PREDICTIVE ANALYSIS OF DIABETES

A Model Evaluation Study

Prepared by:
Sethara Gunawardana
s16309

Abstract

The early and accurate classification of diabetes remains a critical challenge in the field of public health and clinical decision-making. With the rise of machine learning (ML) techniques in the healthcare domain, there has been increasing interest in leveraging predictive algorithms to support early diagnosis and risk stratification for chronic diseases such as diabetes. This study investigates the effectiveness and comparative performance of five supervised machine learning models in predicting diabetes status among individuals, using a dataset that includes both clinical indicators and lifestyle-related attributes.

The models evaluated in this study include Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Extreme Gradient Boosting (XGBoost). These models were selected to represent a diverse set of learning paradigms, including linear modeling, non-linear decision tree-based learning, instance-based classification, kernel-based learning, and ensemble methods. The dataset used for this analysis comprises 128 patient records and 11 features, which cover numerical attributes such as age, BMI, Fasting Blood Sugar (FBS), and HbA1c, as well as categorical variables including gender, blood pressure status, smoking habits, diet, exercise, and family history of diabetes.

The data preprocessing phase included handling categorical variables through one-hot encoding and standardizing numerical features to improve model training efficiency. The dataset was split into training and testing subsets in a 70:30 ratio to ensure robust model evaluation. Each model was trained using the training set and then evaluated on the testing set using several standard performance metrics: accuracy, precision, recall, F1-score, and confusion matrices.

The results of the evaluation revealed that three models—Logistic Regression, Support Vector Classifier, and XGBoost—achieved perfect scores across all evaluation metrics on the test set, indicating a very strong fit to the data. The Decision Tree classifier also performed well but demonstrated a slight reduction in precision, suggesting a potential tendency to overfit. In contrast, the KNN algorithm showed the lowest performance, particularly in terms of recall, highlighting its sensitivity to the feature space and limitations on small datasets.

This report provides an in-depth analysis of the strengths and weaknesses of each algorithm in the context of diabetes prediction. It also offers practical recommendations for deploying these models in real-world healthcare settings, particularly emphasizing model interpretability, scalability, and robustness in clinical environments.

Table of Contents

List of Figures and Tables

1. Introduction

Diabetes is a chronic, progressive metabolic disorder that significantly affects the way the human body regulates blood glucose levels. The disease occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Over time, poorly controlled diabetes can result in serious complications, including cardiovascular disease, neuropathy, nephropathy (kidney failure), retinopathy (leading to blindness), and even limb amputation. As of 2023, the World Health Organization reports that over 422 million people worldwide suffer from diabetes, and this number is expected to rise due to increasing obesity rates, sedentary lifestyles, and aging populations.
In many low and middle-income countries, where access to consistent medical screening and treatment may be limited, the burden of diabetes is even more pronounced. Early diagnosis and effective disease monitoring are essential to reduce the long-term health and economic consequences associated with diabetes. However, traditional diagnostic processes, while effective, can be resource-intensive and reactive rather than preventive.
In recent years, the proliferation of big data in healthcare and advancements in computational technologies have paved the way for data-driven approaches to disease prediction and management. One such approach is **machine learning (ML)**, a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions without being explicitly programmed. ML techniques have proven highly effective in numerous healthcare applications, including disease classification, risk factor analysis, patient outcome forecasting, and treatment optimization.
In the context of diabetes, ML models can analyze a combination of clinical indicators and behavioral factors to predict whether an individual is likely to be diabetic. This can assist clinicians in making faster, more accurate, and potentially earlier diagnoses, improving patient outcomes and reducing healthcare system strain.

Objectives of the Study

This study is designed to explore how machine learning algorithms can be applied to the task of diabetes prediction. The key objectives are as follows:

- To perform data preprocessing and exploratory analysis on a dataset containing clinical and lifestyle attributes relevant to diabetes.
- To implement and train multiple supervised learning models, including both simple and complex classification algorithms.
- To evaluate and compare the models based on standard performance metrics such as accuracy, precision, recall, and F1-score.
- To identify the most effective and reliable model for predicting diabetes.

By addressing these objectives, this study contributes to the growing body of research focused on applying artificial intelligence to improve healthcare diagnostics and support preventive medicine initiatives.

2. Literature Review

Machine learning has become an essential tool in healthcare analytics, especially for chronic disease prediction such as diabetes. Among the many techniques available, logistic regression has been widely adopted due to its simplicity, interpretability, and proven effectiveness in binary classification problems.
A study titled *Prediction of Type 2 Diabetes using Logistic Regression Techniques* demonstrated that logistic regression could accurately predict diabetes status based on basic clinical features such as glucose levels, BMI, and age. The model achieved an accuracy rate of 82%, confirming its suitability for early diabetes detection using routinely collected patient data (ResearchGate, 2024). This reinforces logistic regression as a strong baseline model for structured clinical datasets.

Another relevant study, *Diabetes Prediction using Machine Learning Algorithms*, explored a more advanced approach by integrating traditional clinical features with external lifestyle and hereditary factors. The researchers utilized big data analytics and a comprehensive machine learning pipeline to uncover hidden patterns in large-scale health records. Their model showed improved classification performance compared to traditional approaches, emphasizing the value of feature integration and algorithmic enhancement for predictive accuracy (ScienceDirect, 2020).

Together, these studies demonstrate the growing potential of machine learning in diabetes prediction, from interpretable linear models to more complex data-driven pipelines. However, a critical gap exists in the literature—a lack of comparative studies that evaluate multiple machine learning algorithms side-by-side using the same dataset. Most studies focus on a single model or methodology, limiting insights into which techniques perform best under similar conditions.

This report addresses that gap by systematically comparing five widely used supervised learning algorithms—Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and XGBoost—on a dataset that includes both clinical and behavioral features. This approach offers a more comprehensive understanding of each model's strengths and suitability for real-world diabetes prediction tasks.

References

- ResearchGate. (2024). *Prediction of Type 2 Diabetes using logistic regression techniques*. https://www.researchgate.net/publication/377406725
- ScienceDirect. (2020). *Diabetes Prediction using Machine Learning Algorithms*. https://www.sciencedirect.com/science/article/pii/S1877050920300557

3. Theory and Methodology

3.1 Theoretical Foundations

This section outlines the theoretical basis of the machine learning algorithms used in this study. Each algorithm represents a different approach to supervised classification, offering unique strengths depending on data structure and distribution.

Logistic Regression
Logistic Regression is a type of generalized linear model commonly used for binary classification tasks. Instead of predicting actual values, it estimates the probability that a given input belongs to a particular class. This is achieved using the logistic (sigmoid) function, which transforms the linear combination of input features into values between 0 and 1. It is particularly effective when the relationship between input features and the target variable is approximately linear.

