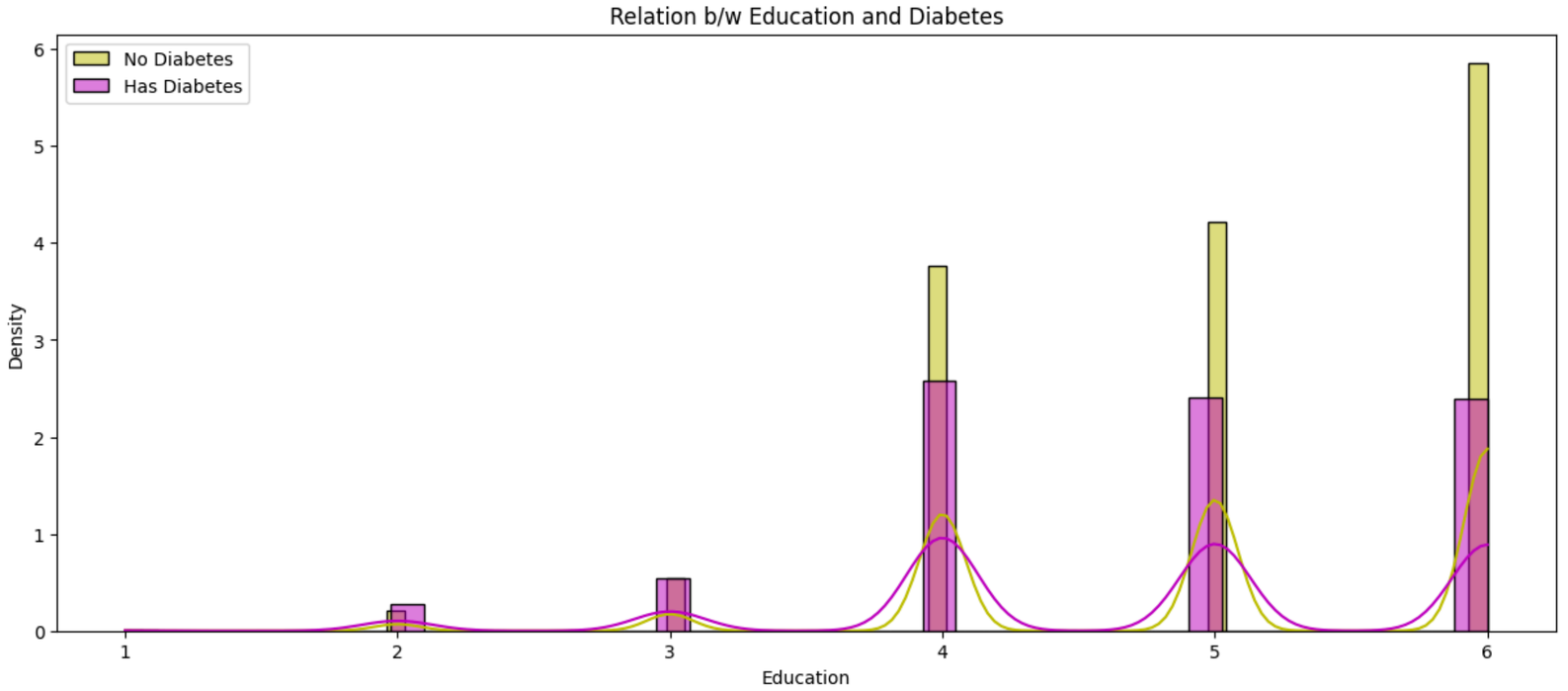
Exploratory Data Analysis(EDA)

Stacked Bar Chart Analysis of Binary Columns vs. Diabetes

- **Key Risk Factors:**
  - High blood pressure
  - High cholesterol
  - Lack of physical activity
- **Lifestyle Factors:**
  - Healthy habits like consuming fruits, vegetables, and engaging in physical activity seem to reduce diabetes risk.
- **Healthcare Access:**
  - Disparities in healthcare access might play a role in diabetes management and diagnosis.



Exploratory Data Analysis (EDA)

