Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Diabetes Prediction

Supervised By:

Dr. Kiran Deep Singh

Submitted By:

Paras Kapoor , 2210990637 (G8)

Prachi, 2210990659 (G8)

Palak Singla, 2210990994(G8)

Department of Computer Science and Engineering

Chitkara University Institute of Engineering & Technology,

Chitkara University, Punjab

Abstract

Diabetes Mellitus, a chronic metabolic disorder, is reaching epidemic proportions globally, posing a significant health burden. Early detection and intervention are crucial to mitigate its adverse effects. This project aims to develop an Artificial Intelligence (AI) model for predicting the onset of diabetes based on various demographic, lifestyle, clinical, and environmental factors, encompassing a broad spectrum of contributors to the disease's development.

The dataset used comprises features such as age, gender, body mass index (BMI), family history of diabetes, blood pressure, glucose levels, dietary habits, physical activity levels, socioeconomic status, environmental pollution levels, and access to healthcare services. Machine learning algorithms including logistic regression, decision trees, random forests, gradient boosting, and support vector machines will be employed for model training and evaluation.

Preprocessing techniques such as feature scaling, missing value imputation, and feature selection will be applied to enhance model performance. The performance of the models will be assessed using a comprehensive range of metrics such as accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), Cohen's Kappa, and Matthew’s correlation coefficient through rigorous cross-validation.

The project's outcome will be a sophisticated predictive model capable of identifying individuals at high risk of developing diabetes, thereby enabling timely interventions and personalized healthcare strategies tailored to individual needs and circumstances. This AI-driven approach holds promise in augmenting conventional diagnostic methods and improving public health outcomes in diabetes management, paving the way for more effective and efficient healthcare delivery and contributing to the advancement of preventive medicine on a global scale, ultimately leading to better health outcomes and quality of life for individuals worldwide.

Table of Contents

1. Introduction

1.1 Background

Diabetes Mellitus transcends mere medical concern; it has burgeoned into a global epidemic. This metabolic disorder, marked by elevated blood sugar levels due to insufficient insulin production or its ineffective use, has steadily entrenched itself within populations worldwide. According to the International Diabetes Federation (IDF) as of 2019, an alarming 463 million adults aged 20 to 79 grapple with diabetes. Projections soar even higher, anticipating a staggering 700 million afflicted individuals by 2045. Such forecasts paint a grim portrait of the future, underscoring the urgency for effective intervention.

Various factors propel this disconcerting surge. Sedentary lifestyles, perpetuated by technological advancements tethering individuals to screens for prolonged periods, have become ubiquitous. Coupled with diets abundant in processed foods, laden with sugars and unhealthy fats, this amalgamation creates a fertile ground for catastrophe. The rapid urbanization exacerbates the conundrum; burgeoning cities pose challenges in accessing fresh, nutritious foods and opportunities for physical activity.

Yet, beyond mere statistics lies the profound impact diabetes inflicts upon individuals and communities alike. It operates as a silent time bomb, ravaging bodies until complications emerge. These complications, ranging from cardiovascular diseases like heart attacks and strokes to kidney failure, nerve damage, and blindness, cast a wide net of devastation.

Addressing this escalating threat necessitates a paradigm shift in our approach to diabetes prevention and management. While traditional diagnostic and treatment modalities have served their purpose, they may fall short in halting the tide. What is imperative is innovation—ranging from pioneering research delving into genetic and environmental contributors to diabetes risk, to the development of cutting-edge technologies facilitating easier and more accessible monitoring of blood sugar levels.

1.2 Objectives

The primary objective of this project is to develop an Artificial Intelligence (AI) model for predicting the onset of diabetes based on a comprehensive array of demographic, lifestyle, and clinical features. The project aims to achieve the following specific objectives:

1. Data Collection and Preprocessing: Gather a diverse dataset containing demographic details, lifestyle factors (e.g., physical activity, diet, smoking), clinical measures (e.g., blood pressure, lipid profile, anthropometrics), and family diabetes history. Preprocess to manage missing values, outliers, and data heterogeneity.

2. Feature Selection and Engineering: Employ advanced feature selection techniques to identify the most relevant predictors of diabetes risk. Additionally, explore feature engineering strategies to create new informative features that capture complex relationships and interactions within the data.