Decision Tree Classifier
A Decision Tree is a non-parametric model that splits the dataset into subsets based on the value of input features. It constructs a tree-like structure where each internal node represents a decision based on a feature, each branch represents an outcome, and each leaf node corresponds to a class label. Decision trees are intuitive, easy to interpret, and can model non-linear relationships. However, they can be prone to overfitting if not properly pruned.

K-Nearest Neighbors (KNN)
KNN is a simple instance-based learning algorithm that classifies new data points by looking at the 'k' closest training examples in the feature space. The class most common among the neighbors is assigned to the new data point. While KNN is easy to understand and implement, its performance heavily depends on the choice of 'k' and the scaling of input features. It can also become computationally expensive with large datasets.

Support Vector Classifier (SVC)
SVC is a powerful classification algorithm that works by finding the optimal hyperplane that separates data points of different classes. It uses support vectors—data points that lie closest to the decision boundary—and maximizes the margin between classes. SVC can be extended to non-linear problems by kernel functions, which project data into higher-dimensional spaces where linear separation is possible.

XGBoost
Extreme Gradient Boosting (XGBoost) is an ensemble learning technique that builds a strong classifier from multiple weak learners, typically decision trees. It uses a gradient boosting framework that adds new models sequentially to correct the errors made by previous ones. XGBoost is known for its efficiency, regularization capability, and superior performance in many classification problems, especially on structured tabular data.

3.2 Methodology Steps

The process followed in this study can be outlined in several key stages:

1. **Data Loading and Cleaning**
   The dataset was first loaded into a pandas DataFrame. A thorough check was performed to identify and handle any missing or inconsistent values. Fortunately, the dataset was found to be complete, requiring minimal cleaning.
2. **Splitting Data into Training and Testing Sets**
   The dataset was divided into training (70%) and testing (30%) subsets using a stratified approach to preserve the class distribution in both sets. This ensures fair and representative model evaluation.
3. **Encoding Categorical Variables**
   To prepare the data for model training, categorical variables were transformed using OneHotEncoding. This method creates binary columns for each category, allowing algorithms to interpret categorical data numerically without imposing ordinal relationships.
4. **Scaling Numerical Variables**
   Numerical features such as Age, BMI, Fasting Blood Sugar (FBS), and HbA1c were scaled using StandardScaler. This transformation standardizes the values so that they have a mean of 0 and a standard deviation of 1, improving the performance of distance-based models like KNN and SVC.
5. **Model Training and Evaluation**
   Each of the five models—Logistic Regression, Decision Tree, KNN, SVC, and XGBoost—was trained on the training data. Predictions were made on the testing data, and performance was assessed using metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.
6. **Comparison of Model Performance**
   Finally, the results from all models were compiled and compared. This comparison allowed us to identify the most effective model for predicting diabetes within the given dataset, considering both predictive accuracy and practical deployment factors.

4. Data

4.1 Dataset Overview

- Records: 128
- Features: 10 predictor variables, 1 target variable (Diagnosis)
- Source: Diabetes-classification-dataset from Kaggle

4.2 Variable Descriptions

Feature	Type	Description
Age	Numeric	Age of individual
Gender	Categorical	Male or Female
BMI	Numeric	Body Mass Index
Blood Pressure	Categorical	Normal, High, or Low
FBS	Numeric	Fasting Blood Sugar
HbA1c	Numeric	Hemoglobin A1c level
Family History of Diabetes	Categorical	Yes or No
Smoking	Categorical	Yes or No
Diet	Categorical	Poor or Healthy
Exercise	Categorical	Regular or No
Diagnosis	Target	Yes (Diabetic), No (Non-Diabetic)

Table 4.2.1: Description of variables

4.3 Data Preprocessing

- No missing values were found.
- Numerical features were standardized.
- Categorical features were encoded using OneHotEncoding.

5. Exploratory Data Analysis (EDA)

5.1 Summary Statistics

	Age	**BMI**	**FBS**	**HbA1c**
count	128.000000	128.000000	128.000000	128.000000
mean	42.031250	35.359375	162.500000	7.887500
std	16.783915	14.981739	61.323975	2.146339
min	12.000000	10.000000	80.000000	5.000000
25%	28.000000	24.000000	120.000000	6.400000
50%	40.000000	34.000000	160.000000	7.800000
75%	55.000000	45.500000	205.000000	9.375000
max	75.000000	67.000000	280.000000	12.000000

1-*Table 5.1.1: Table of Summary Statistics*

The dataset comprises 128 observations, providing insights into four key variables: Age, BMI, Fasting Blood Sugar (FBS), and HbA1c levels.

Age: The average age of participants is 42 years, with a standard deviation of 16.78, indicating substantial variation. The youngest participant is 12 years old, while the oldest is 75 years. The median age is 40 years.

BMI: The dataset reflects a mean BMI of 35.36, with values ranging from 10 to 67. The standard deviation of 14.98 suggests notable variability in body mass distribution.

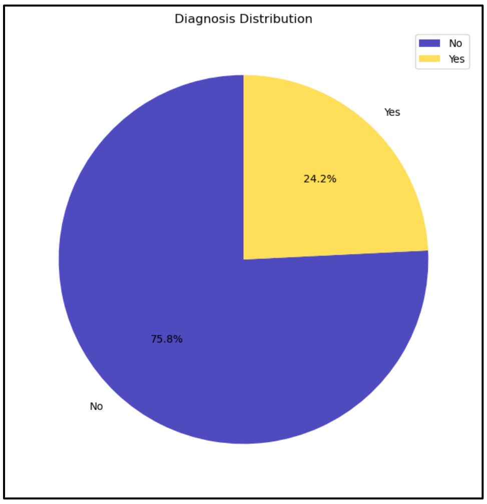
Fasting Blood Sugar (FBS): The average FBS level is 162.5 mg/dL, with a standard deviation of 61.32. The lowest recorded FBS value is 80 mg/dL, while the highest reaches 280 mg/dL, signifying a broad range in fasting glucose levels.

HbA1c: The mean HbA1c value is 7.89%, with a minimum of 5% and a maximum of 12%. The standard deviation of 2.15 indicates variability in long-term blood sugar levels.

These summary statistics highlight key trends and variability across the dataset, providing valuable insight into the distribution of demographic and health-related metrics.