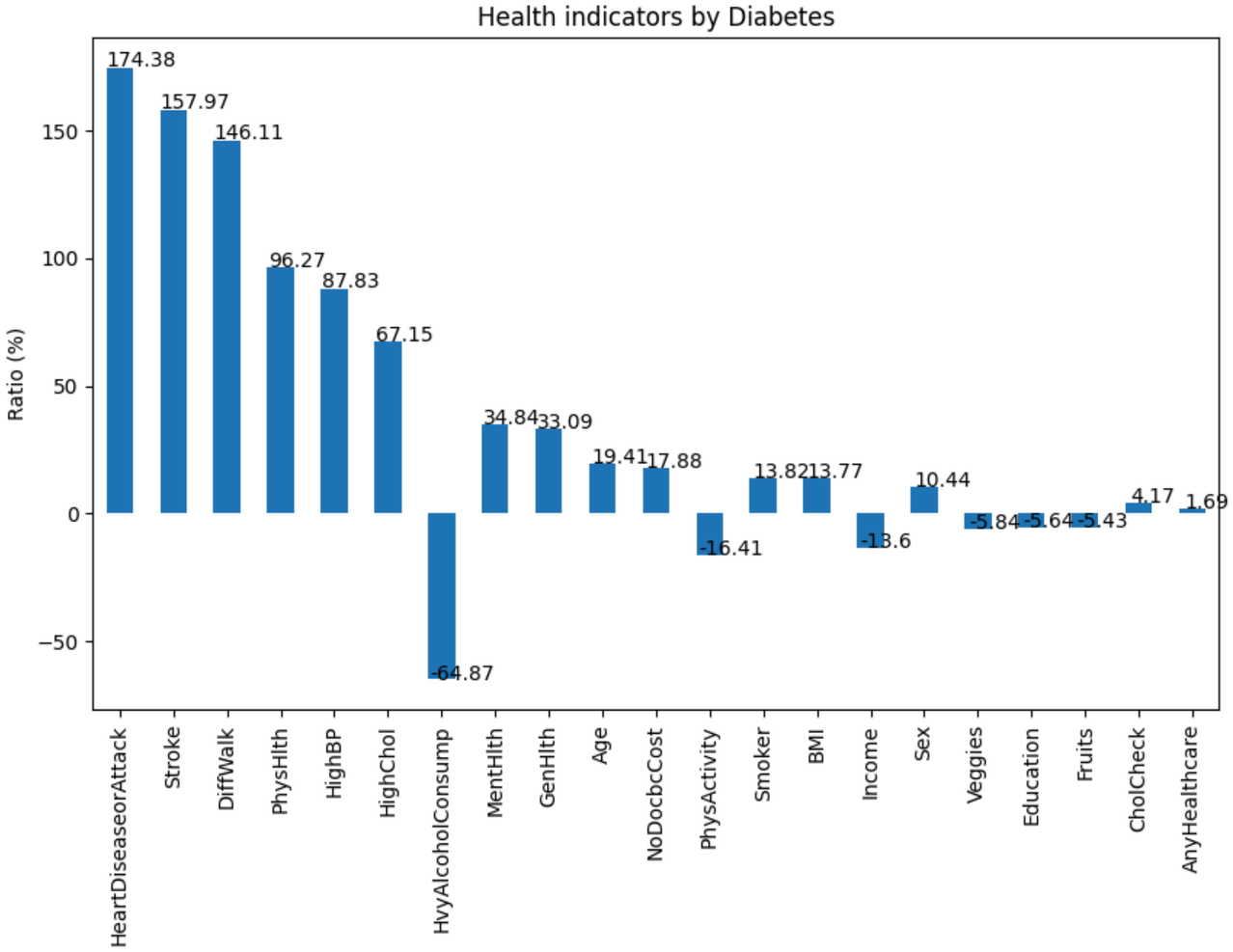


- There are more people with higher levels of education.
- There are more people without diabetes who have higher levels of education.

Exploratory Data Analysis (EDA)

One interesting plot was to compare the avg. of values of the indicators with and without diabetes, that showed great difference (50%>) in values for the following features:

HeartDiseaseorAttack, Stroke, DiffWalk, PhysHlth, HighBP, HighChol and HvyAlcoholConsump.

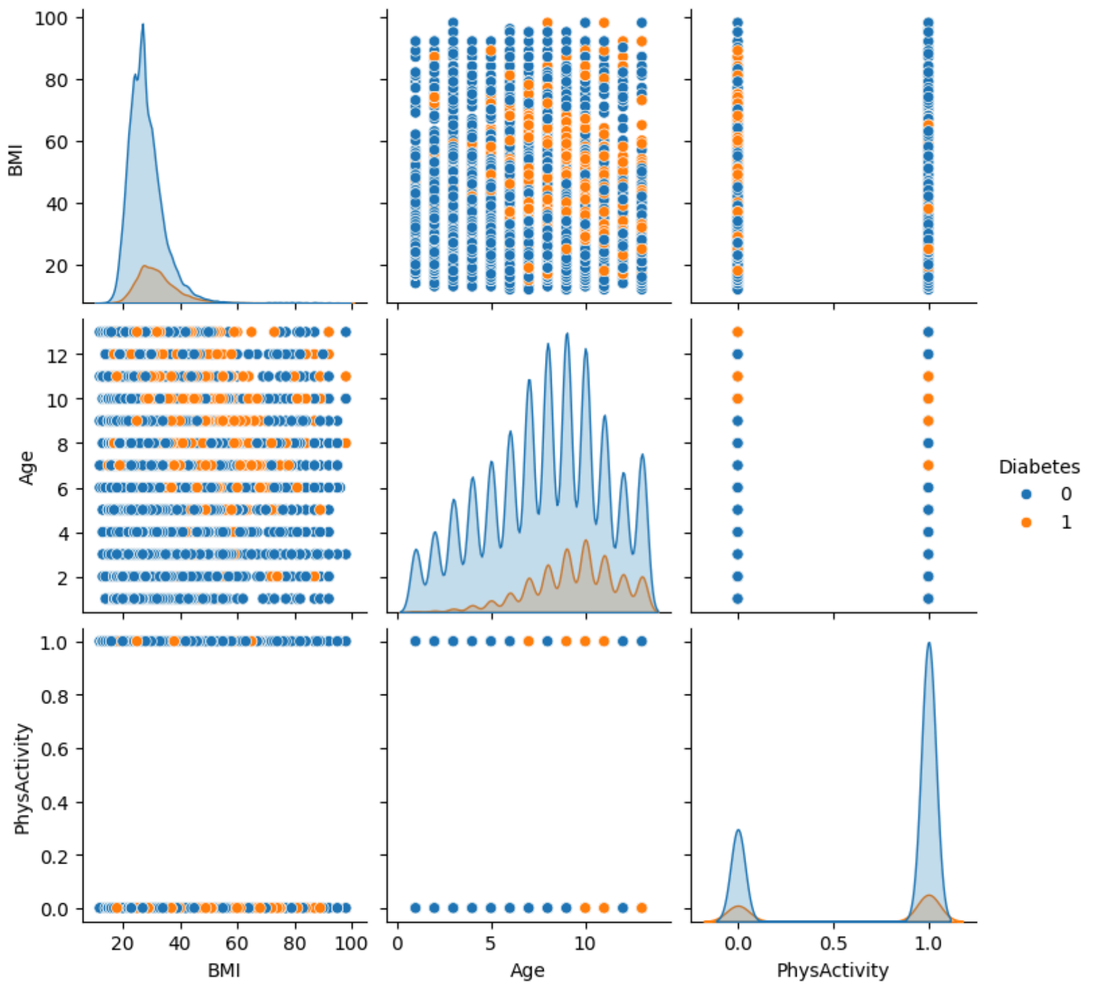


Exploratory Data Analysis (EDA)

Feature correlation

Analyzing 3 quantitative variables, namely **BMI**, **Age** and **PhysActivity**, it can be observed that:

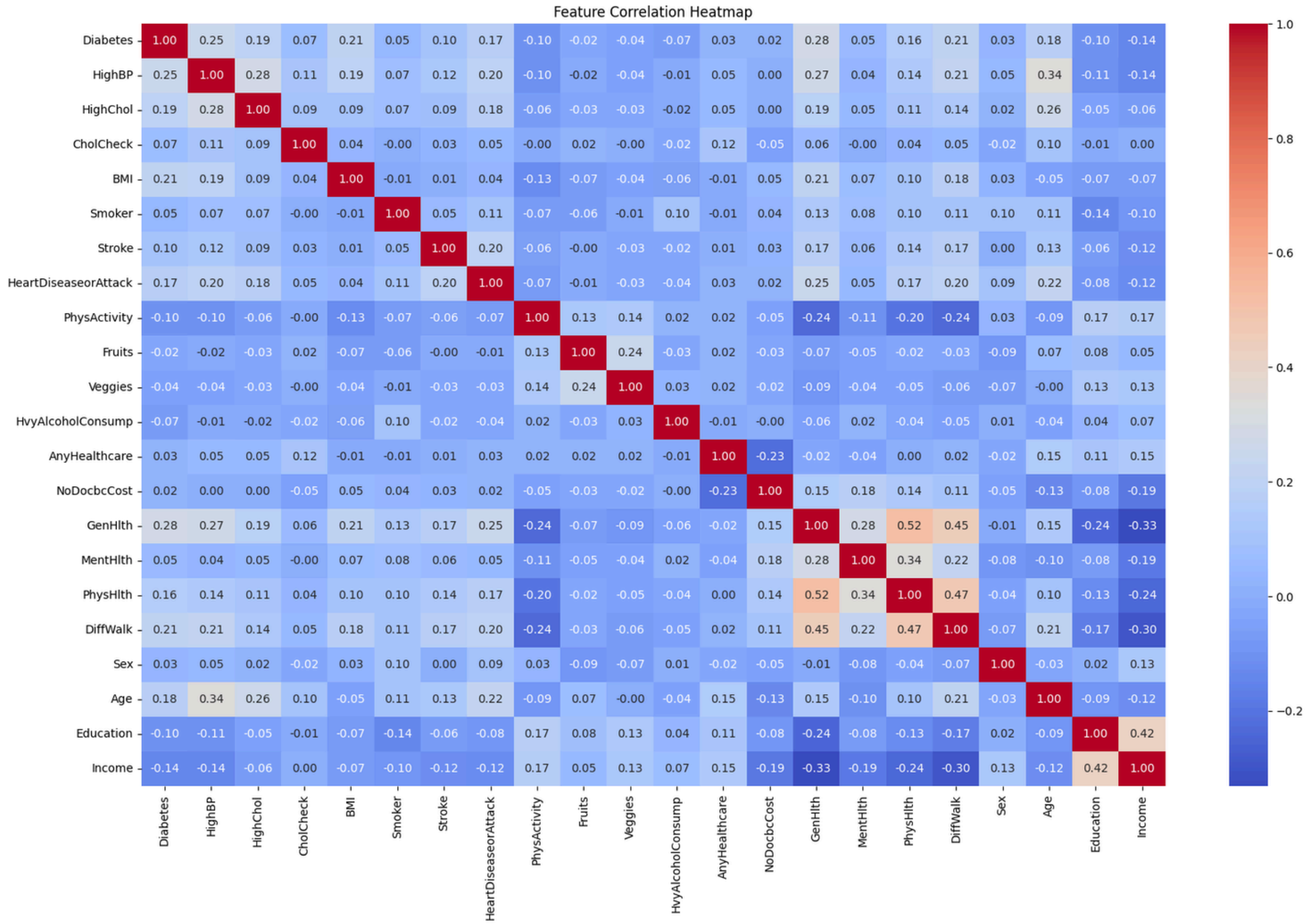
- Higher Age correlates more with higher chance of diabetes,
- Higher BMI shows greater proportion of diabetes, and combined with Age shows more instances of diabetes.
- PhysActivity is not directly related to any of the other features, although it also lowers the likelihood of diabetes.



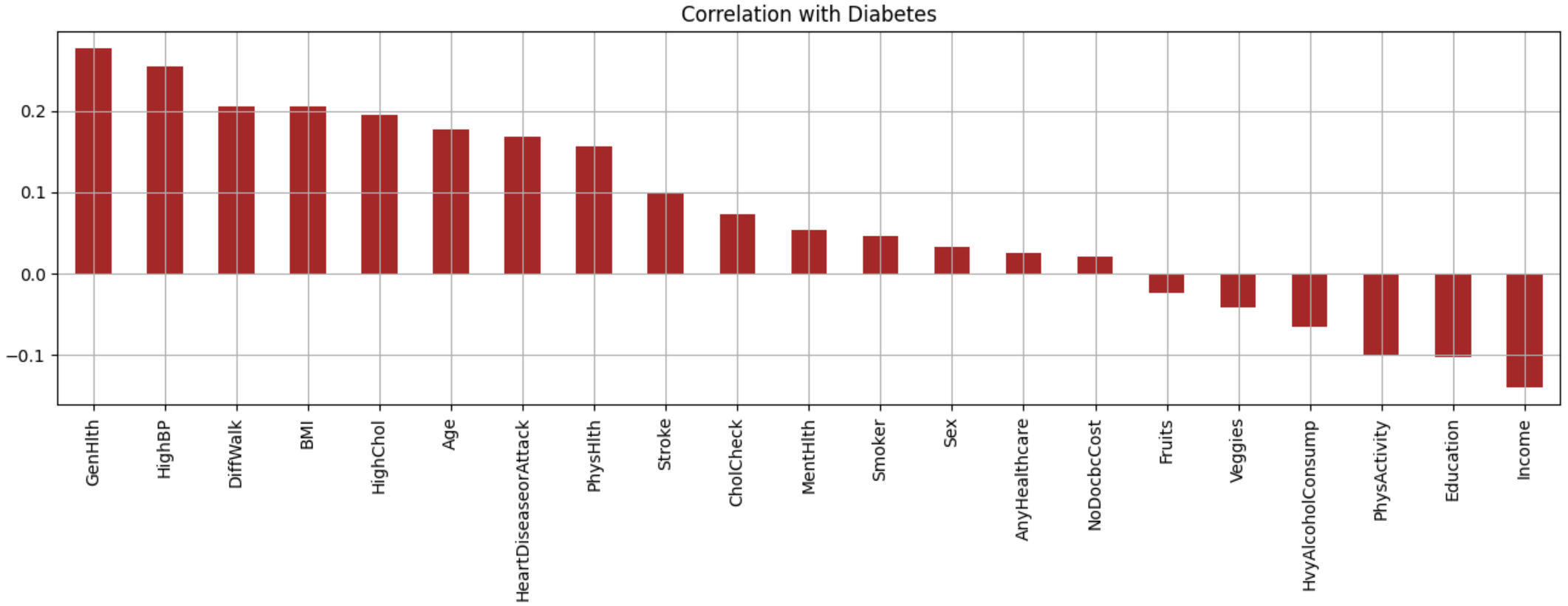
HeatMap

- **Strong Positive Relation:**
  - (GenHlth, PhysHlth)
  - (PhysHlth, DiffWalk)
  - (GenHlth, DiffWalk)

