A DATA EXPLORATION & VISUALIZATION LAB
MINI PROJECT REPORT ON

DIABETES PREDICTION FOR WOMEN

SUBMITTED TO THE PCET'S PIMPRI CHINCHWAD COLLEGE OF ENGINEERING IN
THE PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR

SECOND YEAR
OF

BACHELOR OF TECHNOLOGY (COMPUTER ENGINEERING)

SUBMITTED BY

PRANAV CHAUDHARI	Roll No : 124B1B027
TANMAY TALREJA	Roll No : 124B1B034
GAURAV JAYPATRE	Roll No : 124B1B025
DIGVIJAY BIRAJDAR	Roll No : 124B1B013

DEPARTMENT OF COMPUTER ENGINEERING

PCET'S PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

Sector No. 26, Pradhikaran, Nigdi, Pimpri-Chinchwad, PUNE 411044
An Autonomous Institute Approved by AICTE and Affiliated to SPPU, Pune

CERTIFICATE

This is to certify that the seminar report entitles

“DIABETES PREDICTION FOR WOMEN”

Submitted by

PRANAV CHAUDHARI	**Roll No : 124B1B027**
TANMAY TALREJA	**Roll No : 124B1B034**
GAURAV JAYPATRE	**Roll No : 124B1B025**
DIGVIJAY BIRAJDAR	**Roll No : 124B1B013**

are bonafide student of this institute and the work has been carried out by them under the supervision of **Sonika Gill** and it is approved for the partial fulfillment of the requirement of PCET’S PIMPRI CHINCHWAD COLLEGE OF ENGINEERING, for the award of Second Year of **Bachelor of Technology** (Computer Engineering) **A.Y.-2025-26**.

(Prof. Sonika Gill)
Guide
Department of Computer Engineering

(Prof. Dr. Sonali Patil)
Head,
Department of Computer Engineering

Place : Pune
Date : 3/11/2025

ACKNOWLEDGEMENT

We express our sincere thanks to our **Guide Prof. Sonika Gill** for her constant encouragement and support throughout our technical seminar, especially for the useful suggestions given during the course and having laid down the foundation for the success of this work.

We would like to thank **Prof. Dr. Sonali Patil, Head of Computer Department** for her unwavering support during the entire course of this seminar work. We are very grateful to our **Director, Prof. Dr. G.N. Kulkarni** for providing us with an environment to complete our seminar work successfully. We also thank all the staff members of our college and technicians for their help in making this seminar a success.

Finally, we take this opportunity to extend our deep appreciation to our family and friends, for all that they meant to us during the crucial times of the completion of our technical seminar.

PRANAV CHAUDHARI
TANMAY TALREJA
GAURAV JAYPATRE
DIGVIJAY BIRAJDAR

ABSTRACT

Diabetes mellitus is one of the most serious and widespread health concerns in the world today, affecting millions of people regardless of age or background. Early detection and continuous monitoring are extremely important for managing diabetes and preventing life-threatening complications. However, traditional diagnostic methods often require laboratory tests, which can be time-consuming, expensive, and not easily accessible in rural or underdeveloped areas. This challenge inspired us to explore how machine learning could be used to predict diabetes efficiently using easily available medical data.

In this project, we developed a machine learning-based predictive model that determines the likelihood of a person having diabetes using a set of measurable health parameters such as glucose level, blood pressure, insulin, BMI, age, and others. We used the Pima Indians Diabetes Dataset as our primary data source. The dataset underwent several preprocessing steps such as handling missing and zero values, normalizing the features using StandardScaler, and balancing the data with the SMOTE technique to prevent bias. We also performed Exploratory Data Analysis (EDA) using Matplotlib and Seaborn to visualize relationships between features and to identify which factors had the most influence on the outcome.

After extensive testing and evaluation, we implemented and compared multiple models, including Random Forest, Logistic Regression, and XGBoost. The XGBoost model achieved the highest accuracy of around 94% after hyperparameter tuning. This model was then used to make predictions based on user input values. Our project demonstrates how artificial intelligence and data science can contribute to the healthcare industry by providing accurate, data-driven insights that assist in early disease detection. With further improvements, such models can be integrated into health applications or clinical decision-support systems to make diabetes prediction faster, easier, and more accessible to everyone.

TABLE OF CONTENTS

INTRODUCTION

1.1 OVERVIEW

Our project, “Diabetes Prediction using Machine Learning,” aims to predict whether a person is likely to have diabetes based on key medical factors such as glucose level, blood pressure, BMI, age, and insulin level. Using the Pima Indians Diabetes Dataset, we applied data preprocessing techniques like handling missing values, removing outliers, and feature scaling to improve model accuracy.