5.2 Univariate Data Analysis

5.2.1. Diagnosis Distribution



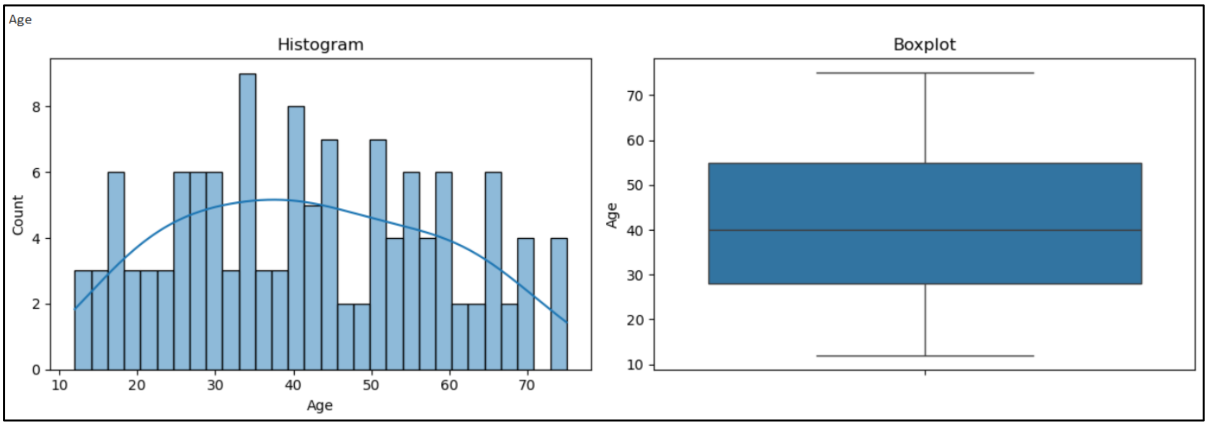
The pie chart represents the distribution of diagnoses in the dataset. The majority of individuals (75.8%) fall under the "No" category, indicating that they have not been diagnosed with the condition being analyzed. Conversely, 24.2% are classified as "Yes," representing the proportion of diagnosed cases.

Diabetic: 31 (24.2%)
non-Diabetic: 97 (75.8%)

1-*Figure 5.2.2.1: Distribution of Diagnosis*

5.2.2. Distributions of Quantitative Variables

1. Age

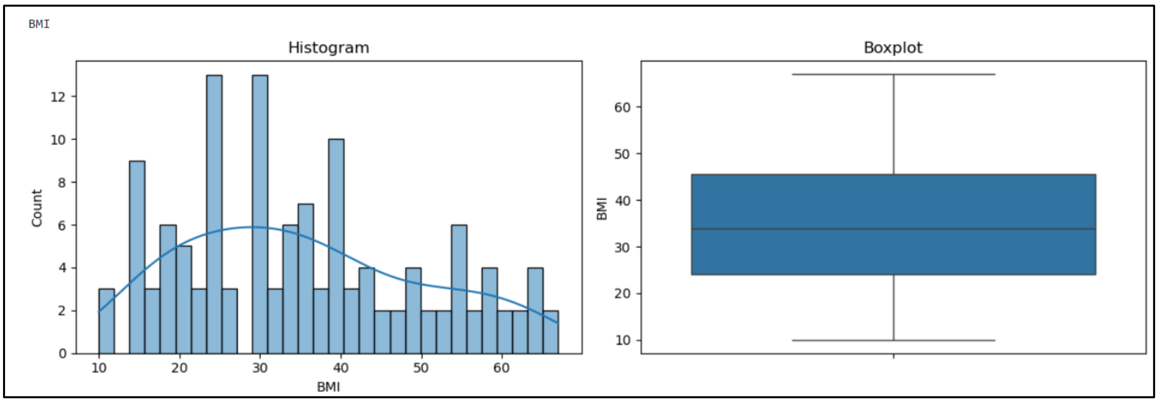


2-*Figure 5.2.2.1: Distributions of Age*

- **Histogram**: The age distribution spans approximately 12 to 75 years. The data appears fairly uniform, with a slight concentration in the 30–40 age range. The KDE line suggests smooth, non-skewed distribution.
- **Boxplot**:
    - Median: Around 40 years
    - Interquartile Range (IQR): ~27 (Q1) to ~55 (Q3)
    - Minimum and Maximum: ~12 and ~75
    - No visible outliers were detected.

Interpretation:
The dataset shows a broad and balanced distribution of ages, indicating a diverse sample. The lack of skewness and outliers suggests the age variable is suitable for further statistical analysis without the need for outlier handling.

2. BMI

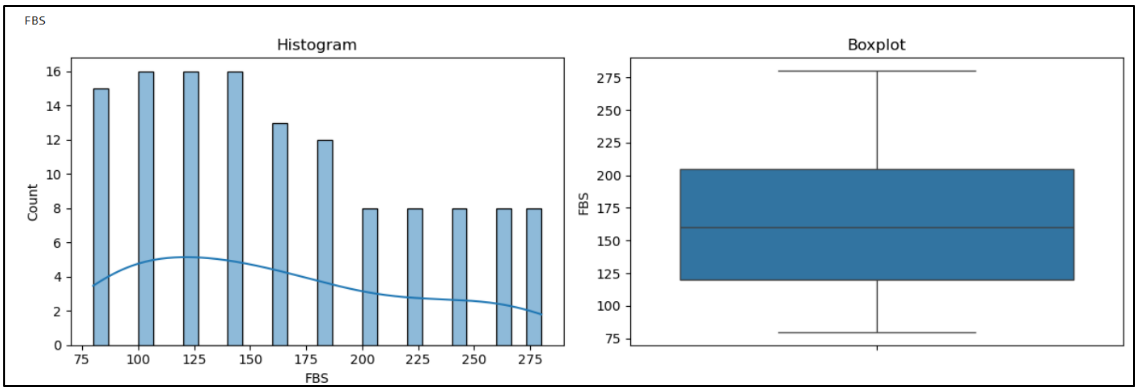


3-Figure 5.2.2.2: Distributions of BMI

- **Histogram**:
  BMI values range from about 10 to 67. The distribution is moderately right skewed, with a higher concentration of individuals between 20 and 40 BMI. The KDE curve supports this skew, showing a gradual decline beyond 40.
- **Boxplot**:
  - Median BMI: ~33
  - IQR: ~24 (Q1) to ~45 (Q3)
  - Range: ~10 to ~67
  - No extreme outliers are visible.

Interpretation:
The dataset includes a broad range of BMI values, with most individuals falling into the overweight to obese categories. The slight right skew indicates a tail of higher BMI values, though no extreme outliers are present. This distribution is suitable for analysis but may benefit from transformations.