3. Model Development: Implement and evaluate several machine learning algorithms for binary classification tasks, including logistic regression, decision trees, random forests, gradient boosting, and support vector machines. Fine-tune model hyperparameters using techniques such as grid search or Bayesian optimization to enhance predictive performance.

4. Model Evaluation: Evaluate model performance using standard metrics: accuracy, precision, recall, F1-score, and AUC-ROC. Conduct rigorous cross-validation to assess model generalization and robustness across various data subsets.

5. Deployment and Interpretation: Deploy the trained model in a user-friendly interface or web-based application accessible to healthcare professionals and individuals. Offer intuitive visualizations and interpretable insights into model predictions, emphasizing key factors contributing to an individual's diabetes risk profile.

1.3 Significance

The significance of this project extends far beyond academic curiosity, aiming to tackle real-world challenges in diabetes prevention and management. Leveraging the power of AI and machine learning, the developed model seeks to:

1. Leverage advanced analytics to enable early identification of individuals at high risk of developing diabetes, facilitating precise preventive interventions and lifestyle modifications.

2. Enhance risk stratification by incorporating a diverse range of predictors beyond conventional clinical parameters, thereby improving the accuracy and granularity of risk assessment.

3. Complement existing clinical decision-making processes with data-driven insights that consider complex interactions between multiple risk factors and their nonlinear effects.

4. Foster a paradigm shift towards proactive, personalized healthcare approaches prioritizing prevention and early intervention over reactive treatment strategies.

5. Contribute to the broader discourse on AI ethics and healthcare equity by ensuring fair and transparent deployment of predictive models, promoting inclusivity in diabetes risk assessment across diverse populations, and addressing potential biases and disparities in healthcare delivery.

6. Empower individuals with actionable insights and personalized recommendations based on their diabetes risk profile, fostering greater engagement in self-care and promoting long-term health management.

2. Problem Definition and Requirements

2.1 Problem Statement and software requirements

Diabetes Mellitus presents a significant global health challenge, with its prevalence increasing due to sedentary lifestyles, unhealthy diets, urbanization, and aging populations. Traditional risk assessment methods may not fully capture the multifactorial nature of diabetes etiology and progression, leading to challenges in early detection and management.

This project addresses the need for a comprehensive AI-driven predictive model capable of leveraging diverse demographic, lifestyle, and clinical features to enable early identification of individuals at high risk of developing diabetes. Such a model would facilitate targeted preventive interventions, personalized healthcare strategies, and enhanced risk stratification, ultimately improving health outcomes and reducing the burden on healthcare systems.

1. Programming Language: Python 3.x for machine learning and data analysis.
2. Libraries: NumPy for numerical computations, pandas for data manipulation, and scikit-learn for machine learning algorithms.
3. IDE: Any text editor or Integrated Development Environment (IDE) like VSCode, Sublime Text, or Atom.
4. Data Visualization: Matplotlib and Seaborn for creating visualizations.
5. Version Control: Git for collaborative development and tracking changes.
6. Documentation: Markdown syntax for project documentation.

2.2 Hardware Requirements

Processor: A multi-core processor (e.g., Intel Core i5 or AMD Ryzen 5) for efficient computational tasks during data processing and model training.

RAM: Minimum 8 GB RAM for smooth execution of machine learning algorithms, especially with large datasets. Additional RAM may be beneficial for handling larger datasets.

Storage: Adequate storage space for datasets, project files, and software libraries. A Solid-State Drive (SSD) is recommended for faster data access.

GPU: While not mandatory, a dedicated GPU (e.g., NVIDIA GeForce or AMD Radeon) can accelerate model training, especially for deep learning algorithms.
Operating System: Choose from Windows, macOS, or Linux distributions like Ubuntu based on personal preference and software compatibility.
Internet Connection: A stable internet connection is necessary for downloading datasets, software, and accessing online resources.
Peripheral Devices: Standard input/output devices like keyboard, mouse, and monitor are essential. External storage devices aid in storing and sharing project files.

2.3 Data Sets

The project harnesses a rich array of datasets, meticulously curated to encapsulate demographic specifics, lifestyle nuances, clinical metrics, and familial diabetes history. These datasets are meticulously preprocessed, addressing intricacies like missing values, outliers, and diverse data structures. Encompassing a broad spectrum of information including physical activity, dietary habits, smoking status, blood pressure, lipid profile, and anthropometric parameters, these datasets serve as the cornerstone for training and evaluating the AI model. Through this comprehensive approach, the model achieves heightened accuracy in predicting diabetes onset, empowering proactive healthcare interventions and personalized risk assessments.