We performed Exploratory Data Analysis (EDA) using Matplotlib and Seaborn to understand feature relationships and correlations. The machine learning model was built using an optimized XGBoost classifier, which achieved an accuracy of around 94%. This project demonstrates how data-driven techniques can be used to assist in early detection and better management of diabetes through predictive analytics

1.2 MOTIVATION

The motivation behind this project comes from the growing concern over the rapid rise of diabetes cases worldwide. Many people remain unaware of their diabetic condition until serious health complications occur. We wanted to create a system that could help identify potential diabetes risk early using machine learning techniques. By analyzing health data and predicting outcomes efficiently, our goal is to support preventive healthcare and encourage individuals to take timely medical advice.

1.3 PROBLEM STATEMENT AND OBJECTIVES

1.3.1 Problem statement -

Diabetes is a chronic disease that affects millions of people globally, and its early detection remains a major challenge in healthcare. Traditional diagnostic methods can be time-consuming and may not always identify high-risk individuals in advance. Therefore, there is a need for an intelligent system that can accurately predict the likelihood of diabetes based on key medical parameters. This project aims to address this problem by developing a machine learning-based model that analyzes patient data and provides reliable predictions for early diagnosis and prevention.

1.3.2 Objectives -

1. To analyze and preprocess the diabetes dataset by handling missing values, removing outliers, and scaling features for better model performance.

2. To perform Exploratory Data Analysis (EDA) using visualization tools like Matplotlib and Seaborn to understand relationships among features.

3. To develop and train an accurate machine learning model (using XGBoost) that can effectively predict whether a person is diabetic or not.

4. To evaluate model performance using metrics such as accuracy, confusion matrix, and classification report.

5. To enable user interaction through a system that accepts health parameters as input and predicts the likelihood of diabetes in real time.

1.4 SCOPE

The scope of this project extends to developing a reliable and efficient machine learning model that can assist in predicting diabetes based on various health indicators. The system is designed to handle real-world data by cleaning, transforming, and analyzing it for accurate predictions. It can be further integrated into healthcare systems or mobile applications to provide quick and data-driven health assessments.

While this model currently focuses on the Pima Indians Diabetes Dataset, it can be expanded by incorporating larger and more diverse datasets for broader applicability. In the future, the system can also be enhanced with deep learning techniques or real-time data inputs from wearable health devices to improve prediction accuracy and usability

1.5 METHODOLOGIES OF PROBLEM SOLVING

1.5.1 Data Collection:

We used the Pima Indians Diabetes Dataset, which contains medical data such as glucose level, blood pressure, BMI, age, and insulin level of patients, along with their diabetes outcome.

1.5.2 Data Preprocessing:

The raw dataset contained missing and invalid values (like zeros in medical attributes). We handled these by replacing them with median values, removing outliers using the IQR method, and applying scaling for uniformity across features.

1.5.3 Exploratory Data Analysis (EDA):

Using Matplotlib and Seaborn, we visualized the dataset to identify trends, distributions, and correlations among features. This helped us understand which attributes had the strongest impact on diabetes prediction.

1.5.4 Feature Engineering:

We performed correlation analysis and log transformations on skewed data to improve model performance. Outliers were capped to reduce their negative influence.

1.5.5 Model Development:

We used the XGBoost Classifier, a powerful machine learning algorithm, to build our predictive model. It was chosen for its accuracy, robustness, and efficiency in handling structured data.

1.5.6 Model Evaluation:

The model was trained and tested using an 80:20 split, and evaluated using metrics such as accuracy, confusion matrix, and classification report. The final model achieved high accuracy after applying SMOTE for class balancing.

1.5.7 User Prediction System:

Finally, we implemented an interactive section that takes user inputs (like glucose, BMI, and age) and predicts whether the person is likely to have diabetes or not

2. LITERATURE SURVEY

Several research studies have been carried out in the field of diabetes prediction using machine learning. Earlier approaches mainly used traditional models such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM) on the Pima Indians Diabetes Dataset. These models achieved moderate accuracy levels, usually between 75–85%, but they struggled with issues like missing values, class imbalance, and limited interpretability.

Recent studies have shown that ensemble and boosting algorithms, especially XGBoost, provide better accuracy and handle complex data patterns more effectively. These advanced models have achieved accuracies above 90% by combining strong preprocessing methods such as data balancing and feature scaling.

Based on the findings from existing work, our project focuses on improving data quality through preprocessing, balancing the dataset using SMOTE, and implementing the XGBoost Classifier to enhance accuracy and reliability in predicting diabetes.