3. FBS



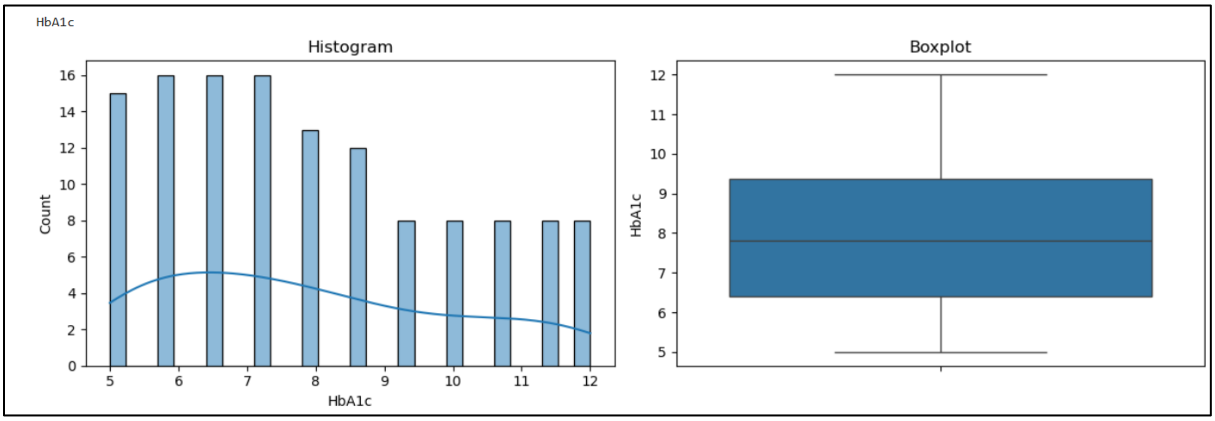
4-Figure 5.2.2.3: Distributions of FBS

- **Histogram**:
  The FBS values range from approximately 80 to 280 mg/dL. The distribution is right-skewed, with more frequent values between 90 and 160, then gradually tapering off at higher levels. This suggests a concentration of individuals with moderately elevated FBS levels, common in prediabetic or diabetic populations.

- **Boxplot**:
  - Median FBS: ~145 mg/dL
  - IQR: ~115 (Q1) to ~205 (Q3)
  - Range: ~80 to ~280 mg/dL
  - No strong outliers are visible, although the upper range is stretched, reflecting the skew.

Interpretation:
The FBS distribution indicates a significant portion of individuals with elevated blood sugar levels. While most data lie within a moderate range, the extended upper tail highlights potential cases of uncontrolled hyperglycemia. This variable may require normalization or transformation for statistical modeling due to its skewness.

4. HbA1c



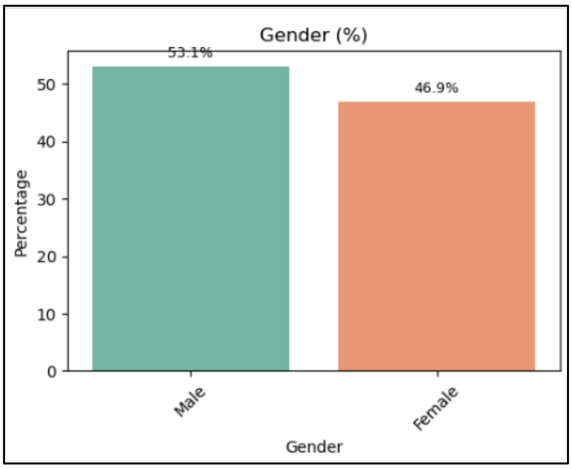
5-*Figure 5.2.2.4: Distributions of HbA1c*

- **Histogram**:
  HbA1c values range from approximately 5% to 12%. The distribution is right-skewed, with a higher concentration of values between 5% and 8%. The frequency declines gradually after 8%, indicating fewer individuals with very high HbA1c levels.
- **Boxplot**:
  - **Median HbA1c**: ~7.8%
  - **IQR**: ~6.5% (Q1) to ~9.3% (Q3)
  - **Range**: ~5% to ~12%
  - No outliers are apparent.

Interpretation:
Most individuals in the dataset fall within or above the prediabetic threshold (≥5.7%). The distribution suggests a high prevalence of abnormal glycemic control, with a noticeable tail toward higher HbA1c levels. While the data is skewed, it remains within a manageable range for analysis and may benefit from transformation into predictive modeling.

5.2.3. Distributions of Qualitative Variables

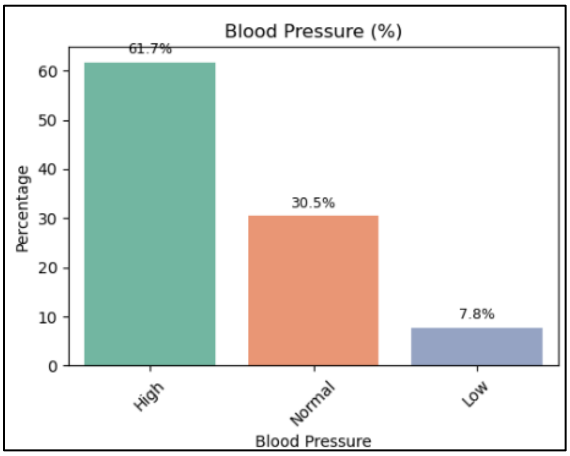
1. Gender



6-Figure 5.2.3.1: Distribution of Gender

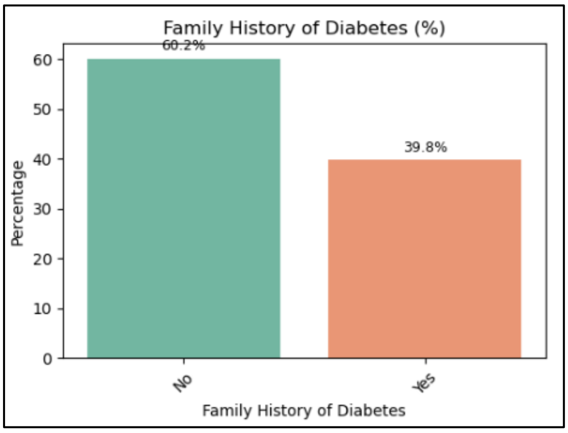
The dataset has a fairly balanced gender composition, with 53.1% males and 46.9% females. This near-even distribution indicates a representative sample in terms of gender, which helps minimize gender-based bias in subsequent analyses. It also allows for meaningful comparison of gender-related health patterns such as prevalence of diabetes, blood pressure levels, and lifestyle behaviors.

2. Blood Pressure



7-Figure 5.2.3.2: Distribution of BP

Blood pressure status shows a notable skew, with 61.7% of individuals exhibiting high blood pressure, 30.5% with normal levels, and only 7.8% with low blood pressure. The dominance of high blood pressure is clinically significant, as it is a key risk factor for cardiovascular disease and diabetes. This suggests that hypertension is prevalent in the study population and should be considered a major factor in risk modeling and health intervention planning.