- **Strong Negative Relation:**
  - (GenHlth, Income)
  - (DiffWalk, Income)



Features Correlation with Target



Showing the features ordered gives us a better look at the overall correlation between them and the target feature, with **GenHlth, HighBP** being the highest positively correlated features, followed closely by **DiffWalk** and **BMI**, while **Income, Education** and **PhysActivity** are the most inversely correlated features, although at a smaller rate.

High vs Low Correlated Features

High Correlated Features

>= 0.1

['Diabetes', 'HighBP', 'HighChol', 'BMI', 'HeartDiseaseorAttack', 'PhysActivity', 'GenHlth', 'PhysHlth', 'DiffWalk', 'Age', 'Education', 'Income'],

Low Correlated Features

< 0.05

['Smoker', 'Fruits', 'Veggies', 'AnyHealthcare', 'NoDocbcCost', 'Sex']

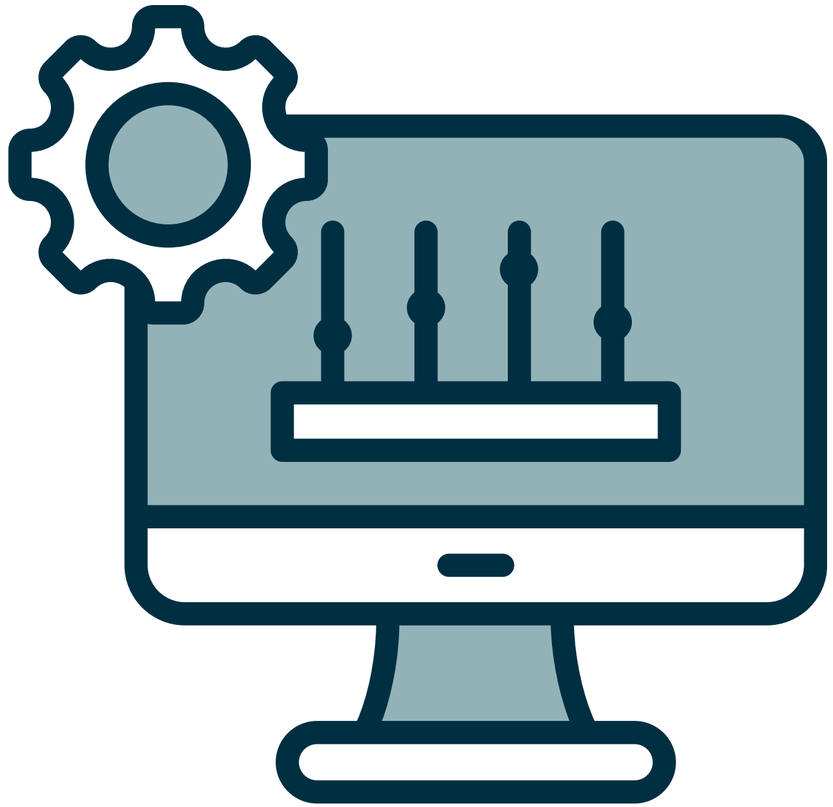
EDA Conclusions

- There are more men diagnosed with diabetes than women.
- Diabetes prevalence increases with age.
- The highest number of diabetes cases is in the age group 65-69 years.
- Higher education levels are associated with a larger proportion of non-diabetic individuals.
- Individuals with lower education levels tend to have higher diabetes prevalence.
- Individuals with high blood pressure are significantly more likely to have diabetes.
- High cholesterol is a strong risk factor for diabetes.
- Regular physical activity is associated with a lower prevalence of diabetes.
- Regular consumption of fruits and vegetables is linked to a lower risk of diabetes.
- Smoking status shows minimal association with diabetes risk in this dataset.
- Heavy alcohol consumption does not show a strong relationship with diabetes.
- Individuals with access to healthcare have slightly higher diagnosed diabetes rates, likely due to increased detection and diagnosis.
- Poor general health is strongly correlated with diabetes.
- Difficulty walking is a significant indicator of diabetes risk.

Machine Learning Models

1. **Logistic Regression**
2. **Random Forest**
3. **Decision Trees**
4. **KNN**
5. **SVM**
6. **Gradient Boosting**

(with HalvingRandomSearchCV HyperParameter Tuning)



Model Performance (Before Tuning)