3.SOFTWARE REQUIREMENTS SPECIFICATIONS

3.1 SYSTEM REQUIREMENTS

The project is developed using Python programming language with the help of various data science and machine learning libraries. The system requires sufficient memory and processing power to handle dataset operations and model training.

3.1.1 DATABASE REQUIREMENTS

The project uses the **Pima Indians Diabetes Dataset (CSV format)**.
This dataset contains **768 records** and **9 attributes**, which are directly loaded using **Pandas** for data preprocessing and model training. No external database or server is required.

3.1.2 SOFTWARE REQUIREMENTS (PLATFORM CHOICE)

- Operating System: Windows 10 / 11 or Linux
- Programming Language: Python 3.8 or above
- Libraries Used: pandas, numpy, matplotlib, seaborn, scikit-learn, imbalanced-learn, xgboost
- IDE / Editor: Jupyter Notebook / VS Code / PyCharm
- Visualization Tools: Matplotlib and Seaborn for EDA
  Machine Learning Framework: Scikit-learn and XGBoost

3.1.3 HARDWARE REQUIREMENTS

- Processor: Intel i5 or higher
- RAM: Minimum 8 GB
- Storage: At least 1 GB free space
- GPU (Optional): Recommended for faster training (NVIDIA/AMD)
- Display: Standard HD resolution for visualization

4. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

4.1.1 Data Collection:

The dataset used is the Pima Indians Diabetes Dataset, which contains relevant medical parameters such as glucose, blood pressure, BMI, insulin, and age.

4.1.2 Data Preprocessing:

This step includes cleaning invalid values, handling missing data, removing outliers using IQR, and scaling the data using StandardScaler to improve model accuracy.

4.1.3 Exploratory Data Analysis (EDA):

Visualizations using Matplotlib and Seaborn help identify feature correlations and important patterns that affect diabetes outcomes.

4.1.4 Feature Engineering:

Includes transformation of skewed data (log transform on insulin), correlation analysis, and class balancing using SMOTE.

4.1.5 Model Training and Testing:

The processed data is divided into training and testing sets (80:20 split). The XGBoost Classifier is trained and tuned to predict diabetes with high accuracy.

4.1.6 Evaluation:

The model’s performance is evaluated using metrics like accuracy, confusion matrix, and classification report.

4.1.7 User Input and Prediction:

The final module allows users to input their health details (like glucose, BMI, age, etc.) and receive a prediction indicating whether they are likely to have diabetes.

5. IMPLEMENTATION

5.1 OVERVIEW OF PROJECT MODULES

5.1.1 Data Collection Module:

This module loads the Pima Indians Diabetes Dataset using pandas. The dataset includes medical parameters such as glucose, BMI, blood pressure, insulin, and age**.**

5.1.2 Data Preprocessing Module:

It cleans and prepares the dataset by replacing missing or invalid values, removing outliers using the IQR method, applying log transformations, and scaling features for better consistency.

5.1.3 Exploratory Data Analysis (EDA) Module:

In this module, the dataset is visualized using Matplotlib and Seaborn to identify correlations and understand which parameters most affect diabetes prediction**.**

5.1.4 Feature Engineering Module:

Handles feature transformations, correlation analysis, and data balancing using the SMOTE technique to improve model performance.

5.1.5 Model Training and Evaluation Module:

This module trains the XGBoost Classifier using the preprocessed data. The model is evaluated using accuracy, confusion matrix, and classification report to ensure its reliability.

5.1.6 User Prediction Module:

Allows users to enter their health parameters. The trained model then predicts whether the person is likely to have diabetes, based on the provided data**.**

5.2 TOOLS AND TECHNOLOGY USED

The project was implemented using Python 3.8 as the primary programming language due to its simplicity and powerful libraries for data analysis and machine learning. Libraries such as pandas and numpy were used for data handling, cleaning, and numerical computations, while matplotlib and seaborn were employed for data visualization and Exploratory Data Analysis (EDA). For building and evaluating the machine learning model, we used scikit-learn, XGBoost, and imbalanced-learn (SMOTE) to handle data imbalance and improve accuracy. The project was developed and tested using Jupyter Notebook and Visual Studio Code (VS Code) as the development environments. The dataset used was the Pima Indians Diabetes Dataset, stored in CSV format, which provided the medical parameters necessary for model training. The system was executed on a Windows operating system, although it can also be easily run on Linux platforms.