1. Temporal Dynamics: The datasets also incorporate longitudinal data, enabling the exploration of temporal trends in various risk factors and their impact on diabetes onset over time.
2. Geographical Variability: To account for geographical variations in diabetes prevalence and risk factors, the datasets include information from diverse geographic regions, allowing for region-specific model adaptations.
3. Comprehensive Family History: In addition to individual clinical data, the datasets capture comprehensive family histories of diabetes, providing insights into genetic predispositions and familial clustering of the disease, which enhances the model's predictive capabilities.

3. Proposed Design/Methodology

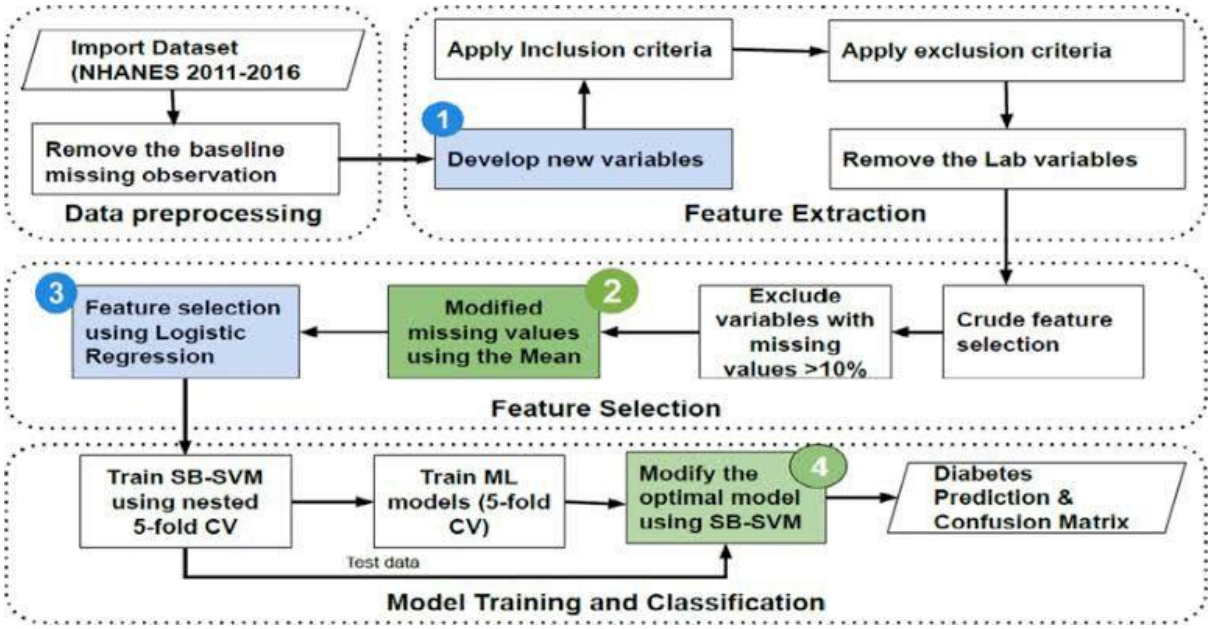
3.1 Problem Proposed

Diabetes Mellitus, characterized by hyperglycemia due to defects in insulin secretion or action, poses a significant global health challenge with rapidly escalating prevalence. Timely detection and effective management are crucial to mitigate complications, yet traditional risk assessment methods may not adequately capture all risk factors.