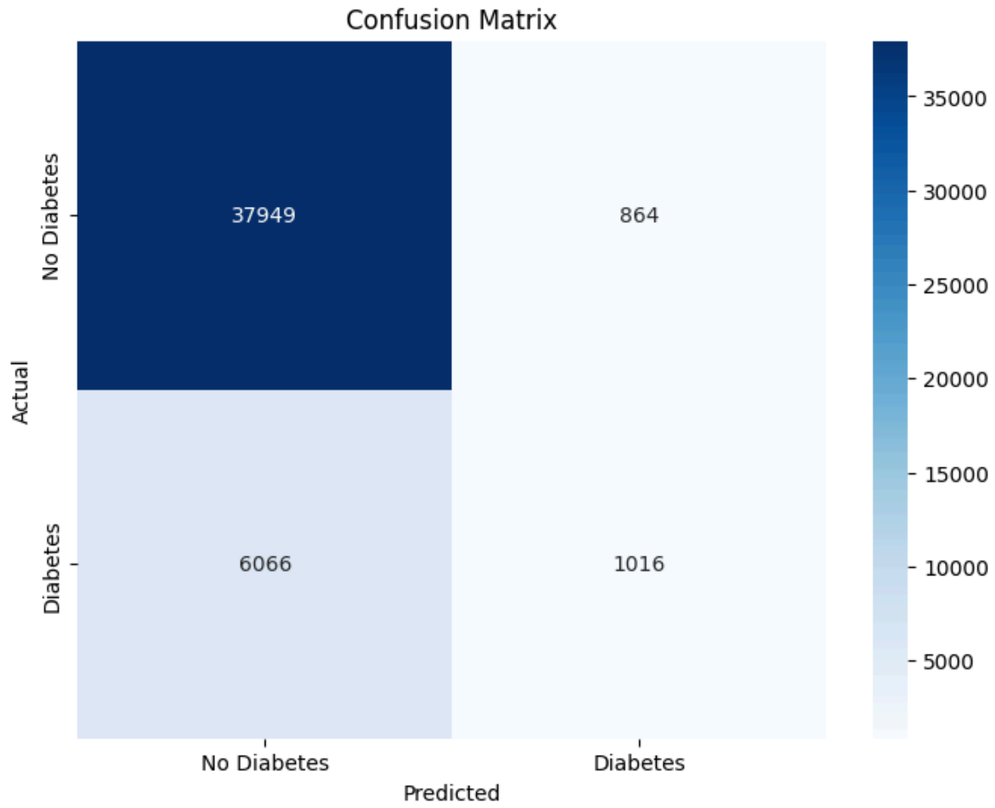
Model	Accuracy
GradientBoosting	0.851313
LogisticRegression	0.849003
SVM	0.84569
RandomForest	0.836322
KNN	0.830592
DecisionTree	0.787210

Metrics Key Conclusions:

- Gradient Boosting is overall the best-performing model with an accuracy of 0.8513.
- Logistic Regression and Gradient Boosting were very good at identifying instances of Class 0 (No Diabetes), with a recall of 0.98.
- All models were able to perform very well for Class 0 (Diabetes) but poorly for Class 1 (Diabetes).
- SVM had trouble identifying instances of Class 1 (Diabetes) showing bias to predict only Class 0 (No diabetes) instances.
    - This is probably due to class imbalance since there are more people without diabetes in the dataset.

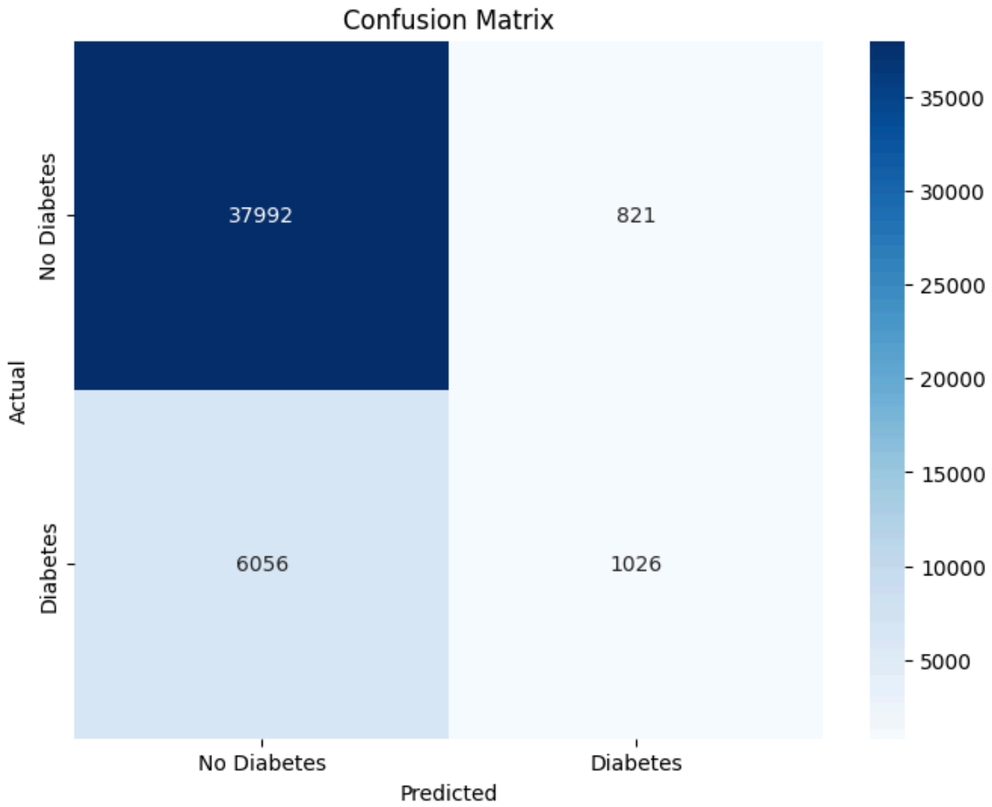
Logistic Regression Performance

- **Best Model Parameters:**
  - {'solver': 'liblinear', 'penalty': 'l1', 'max_iter': 200, 'C': 0.1}
- **Best LogisticRegression Score:**
  - 0.8773
- **Best LogisticRegression Estimator:**
  - LogisticRegression(C=0.1, max_iter=200, penalty='l1', solver='liblinear')
- **Evaluation Metrics on Test Set:**
  - **accuracy**: 0.8490
  - **precision**: 0.8125
  - **recall**: 0.8490
  - **f1_score**: 0.8099



Random Forest Performance

- **Best Model Parameters:**
  - {'n_estimators': 250, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_depth': None, 'bootstrap': True}
- **Best RandomForestClassifier Score:**
  - 0.8829
- **Best RandomForestClassifier Estimator:**
  - RandomForestClassifier(min_samples_leaf=4, min_samples_split=5, n_estimators=250, random_state=42)
- **Evaluation Metrics on Test Set:**
  - **accuracy**: 0.8502
  - **precision**: 0.8151
  - **recall**: 0.8502
  - **f1_score**: 0.8110