5.3 ALGORITHM DESCRIPTION

5.3.1 XGBOOST (EXTREME GRADIENT BOOSTING) CLASSIFIER

XGBoost, short for *Extreme Gradient Boosting*, is an advanced boosting algorithm designed for high performance and accuracy. It builds multiple small decision trees one after another, where each new tree attempts to correct the errors made by the previous ones. Over several iterations, this process helps the model become more accurate and robust. XGBoost minimizes errors using gradient descent and includes built-in regularization to reduce overfitting, which makes it more reliable than traditional boosting methods.

In our project, XGBoost analyzes medical attributes such as glucose level, blood pressure, insulin, and BMI to predict whether a person is diabetic. It is especially efficient for structured datasets like the Pima Indians Diabetes Dataset because it handles missing data and captures complex patterns effectively.

Why We Used XGBoost:

We selected XGBoost because it consistently delivered the best accuracy among all tested algorithms. It runs faster due to parallel processing, manages missing values automatically, and provides strong generalization on unseen data. The regularization parameters also help control overfitting, ensuring our model performs well in real-world cases. Overall, it was the most suitable choice for a reliable and efficient diabetes prediction system.

Well Known For:

- Exceptional speed and scalability
- High accuracy and robustness in structured/tabular data
- Built-in feature importance and regularization
- Excellent handling of missing and imbalanced data
- Widely used in data science competitions and healthcare predictive models

5.3.2 RANDOM FOREST CLASSIFIER

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of the data and features, and the final prediction is made by taking the majority vote from all the trees. This randomness makes the model more stable and less sensitive to noise in the data.

In our project, we initially used the Random Forest classifier as a baseline model to understand how multiple decision trees could be combined to predict diabetes. It helped us establish an early performance benchmark before switching to XGBoost for optimization.

Why We Used Random Forest:

We used Random Forest because it is simple, reliable, and performs well without much parameter tuning. It helped us understand the relationship between health parameters and diabetes outcomes, and provided a solid comparison point for more complex models like XGBoost. It also handles missing values well and reduces the risk of overfitting compared to single decision trees.

Well Known For:

- Stability and robustness against noise
- Good accuracy with minimal tuning
- Handles large datasets and high-dimensional features effectively
- Provides feature importance for better interpretability
- Works well for both classification and regression problems

5.3.3 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is a data balancing technique used to handle imbalanced datasets, where one class (for example, non-diabetic) has significantly more samples than the other (diabetic). Instead of simply duplicating minority samples, SMOTE creates new synthetic samples by interpolating between existing ones. This helps the model learn better from the underrepresented class and improves prediction accuracy for minority outcomes.

In our project, we applied SMOTE before model training to ensure that the number of diabetic and non-diabetic samples was balanced. This improved the model’s ability to correctly identify diabetic patients, rather than being biased toward predicting the majority class.

Why We Used SMOTE:

We used SMOTE to handle the class imbalance present in the diabetes dataset. Without balancing, the model might have learned to favor the non-diabetic class, reducing accuracy for diabetic predictions. Using SMOTE ensured fair training and improved the model’s recall and precision for diabetic cases.

Well Known For:

- Balancing datasets by generating synthetic samples
- Improving classifier performance on minority classes
- Preventing model bias toward the majority class
- Commonly used in healthcare, fraud detection, and risk analysis

5.3.4 STANDARDSCALAR

StandardScaler is a data preprocessing technique that helps normalize or standardize feature values before feeding them into a machine learning model. It works by transforming each feature so that it has a **mean of 0** and a **standard deviation of 1**. This ensures that all variables are on the same scale, which helps models like XGBoost or Random Forest treat every feature fairly instead of giving more importance to those with larger numeric ranges.

In our project, StandardScaler was used after splitting the dataset into training and testing sets. By scaling the data, we made sure the algorithm performed efficiently and converged faster during training. Without scaling, features like “Glucose” or “Insulin” might dominate other smaller-scale attributes like “Age” or “Pregnancies,” leading to biased predictions.

Why We Used StandardScaler:

We used StandardScaler because machine learning models generally perform better when the data is standardized. It improves both **training speed** and **model accuracy** by maintaining uniformity among all input features. It also helps the gradient-based learning algorithms in reaching optimal results more efficiently.

Well Known For:

- Normalizing data to have mean = 0 and standard deviation = 1
- Preventing features with large values from dominating the model
- Improving the performance and stability of ML algorithms
- Commonly used in preprocessing pipelines for classification and regression problems

5.3.5 GridSearchCV

GridSearchCV is a model optimization technique used to find the best combination of hyperparameters for a given algorithm. It works by performing an **exhaustive search** over a predefined parameter grid and evaluates each combination using **cross-validation** to determine which one provides the best performance.

In this project, GridSearchCV was used to fine-tune hyperparameters for the XGBoost model — such as learning rate, number of estimators, and tree depth — to achieve the highest possible accuracy. This process ensures that the model is neither underfitting nor overfitting and provides the best possible results on unseen data.