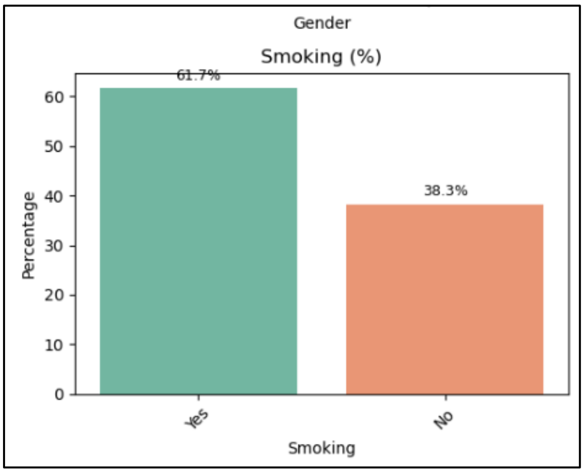


8-Figure 5.2.3.3: Distribution of Family History of Diabetes

3. Family History of Diabetes

Approximately 39.8% of individuals have a family history of diabetes, while the remaining 60.2% do not. This indicates a substantial genetic predisposition within the population. Family history is a known non-modifiable risk factor, and its presence in nearly 40% of the sample emphasizes the importance of early screening and preventive measures for high-risk individuals.

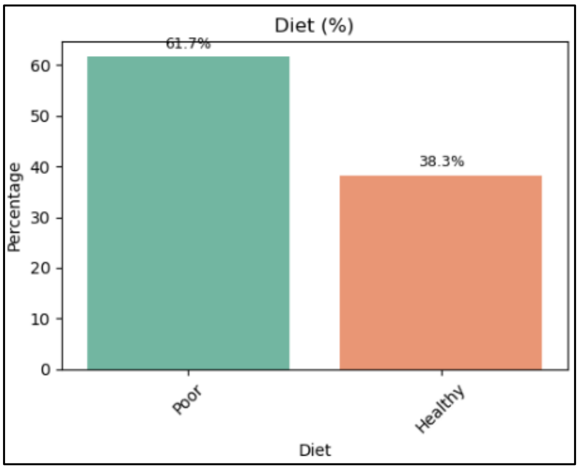
4. *Smoking Status*



9-*Figure 5.2.3.4: Distribution of Smoking Status*

A concerning 61.7% of participants are smokers, with only 38.3% non-smokers. This high rate of smoking suggests a major public health concern, given its association with numerous chronic diseases including diabetes, cardiovascular conditions, and respiratory disorders. The variable may also have strong interactions with other lifestyle factors such as exercise and diet, warranting further investigation.

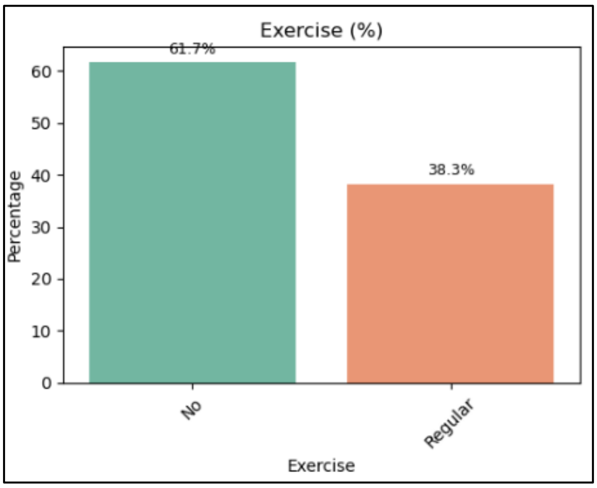
5. *Diet Quality*



10-*Figure 5.2.3.5: Distribution of Diet quality*

Diet habits lean heavily toward unhealthy choices, with 61.7% of individuals consuming a poor diet, compared to only 38.3% who follow a healthy one. Poor diet is a leading contributor to obesity, insulin resistance, and elevated blood glucose levels. This finding highlights the need for nutritional education and dietary interventions to mitigate health risks in the population

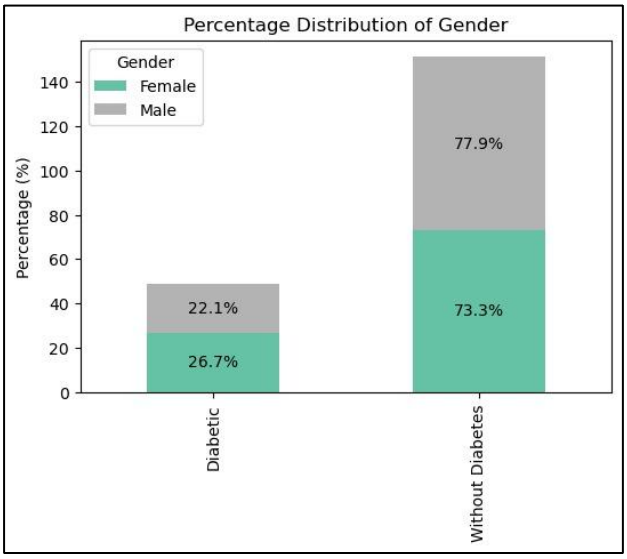
6. *Physical Activity*



11-*Figure 5.2.3.6: Distribution of Physical activity*

Exercise habits are similarly unfavorable, with 61.7% of individuals reporting no regular physical activity, while only 38.3% engage in regular exercise. This sedentary behavior likely contributes to the high incidence of other risk factors observed, such as obesity, hypertension, and poor glycemic control. Promoting active lifestyles could be a key strategy in improving overall health outcomes.