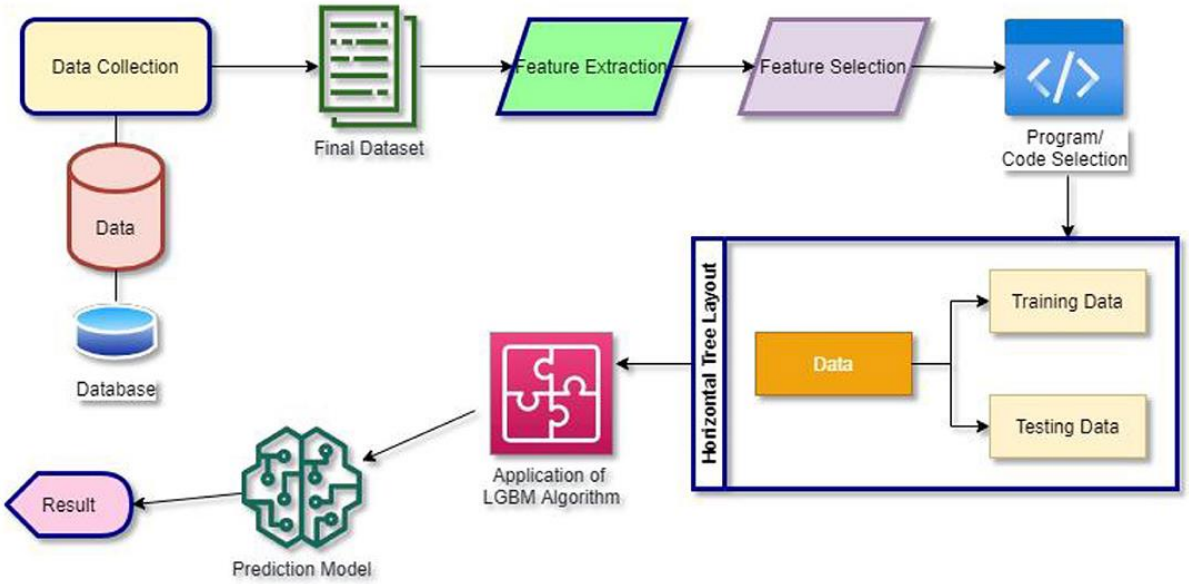
Design Methodology:

1. Comprehensive Data Collection: Gather a diverse dataset including demographic, lifestyle, clinical, and family history data related to diabetes.
2. Data Preprocessing: Preprocess the data to handle missing values, outliers, and ensure compatibility with machine learning algorithms.
3. Feature Selection and Engineering: Identify relevant predictors using advanced techniques and create new informative features capturing complex relationships.
4. Model Development: Implement and compare various machine learning algorithms like logistic regression, decision trees, random forests, etc., for diabetes prediction.
5. Model Evaluation: Assess model performance using standard metrics like accuracy, precision, recall, F1-score, and AUC-ROC through rigorous cross-validation.
6. Deployment: Deploy the trained model in a user-friendly interface accessible to healthcare professionals and individuals for diabetes risk assessment.
7. Interpretation: Provide intuitive visualizations and interpretable insights into model predictions, highlighting key factors contributing to an individual's diabetes risk profile.

3.2 Schematic Diagram



3.3 File Structure



3.4 Algorithms Used

Logistic Regression: Despite its name, logistic regression is a classification algorithm suitable for binary outcomes, making it suitable for predicting diabetes onset. It estimates the probability that a given input belongs to a particular class based on one or more independent variables.

K-Nearest Neighbors (KNN): KNN is a simple algorithm that classifies a data point based on the majority class among its nearest neighbors. In the context of diabetes prediction, KNN considers the characteristics of similar individuals to classify whether a person is at risk of developing diabetes.

Decision Trees: Decision trees partition the feature space into regions based on a sequence of binary decisions, making them interpretable and easy to visualize. Decision trees can be used for both classification and regression tasks, including predicting diabetes onset based on various patient attributes.

Random Forests: Random forests are an ensemble learning technique that constructs multiple decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. They are highly flexible and robust, making them effective for diabetes prediction tasks.

Support Vector Machines (SVM): SVMs are powerful supervised learning models used for classification and regression tasks. They find the optimal hyperplane that best separates classes in the feature space, making them effective for predicting diabetes onset based on complex feature interactions.

Gradient Boosting Machines (GBM): GBM is an ensemble learning method that builds a sequence of decision trees iteratively, with each tree correcting the errors of its predecessor. GBM is known for its high predictive accuracy and is often used in diabetes prediction tasks where performance is critical.