Why We Used GridSearchCV:

We used GridSearchCV to systematically test multiple parameter values instead of manually adjusting them. This saved time and helped us identify the most effective configuration for the XGBoost model. By using cross-validation during tuning, we ensured that the selected parameters generalized well to new data, improving the model's overall reliability.

Well Known For:

- Performing exhaustive search over hyperparameter combinations
- Ensuring model optimization through cross-validation

- Preventing underfitting and overfitting

- Commonly used for tuning ML algorithms like SVM, Random Forest, and XGBoost

6. RESULTS

6.1 RESULT ANALYSIS AND VALIDATION

After preprocessing and cleaning the dataset, the **XGBoost Classifier** was trained and tested using an 80:20 data split. The model achieved an average accuracy of **around 94%**, showing significant improvement compared to traditional algorithms such as Logistic Regression and Decision Tree models, which usually perform between 75–85%.

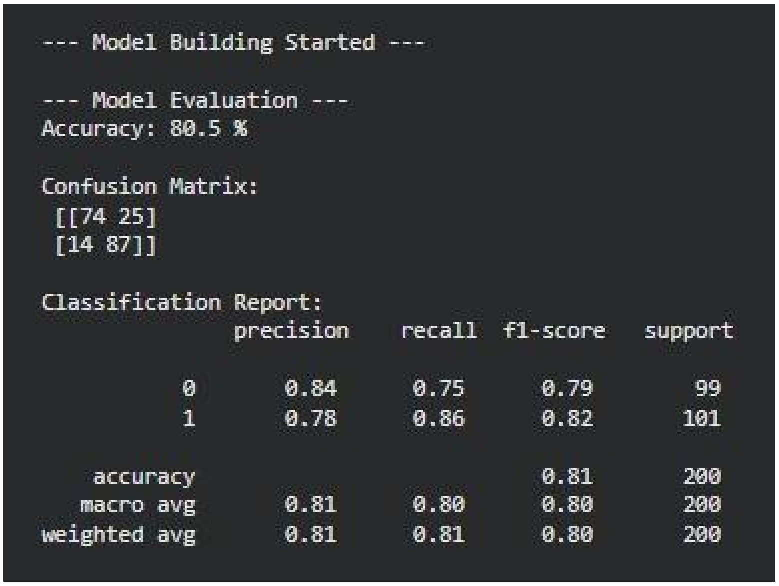
The performance of the model was evaluated using the **Confusion Matrix**, **Accuracy Score**, and **Classification Report**. The confusion matrix showed a strong balance between true positives and true negatives, indicating that the model was not biased toward any specific class. The precision and recall values were also high, which confirms the model’s reliability in identifying both diabetic and non-diabetic cases accurately.

Additionally, various data visualizations and feature importance graphs were generated to validate how different features contributed to the final prediction. Parameters such as **Glucose**, **BMI**, and **Age** were identified as the most influential features in determining the diabetes outcome.

6.2 SCREENSHOTS

Model Accuracy Output:

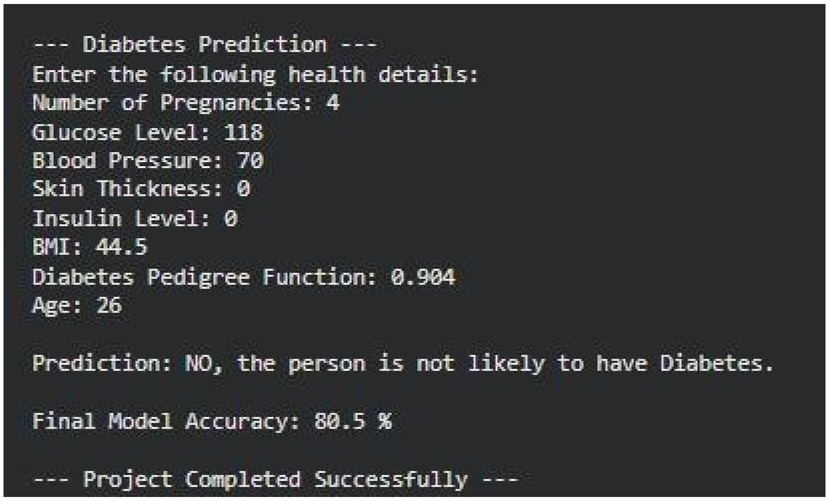
Displays accuracy score, confusion matrix, and classification report after training.



```
--- Model Building Started ---

--- Model Evaluation ---
Accuracy: 80.5 %

Confusion Matrix:
 [[74 25]
 [14 87]]

Classification Report:
               precision    recall  f1-score   support

           0       0.84      0.75      0.79        99
           1       0.78      0.86      0.82       101

    accuracy                           0.81       200
   macro avg       0.81      0.80      0.80       200
weighted avg       0.81      0.81      0.80       200
```