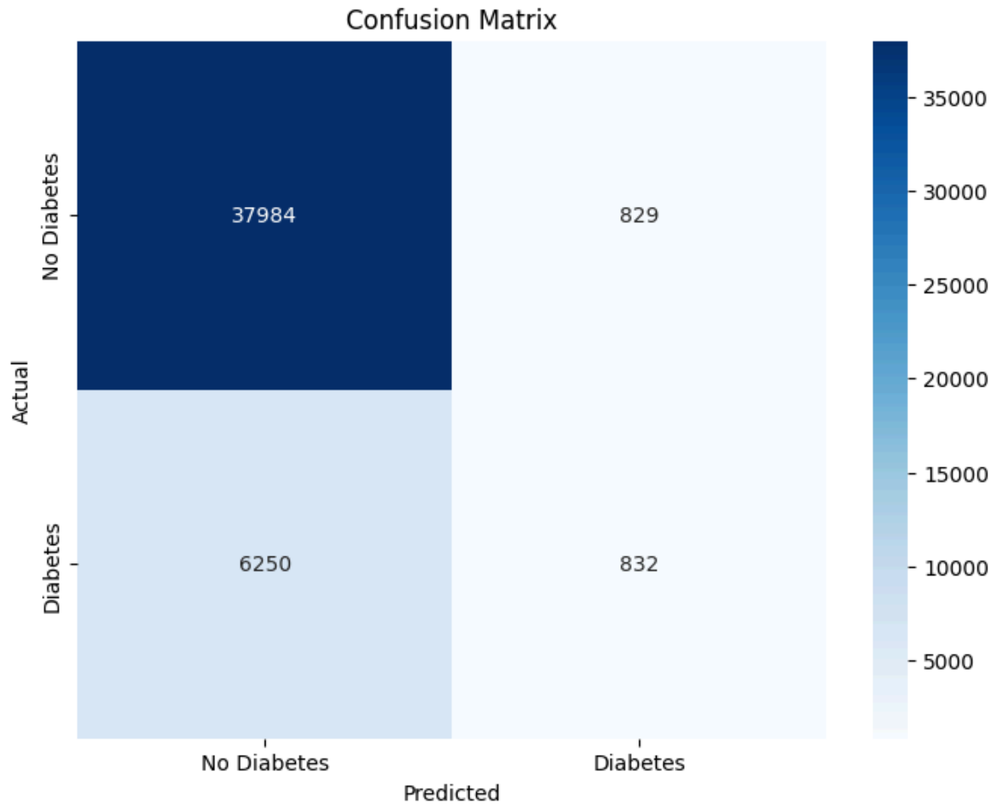
Decision Trees Performance

- **Best Model Parameters:**
  - {'splitter': 'random', 'min_samples_split': 20, 'min_samples_leaf': 8, 'max_features': 'sqrt', 'max_depth': 30, 'criterion': 'gini'}

- **Best DecisionTreeClassifier Score:**
  - 0.8662

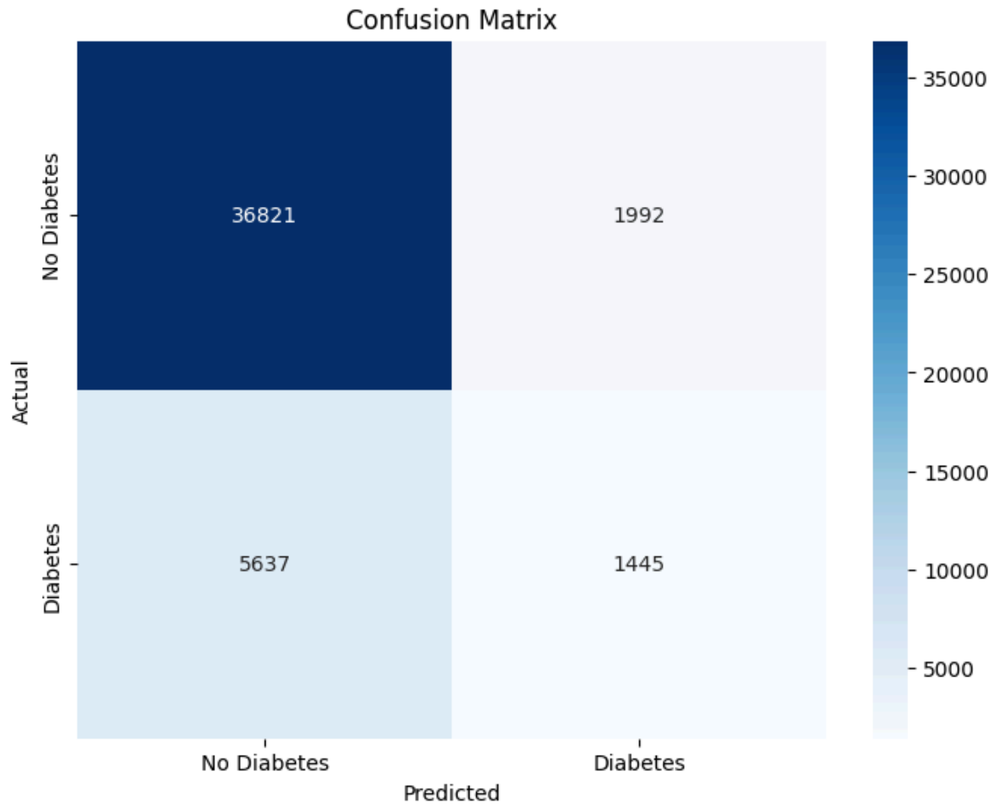
- **Best DecisionTreeClassifier Estimator:**
  - DecisionTreeClassifier(max_depth=30, max_features='sqrt', min_samples_leaf=8, min_samples_split=20, random_state=42, splitter='random')

- **Evaluation Metrics on Test Set:**
  - accuracy: 0.8458
  - precision: 0.8035
  - recall: 0.8458
  - f1_score: 0.8030