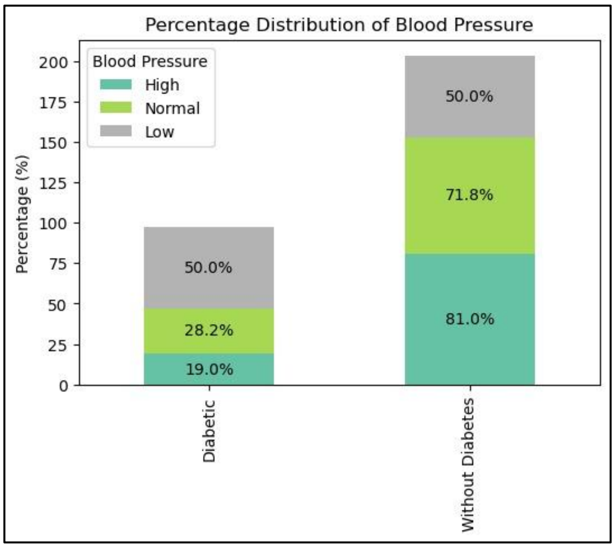
5.3 Bivariate Data Analysis



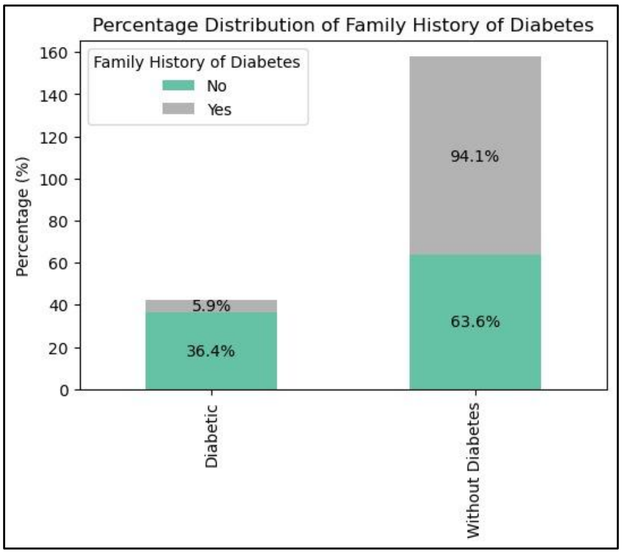
12-Figure 5.3.1: Gender Distribution by Diabetes Status (%)

The chart shows the gender distribution among diabetic and non-diabetic individuals. Among females, 73.3% are not diagnosed with diabetes while 26.7% are diagnosed with diabetes. Out of males, 22.1% are diagnosed with diabetes and 77.9% are not. This indicates that most individuals in the dataset do not have diabetes, and the gender distribution is fairly balanced within each group.

This is distribution of blood pressure categories among diabetic and non-diabetic individuals. Among those with low blood pressure, 50% are diagnosed with diabetes and 50% are not. In the normal blood pressure group, 28.2% are diabetic while 71.8% are not. Among individuals with high blood pressure, 19% are diabetic and 81% are not. This indicates that most individuals with high or normal blood pressure do not have diabetes, while individuals with low blood pressure are equally split between diabetic and non-diabetic.



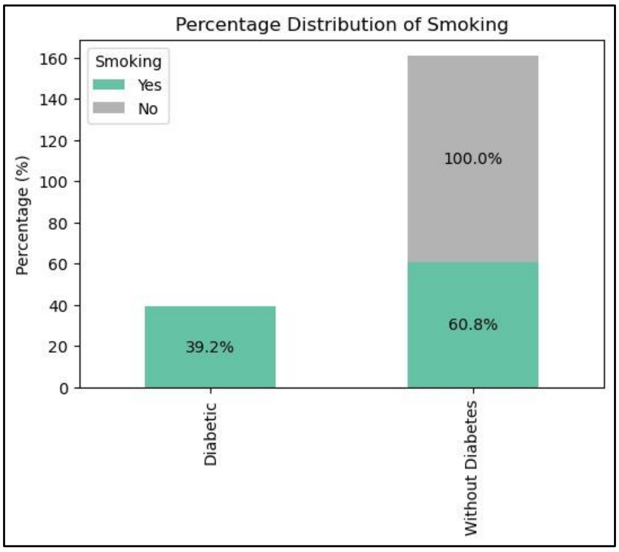
13-Figure 5.3.2: Blood Pressure Distribution by Diabetes Status (%)



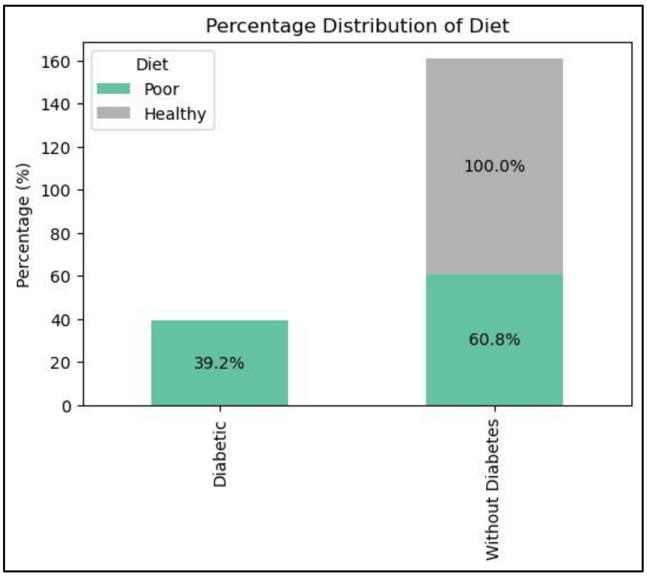
14-Figure 5.3.3: Family History of Diabetes Distribution by Diabetes Status (%)

The chart shows the distribution of family history of diabetes among diabetes and non-diabetic individuals. Among those with a family history, 5.9% are diabetic and 94.1% are not. In contrast, 36.4% of those without a family history are diabetic. This suggests that diabetes in the dataset may be influenced by factors beyond family history.

The chart shows that 39.2% of diabetics in the dataset are smokers, compared to 60.8% of non-diabetics. This indicates that smoking is more common among non-diabetics. The lower smoking rate among diabetics may suggest increased health awareness or medical advice received after diagnosis.

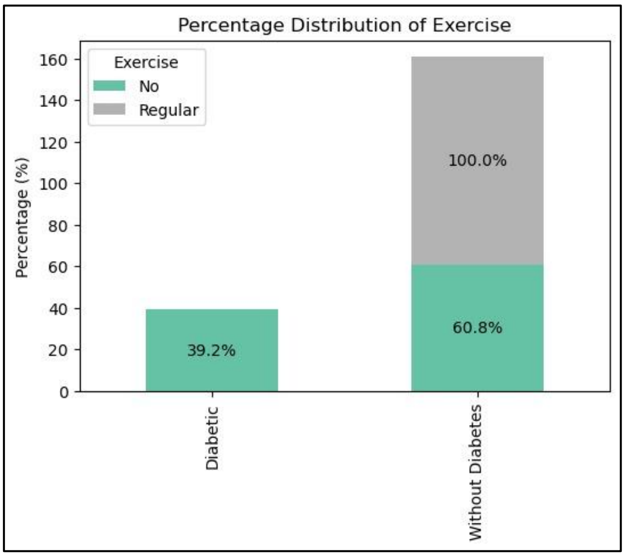


15-Figure 5.3.4: Smoking Distribution by Diabetes Status (%)



16-Figure 5.3.5: Diet Distribution by Diabetes Status (%)

The chart indicates that individuals with a healthy diet are predominantly non-diabetic. Among those with a poor diet, 60.8% do not have diabetes, while 39.2% are diabetic. This suggests that while a poor diet is common among non-diabetics, it may still contribute to an increased risk of developing diabetes. The data also implies that maintaining a healthy diet is associated with a lower likelihood of having diabetes.