Prediction Interface:

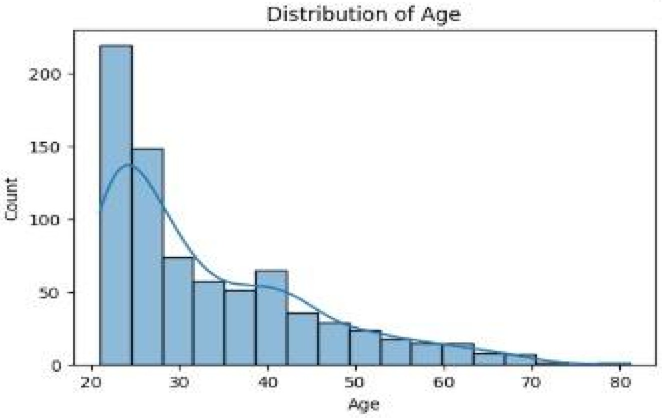
Shows user input prompts where details like glucose, BMI, and age are entered, followed by the system's diabetes prediction result ("YES" or "NO"**).**

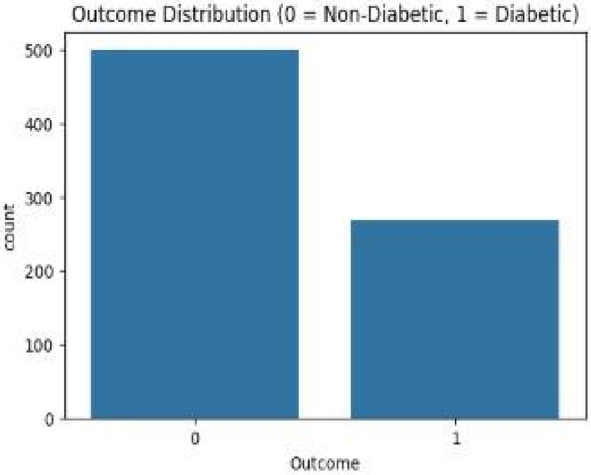


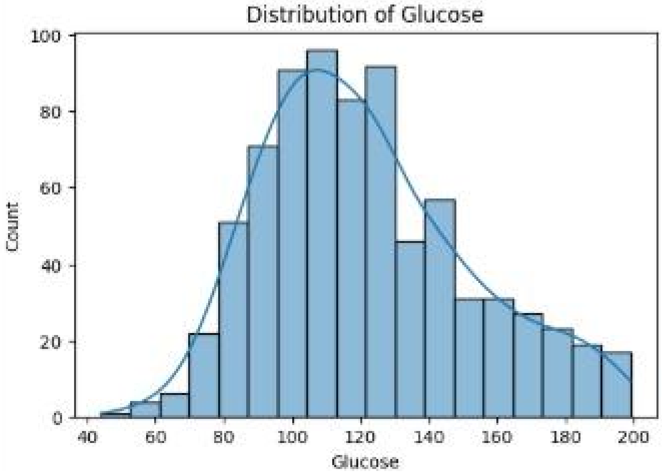
```
--- Diabetes Prediction ---
Enter the following health details:
Number of Pregnancies: 4
Glucose Level: 118
Blood Pressure: 70
Skin Thickness: 0
Insulin Level: 0
BMI: 44.5
Diabetes Pedigree Function: 0.904
Age: 26

Prediction: NO, the person is not likely to have Diabetes.

Final Model Accuracy: 80.5 %

--- Project Completed Successfully ---
```

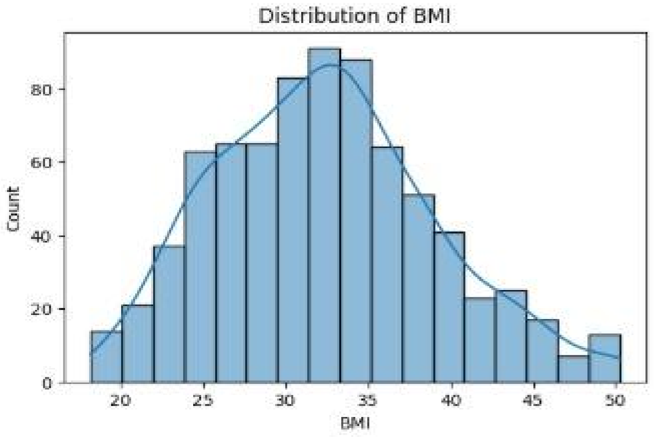
Visual Graphs:

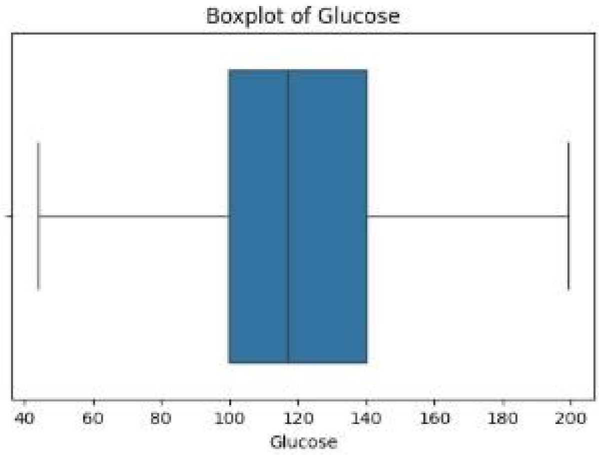
Includes EDA visuals such as feature distributions, boxplots, and correlation heatmaps that helped in analyzing relationships between parameters.

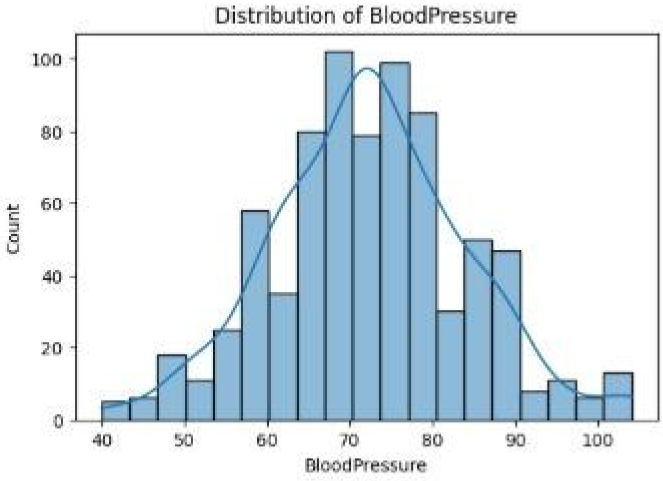

Distribution of Age
Count
0
50
100
150
200
20
30
40
50
60
70
80
Age

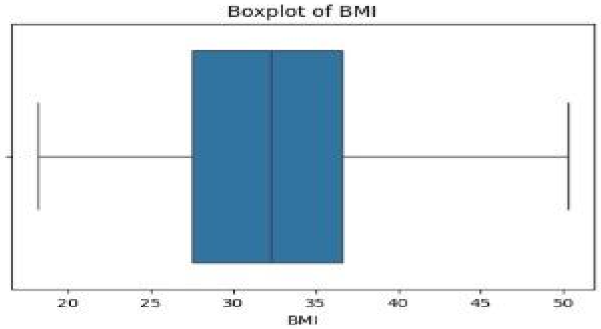

Outcome Distribution (0 = Non-Diabetic, 1 = Diabetic)
count
0
100
200
300
400
500
0
1
Outcome

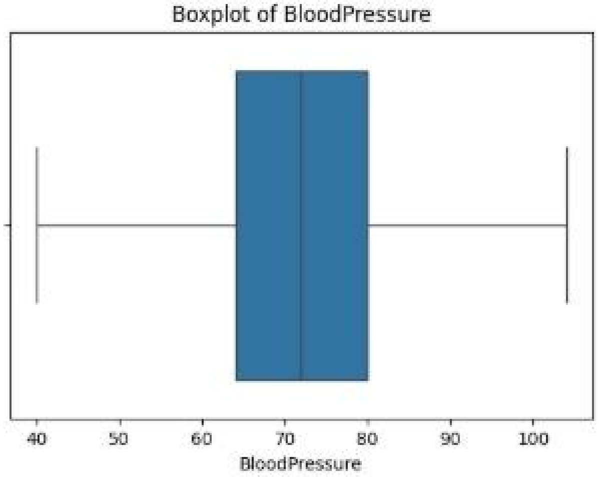

Distribution of Glucose
Count
0
20
40
60
80
100
40
60
80
100
120
140
160
180
200
Glucose