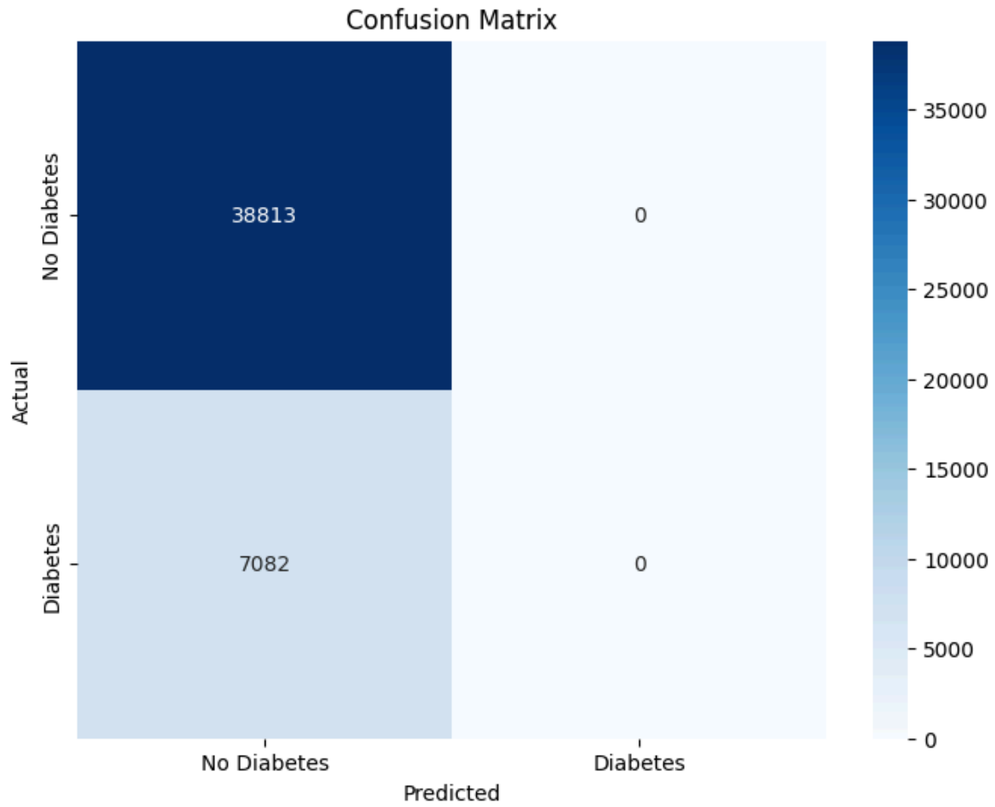
KNN Performance

- **Best Model Parameters:**
  - {'weights': 'distance', 'p': 2, 'n_neighbors': 10, 'leaf_size': 20, 'algorithm': 'ball_tree'}
- **Best KNeighborsClassifier Score:**
  - 0.8830
- **Best KNeighborsClassifier Estimator:**
  - KNeighborsClassifier(algorithm='ball_tree', leaf_size=20, n_neighbors=10, weights='distance')
- **Evaluation Metrics on Test Set:**
  - **accuracy**: 0.8338
  - **precision**: 0.7983
  - **recall**: 0.8338
  - **f1_score**: 0.8087



SVM Performance

- **Best Model Parameters:**
  - {'kernel': 'linear', 'gamma': 'auto', 'degree': 4, 'C': 0.1}
- **Best SVC Score:**
  - 0.8940
- **Best SVC Estimator:**
  - SVC(C=0.1, degree=4, gamma='auto', kernel='linear')
- **Evaluation Metrics on Test Set:**
  - **accuracy**: 0.8457
  - **precision**: 0.7152
  - **recall**: 0.8457
  - **f1_score**: 0.7750



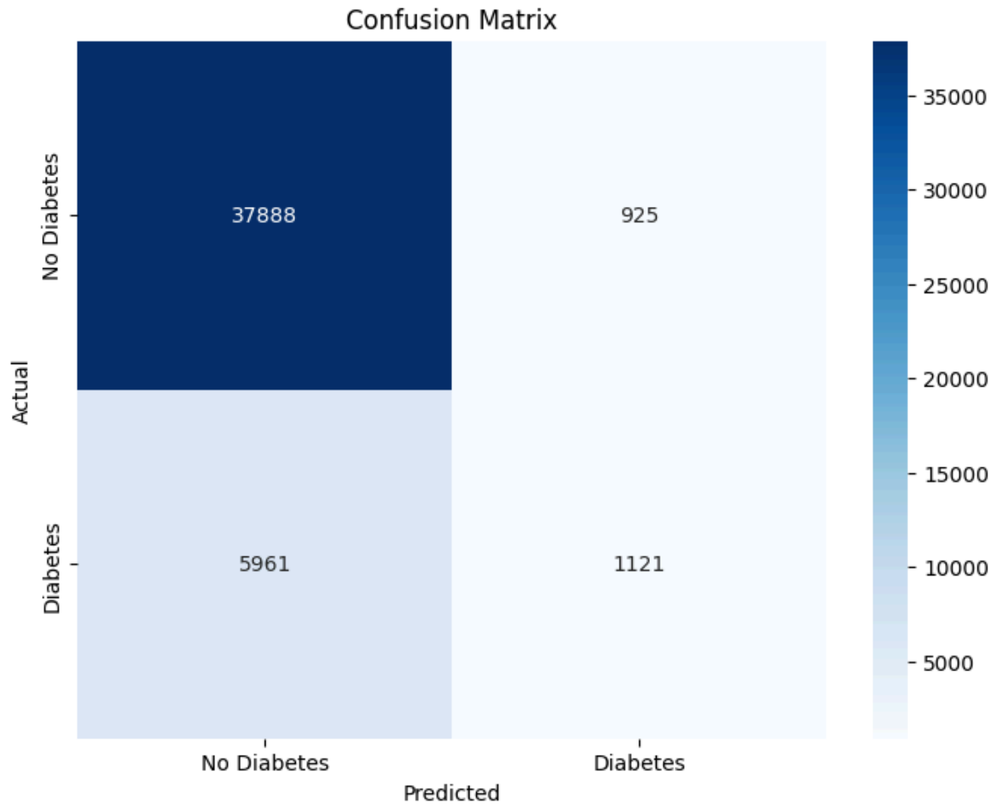
Gradient Boosting Performance

- **Best Model Parameters:**
  - {'warm_start': False, 'subsample': 0.8, 'n_iter_no_change': 5, 'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 7, 'learning_rate': 0.2}

- **Best GradientBoostingClassifier Score:**
  - 0.8830

- **Best GradientBoostingClassifier Estimator:**
  - GradientBoostingClassifier(learning_rate=0.2, max_depth=7, max_features='log2', min_samples_split=10, n_estimators=200, n_iter_no_change=5, subsample=0.8)