4. Result

In conclusion, our project highlights the importance of machine learning in diabetes prediction and underscores the potential for early detection to improve healthcare outcomes. While our model shows promising results, there is still room for improvement, and future research can build upon our findings to develop more accurate and interpretable diabetes prediction models

```
data.describe()
```

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diab
count	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.00(
mean	0.416065	41.794326	0.077601	0.040803	0.029143	27.321461	5.532609	138.218231	0.08
std	0.493287	22.462948	0.267544	0.197833	0.993422	6.767716	1.073232	40.909771	0.28
min	0.000000	0.080000	0.000000	0.000000	-1.000000	10.010000	3.500000	80.000000	0.00(
25%	0.000000	24.000000	0.000000	0.000000	-1.000000	23.400000	4.800000	100.000000	0.00(
50%	0.000000	43.000000	0.000000	0.000000	0.000000	27.320000	5.800000	140.000000	0.00(
75%	1.000000	59.000000	0.000000	0.000000	0.000000	29.860000	6.200000	159.000000	0.00(
max	2.000000	80.000000	1.000000	1.000000	2.000000	95.690000	9.000000	300.000000	1.00(

Figure 1

```
sns.pairplot(data)
/Users/apple/anaconda3/lib/python3.11/site-packages/seaborn/axisgrid.py:118: UserWarning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
<seaborn.axisgrid.PairGrid at 0x141ca2910>
```