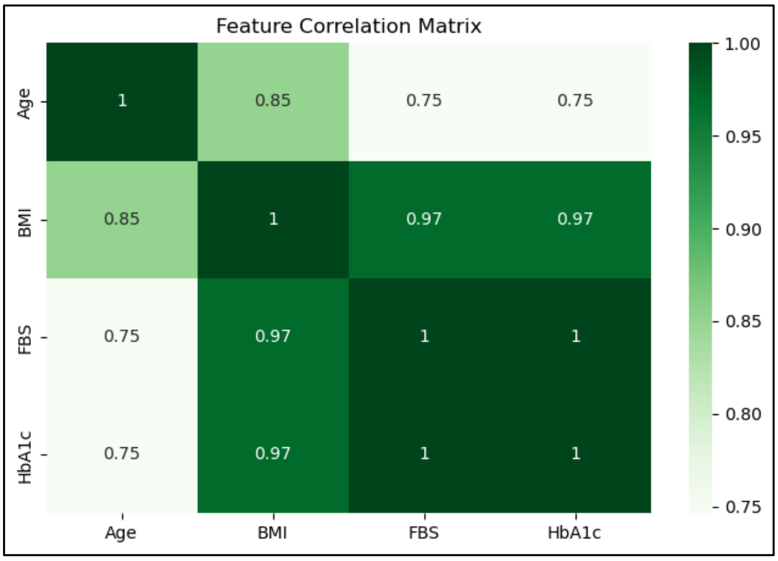
17-Figure 5.3.6: Exercise Distribution by Diabetes Status (%)

The chart indicates that individuals who exercise regularly are mostly non-diabetic. Among those who do not exercise, 60.8% are non-diabetic while 39.2% are diabetic. This suggests that regular physical activity is associated with a lower likelihood of diabetes, highlighting the potential role of exercise in prevention of diabetes and management.

The matrix shows strong positive correlations among all variables:

- BMI is highly correlated with FBS (0.97) and HbA1c (0.97), suggesting that higher body mass is strongly linked to increased blood sugar levels.

- FBS and HbA1c have a perfect correlation (1.00), indicating a close relationship between short-term and long-term glucose levels.



18- Figure 5.3.7: Correlation matrix of quantitative variables

- Age shows moderate to strong correlations with BMI (0.85), FBS (0.75), and HbA1c (0.75), implying that these health indicators tend to worsen with age.

Overall, the variables are closely related, highlighting the interconnected nature of age, weight, and glucose regulation.

6. Advanced Analysis

6.1 Scaling Numeric Features

Before fitting a machine learning model, it is important to scale numeric features to ensure they are on a similar range. This helps algorithms that are sensitive to feature magnitudes—such as logistic regression, SVM, and KNN—perform better and converge faster. In this analysis, the numeric features Age, BMI, FBS, and HbA1c were scaled to standardize their values, improve model accuracy, and avoid bias caused by differing units or value ranges.

```
        Age  Gender       BMI Blood Pressure       FBS     HbA1c  \
0  0.177576    Male -0.694184         Normal -1.023182 -1.023182
1  0.775725  Female -0.359133           High -0.695764 -0.695764
2  1.373875    Male -0.024082           High -0.368345 -0.368345
3  1.972025  Female  0.310969           High -0.040927 -0.040927
4 -0.121499    Male -1.029235         Normal -1.350600 -1.350600

  Family History of Diabetes Smoking     Diet Exercise Diagnosis
0                         No      No  Healthy  Regular        No
1                        Yes     Yes     Poor       No       Yes
2                        Yes     Yes     Poor       No       Yes
3                        Yes     Yes     Poor       No       Yes
4                         No      No  Healthy  Regular        No
```

6.2 Categorical Variable Encoding

Before training machine learning models, categorical variables must be converted into a numerical format since most algorithms cannot interpret text-based data. One effective method is One-Hot Encoding, which creates binary columns for each category, ensuring that no ordinal relationship is assumed between them.

In this analysis, categorical features such as Gender, Blood Pressure, Family History of Diabetes, Smoking, Diet, and Exercise were transformed using One-Hot Encoding. This process generated separate binary columns for each category, allowing the models to treat these variables appropriately during training. The original categorical columns were then removed to avoid redundancy and ensure clean input data.

```
One-Hot Encoded Data:
   Gender_Male  Blood Pressure_Low  Blood Pressure_Normal  \
0          1.0                 0.0                    1.0
1          0.0                 0.0                    0.0
2          1.0                 0.0                    0.0
3          0.0                 0.0                    0.0
4          1.0                 0.0                    1.0

   Family History of Diabetes_Yes  Smoking_Yes  Diet_Poor  Exercise_Regular
0                             0.0          0.0        0.0               1.0
1                             1.0          1.0        1.0               0.0
2                             1.0          1.0        1.0               0.0
3                             1.0          1.0        1.0               0.0
4                             0.0          0.0        0.0               1.0
```

6.3 Model Training

To predict diabetes status, five machine learning models were trained on the processed dataset: Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and XGBoost. These models were selected to represent a variety of learning approaches—linear, tree-based, instance-based, and ensemble methods—for a well-rounded comparison.

The dataset was first split into training and testing sets (70:30 ratio) to ensure unbiased evaluation. Numeric features were then scaled, and categorical variables were encoded using One-Hot Encoding, with all transformations fit on the training set and applied to both training and test sets. Model evaluation was based on key classification metrics, including accuracy, precision, recall, and F1-score.