- **Evaluation Metrics on Test Set:**
  - **accuracy**: 0.8500
  - **precision**: 0.8153
  - **recall**: 0.8500
  - **f1_score**: 0.8131



Model Performance (After Tuning)

Model	Accuracy
RandomForest	0.850158
GradientBoosting	0.849962
LogisticRegression	0.849003
DecisionTree	0.845757
SVM	0.845691
KNN	0.833773

Metrics Highlights:

- All models show slightly lower performance on the test set compared to their cross-validation scores which can indicate slight overfitting.
- Logistic Regression, Random Forest, and Gradient Boosting perform similarly in terms of test accuracy, which siggest that feature selection and preprocessing are effective.
- The SVM model's precision (71.52%) is significantly lower than other models, indicating potential issues in handling class imbalance or poor predictions for certain classes.
- Random Forest and Gradient Boosting offer the best balance of accuracy, precision, recall, and F1-score.
- KNN underperforms, probably because of the dataset's size.

Comparative Analysis

Model	Accuracy Before Tuning	Accuracy After Tuning
RandomForest	0.836322	0.850158
GradientBoosting	0.851313	0.849962
LogisticRegression	0.849003	0.849003
DecisionTree	0.787210	0.845757
SVM	0.84569	0.84569
KNN	0.830592	0.833773

Discussion of Results

- Hyperparameter tuning significantly **improved the performance** of **Random Forest** and **Decision Tree.**
- Gradient Boosting Classifier **consistently** performed well before and after tuning, achieving the best test set performance in terms of balanced metrics (accuracy, precision, recall, F1 score).
  - It emerged as the most reliable model overall, with Random Forest as a close second.
- Decision Tree **improved significantly after tuning**.
- While SVM showed a strong performance on the validation set (0.8940), its slightly lower test accuracy (0.8457) may suggest **overfitting or sensitivity to hyperparameters.**
- Logistic Regression and KNN provided **consistent performance** but lacked the flexibility to surpass the tree-based models.
- Gradient Boosting, Random Forest, and Decision Tree exhibited **well-balanced precision and recall scores**, with Gradient Boosting having the **highest F1 score** (0.8131).

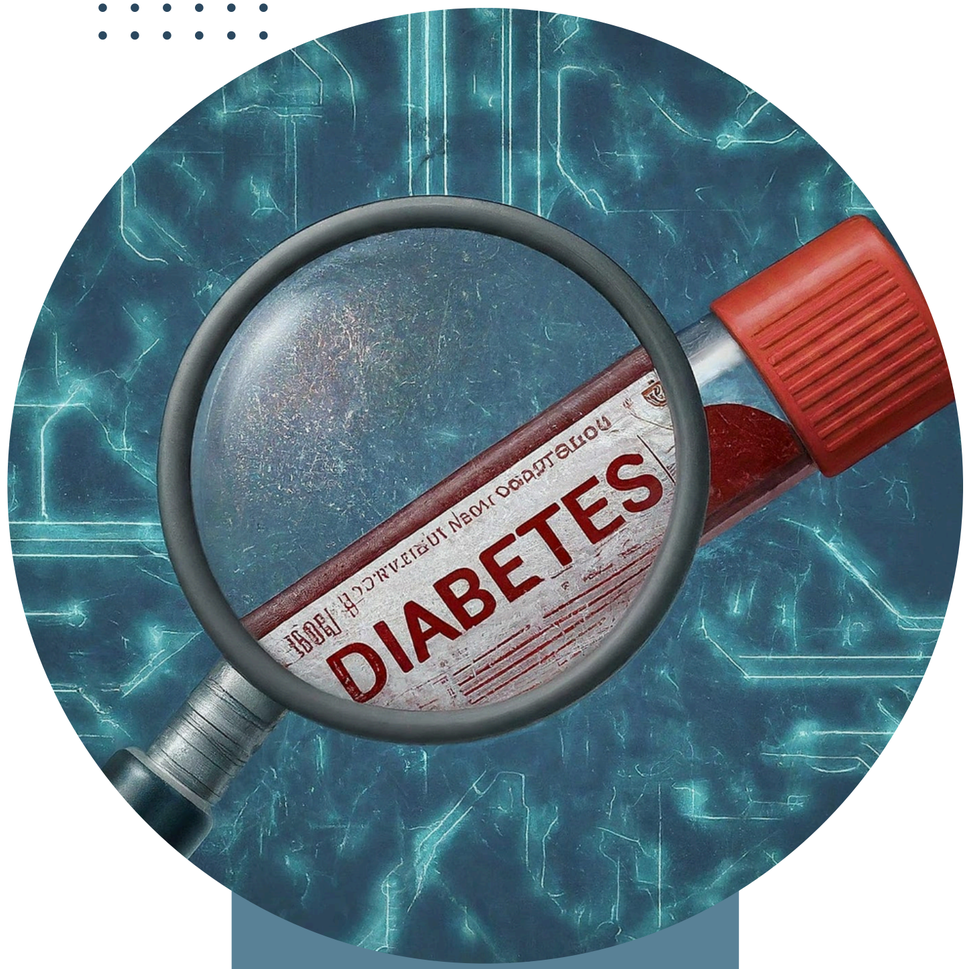
Conclusion

Key Findings:

- This study can contribute to designing more effective interventions to mitigate diabetes risks.
- Key factors that influence diabetes prevalence are **high blood pressure, low general health, lack of physical activity** and **high BMI**.
- ***HalvingRandomSearchCV*** proved to be the most computationally efficient hyperparameter tuning algorithm for such a large dataset.
- Both ***RandomForestClassifier*** and ***GradientBoostingClassifier*** were the best-performing models overall for predicting diabetes risk after parameter tuning.

Future Work:

- More extensive HyperParameter Tuning on the best models.
- Deal with class imbalance.
- Test prediction of diabetes on sample data with the best model.



Università di Bologna

Thank You

for your attention

January 22, 2025

Project by José Ribeiro & Sofia Costa
1900130105 & 1900129396

Project Link: https://github.com/sophie-mc-dev/unibo-big-data-analytics