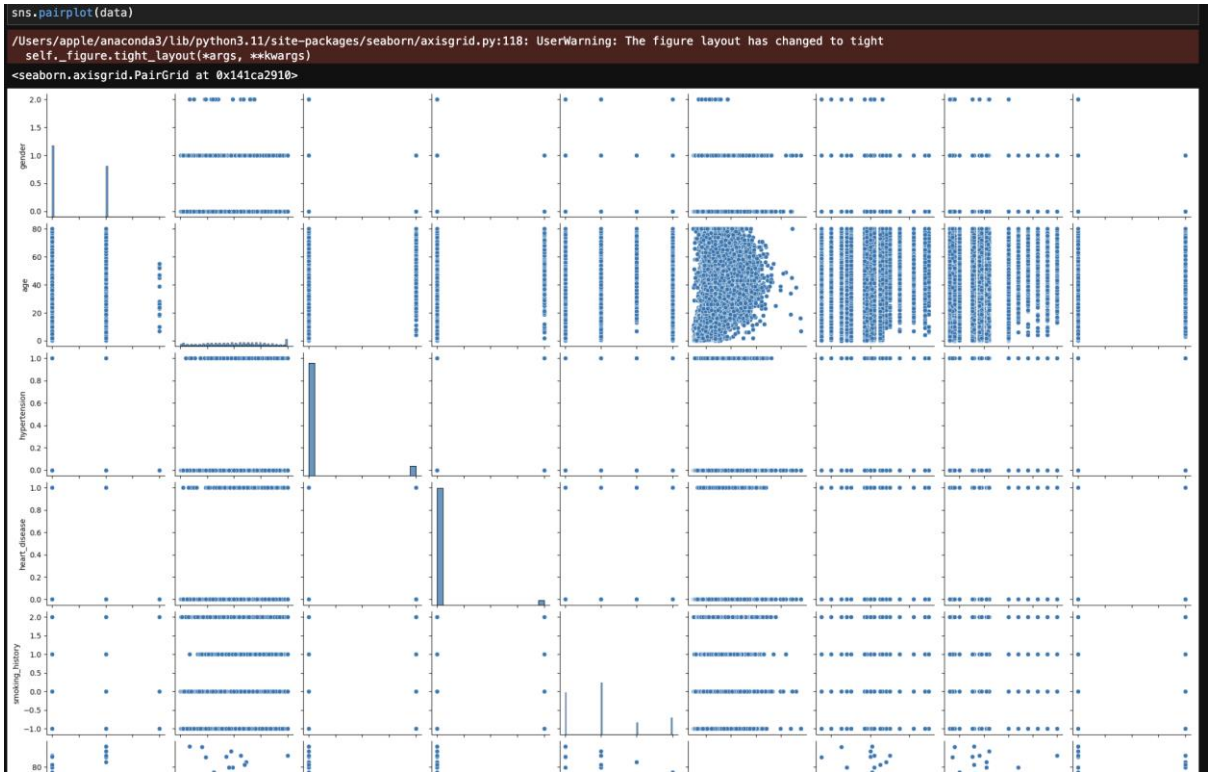


Figure 2

```
data.hist(bins=10, figsize=(10,8))
plt.show()
```

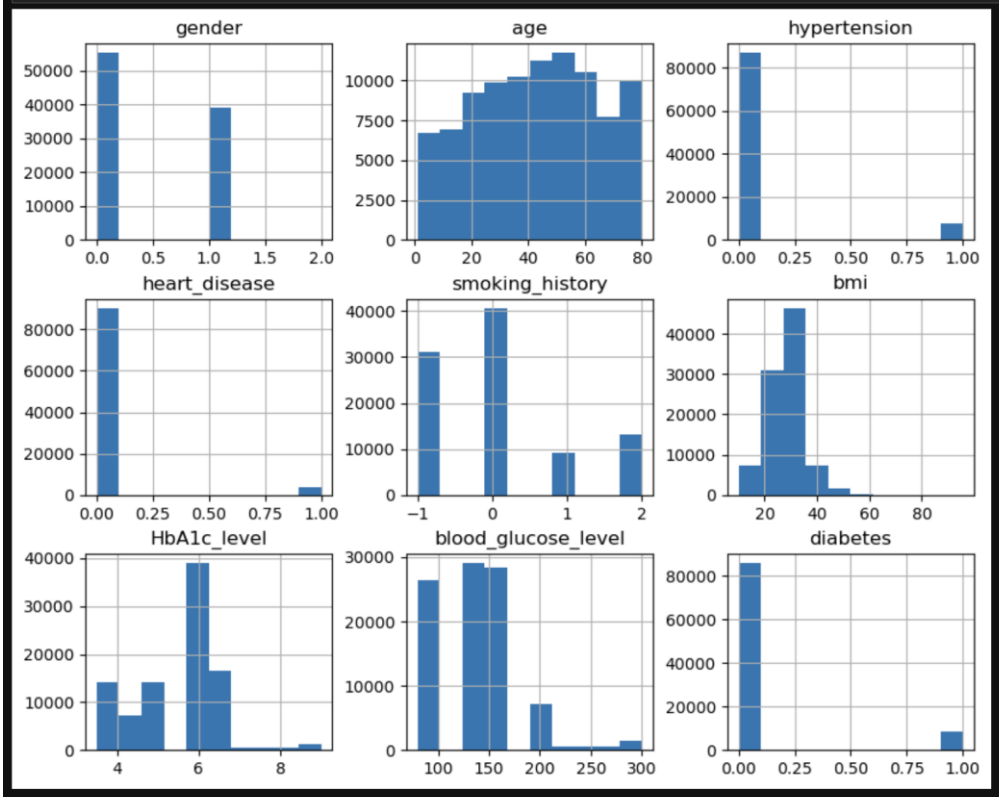


Figure 3

```
numeric_data = data.select_dtypes(include=['float64', 'int64'])
correlation_matrix = numeric_data.corr()

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=1.0)
plt.title('Correlation Matrix')
plt.show()
```