6.4 Model Evaluation

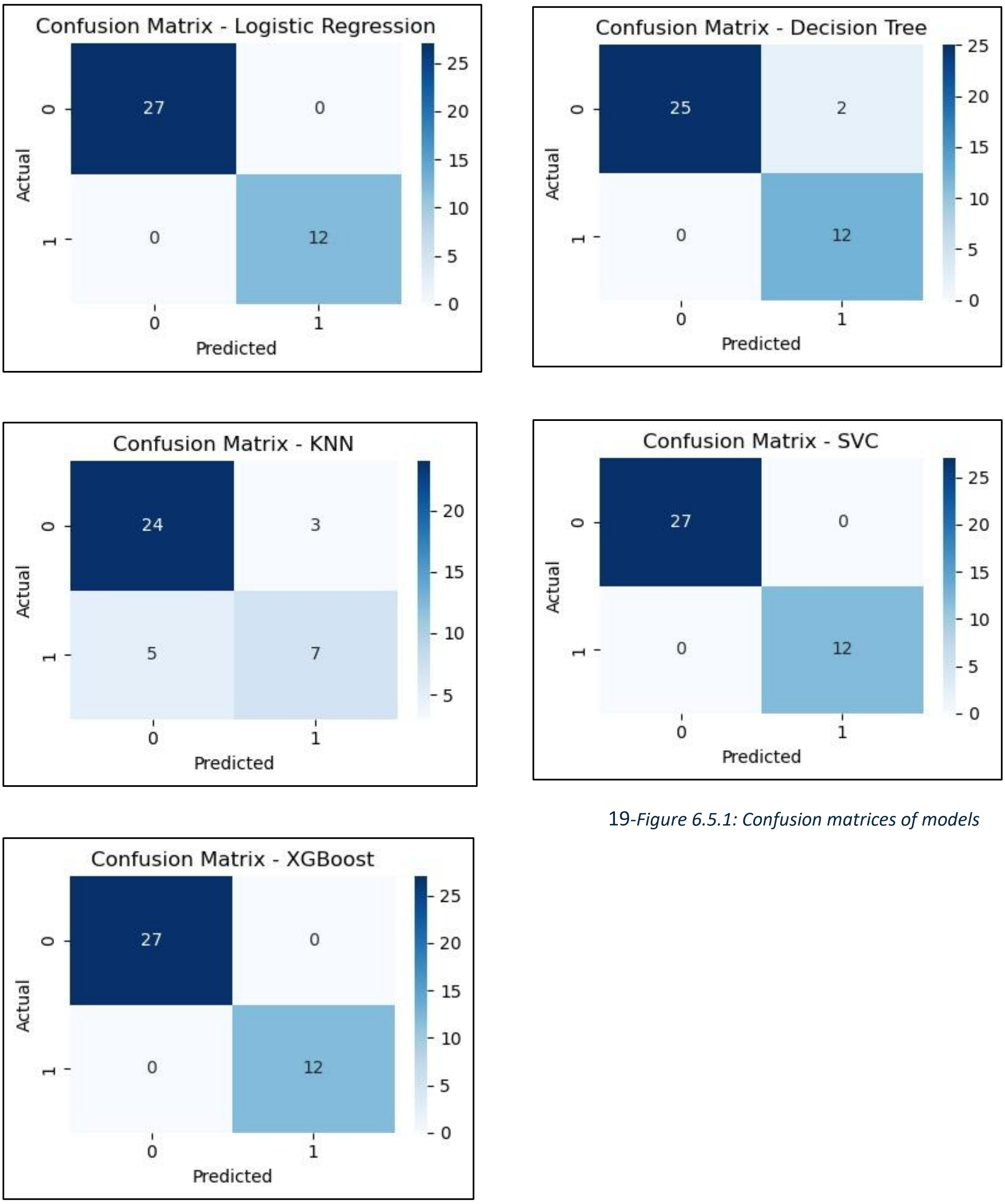
The performance of all trained models was evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. These metrics provide a balanced view of each model's ability to correctly classify outcomes, especially in the presence of class imbalance. Evaluation was performed on the test set to assess how well each model generalizes to unseen data. This comparison helped identify the most effective model for the given problem.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	1.00	1.00	1.00	1.00
Decision Tree	0.95	0.86	1.00	0.92
KNN	0.79	0.70	0.58	0.64
SVC	1.00	1.00	1.00	1.00
XGBoost	1.00	1.00	1.00	1.00

2-*Table 6.4.1: Table of model evaluation*

6.5 Confusion Matrices

A confusion matrix is a table used to evaluate the performance of a classification model. It shows the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This helps in understanding not just overall accuracy, but also how well the model distinguishes between classes. From the confusion matrix, key metrics like precision, recall, and F1-score can be derived, offering deeper insights into the model's strengths and weaknesses.



19-*Figure 6.5.1: Confusion matrices of models*

7. General Discussion and Conclusion

Key Findings

Logistic Regression and Support Vector Classifier (SVC):

- Achieved perfect scores in accuracy, precision, recall, and F1-score.
- Performed well with minimal parameter tuning.
- Demonstrated strong generalization on the available data.

XGBoost:

- Also achieved perfect classification results.
- Effective due to its robust ensemble learning mechanism.
- May require more tuning in larger or more complex datasets.

K-Nearest Neighbors (KNN):

- Recorded the lowest performance, especially in recall.
- Struggled to correctly identify diabetic cases.
- Sensitive to feature scaling and not well-suited for small datasets with complex boundaries.

Strengths and Limitations

Strengths:

- Dataset was complete, with no missing values—ensuring clean preprocessing.
- Included both clinical (e.g., HbA1c, FBS) and behavioral (e.g., diet, exercise) variables for comprehensive analysis.
- Employed multiple evaluation metrics (accuracy, precision, recall, F1-score) for well-rounded model assessment.
- Compared diverse algorithm types (linear, tree-based, instance-based, ensemble) for broader insight.

Limitations:

- Small sample size (128 records), which may not capture population variability.
- Perfect scores may indicate overfitting, especially without cross-validation.
- No use of an external test dataset to validate the results.
- Feature importance or interpretability was not explored in depth.

Recommendations

Model Use:

- Prefer SVC or Logistic Regression for similar small-scale healthcare datasets due to ease of use and strong results.
- Use XGBoost if more complex modeling or ensemble methods are justified by data scale and resources.

Future Improvements:

- Implement cross-validation to improve model robustness and reduce overfitting risk.
- Apply feature selection techniques to simplify models and enhance interpretability.
- Expand the dataset with more records and diversity for better generalization.
- Validate model performance using external datasets for more reliable deployment in real-world settings.

8. References

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
  https://doi.org/10.1145/2939672.2939785

- Tripathy, N., Nayak, S. K., Moharana, B., Pati, A., Balabantaray, S. K., & Panigrahi, A. (2024). *A Comparative Analysis of Diabetes Prediction Using Machine Learning and Deep Learning Algorithms in Healthcare*. Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar.
  https://www.researchgate.net/publication/380212284

- Nuthakki, P., & Pavankumar, T. (2024). Comparative assessment of machine learning algorithms for effective diabetes prediction and care. *International Journal of Computational and Experimental Science and Engineering (IJCESEN)*, 10(4), 1337–1343.
  https://doi.org/10.22399/ijcesen.606

- Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *[Journal Name if available]*. Available online February 20, 2021.
  https://doi.org/10.1016/j.icte.2021.02.004

- World Health Organization (WHO). (2023). Diabetes fact sheet.
  https://www.who.int/news-room/fact-sheets/detail/diabetes