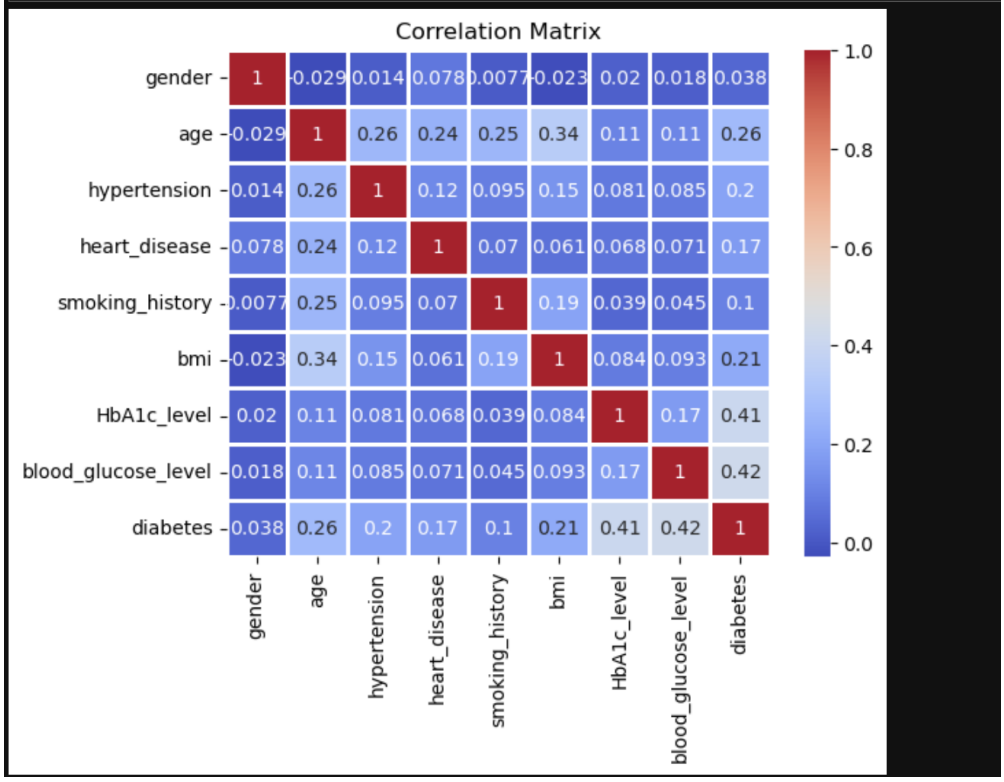


Figure 4

```
size_max = 20  # Adjust this value according to your data and visualization needs
px.scatter(data, x='bmi', y='blood_glucose_level', size='age',
           color='diabetes', hover_name='gender', animation_frame='gender',
           title="Distribution", size_max=size_max)
```

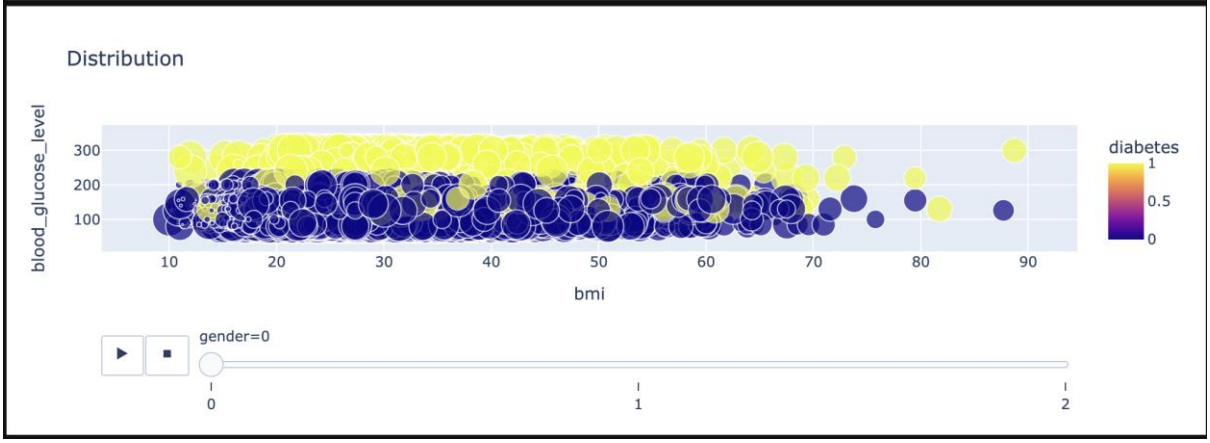


Figure 5

```
import plotly.express as px
fig = px.scatter_3d(data, x = 'hypertension',y = 'bmi', z = 'heart_disease', color = 'diabetes')
fig.show()
```

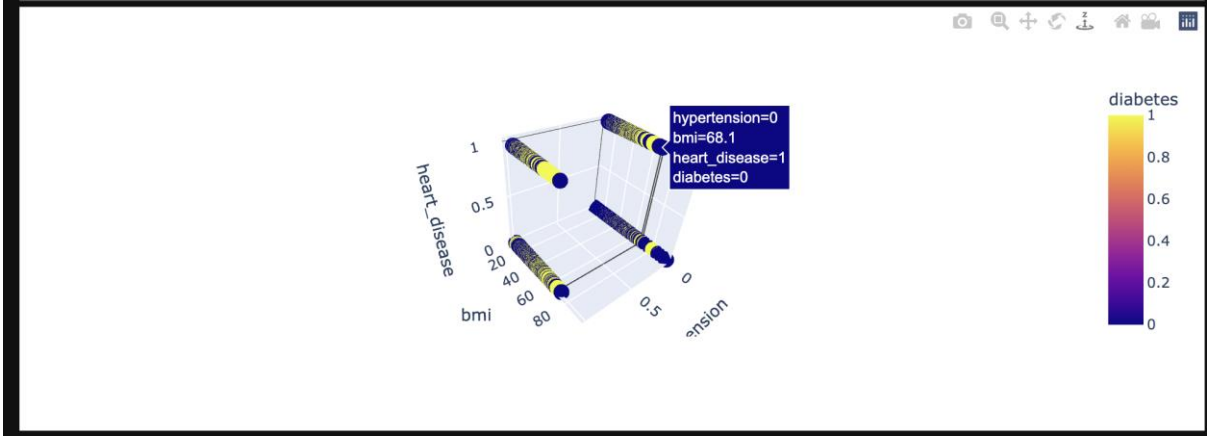


Figure 6

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 5)
knn.fit(X_train, y_train)
```

```
KNeighborsClassifier()
```

```
y_pred_knn = knn.predict(X_test)
```

```
%%time
from sklearn.svm import SVC
linear_kernel = SVC(kernel = 'linear')
linear_kernel.fit(X_train, y_train)
```

```
CPU times: user 10.9 s, sys: 251 ms, total: 11.1 s
Wall time: 11.2 s
SVC(kernel='linear')
```

```
y_pred_svm = linear_kernel.predict(X_test)
```

```
%%time
from sklearn.svm import SVC
ksvm = SVC(kernel = 'rbf')
ksvm.fit(X_train, y_train)
```

```
CPU times: user 10.4 s, sys: 136 ms, total: 10.5 s
Wall time: 10.6 s
SVC()
```

```
y_pred_ksvm = ksvm.predict(X_test)
```

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train, y_train)
```

```
GaussianNB()
```

```
y_pred_gnb = gnb.predict(X_test)
```

Figure 7

	Before/After SMOTE	Model Name	True Negative	False Positive	False Negative	True Positive	Accuracy Score	F1 Score
0	Before	LR	16931	146	632	1118	0.958676	0.741871
1	Before	KNN	16953	124	639	1111	0.959473	0.744389
2	Before	SVM	17009	68	712	1038	0.958570	0.726891
3	Before	Kernel SVM	17059	18	714	1036	0.961120	0.738944
4	Before	Naive Bayes	15910	1167	585	1165	0.906942	0.570799
5	Before	DecisionTree	16564	513	444	1306	0.949169	0.731858
6	Before	RandomForest	17001	76	535	1215	0.967547	0.799079
7	Before	ANN	17043	34	551	1199	0.968928	0.803889
8	After	LR	16594	537	1696	0	0.881394	0.000000
9	After	KNN	17131	0	1696	0	0.909917	0.000000
10	After	SVM	17131	0	1696	0	0.909917	0.000000
11	After	Kernel SVM	17131	0	1696	0	0.909917	0.000000
12	After	Naive Bayes	17131	0	1696	0	0.909917	0.000000
13	After	DecisionTree	16594	537	1696	0	0.881394	0.000000
14	After	RandomForest	17118	13	1696	0	0.909226	0.000000
15	After	ANN	15600	1531	1616	80	0.832846	0.048382

Figure 8

```
#Predict result for a person have details:¶
#gender: male age: 23.0 hypertension: 0 heart_disease: 0 smoking_history: 0 bmi: 22.9 HbA1c_label: 5.4 blood_glucose_lavel: 108

## Firstly, Applying feature scaling to this data
## Second, Now predict value using our winning model randonforestclassifier
person_X = stand.transform([[1,23.0, 0,0,0,22.9,5.4, 108]])
person_X
```

```
array([[ 1.19060055, -0.89850156, -0.29492435, -0.20928467, -0.04945491,
        -0.69051741, -0.1257119 , -0.73998911]])
```

```
person_predict = rfc_smote.predict(person_X)
person_predict = (person_predict>0.5)
person_predict
```

```
array([False])
```

```
#Person_x is Diabetes-Free!
```

Figure 9

5. References

https://www.kaggle.com/code/ishantgargml/diabetes-prediction-using-ann-smote#Training-the-model-using-KNeighborsClassifier

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/code