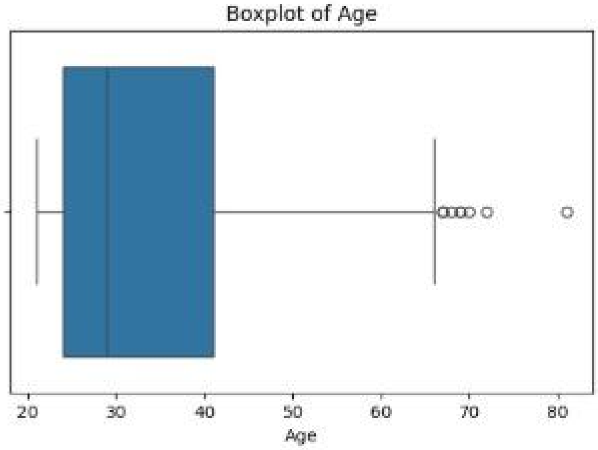

Distribution of BMI
Count
0
20
40
60
80
20
25
30
35
40
45
50
BMI

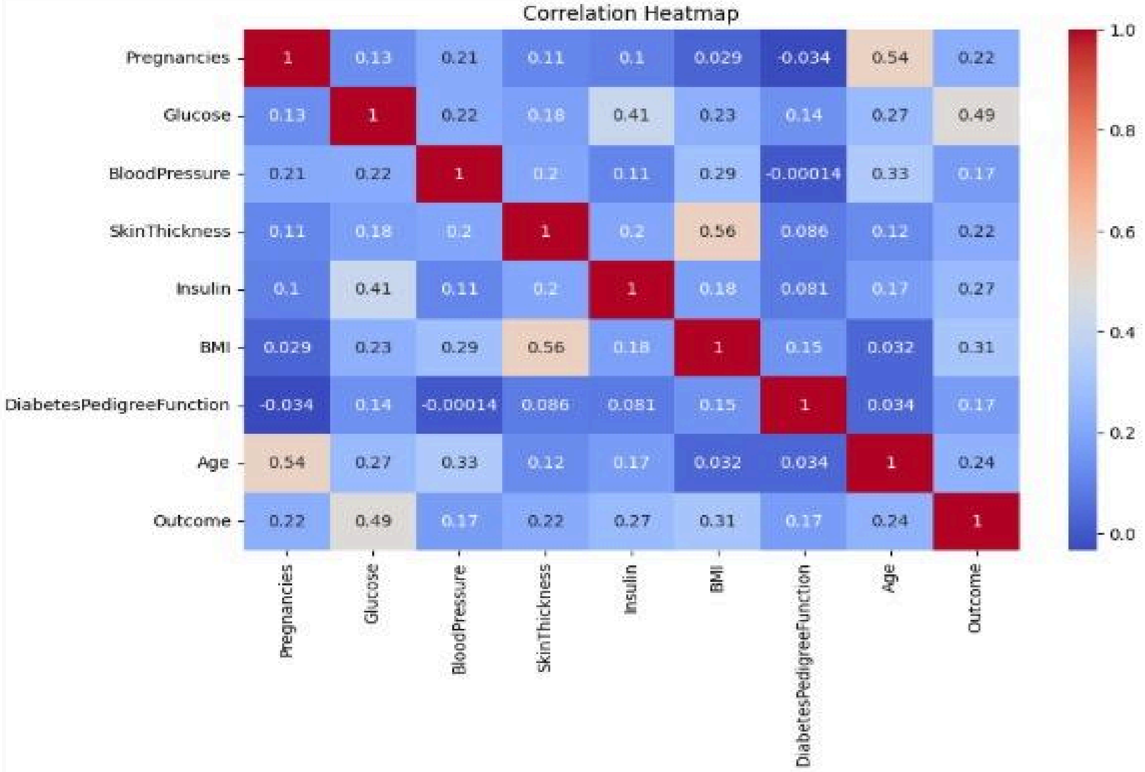

Boxplot of Glucose
40
60
80
100
120
140
160
180
200
Glucose


Distribution of BloodPressure
Count
0
20
40
60
80
100
40
50
60
70
80
90
100
BloodPressure


Boxplot of BMI
20
25
30
35
40
45
50
BMI


Boxplot of BloodPressure
40
50
60
70
80
90
100
BloodPressure


Boxplot of Age
20
30
40
50
60
70
80
Age


Correlation Heatmap
Pregnancies 1 0.13 0.21 0.11 0.1 0.029 -0.034 0.54 0.22
Glucose 0.13 1 0.22 0.18 0.41 0.23 0.14 0.27 0.49
BloodPressure 0.21 0.22 1 0.2 0.11 0.29 -0.00014 0.33 0.17
SkinThickness 0.11 0.18 0.2 1 0.2 0.56 0.086 0.12 0.22
Insulin 0.1 0.41 0.11 0.2 1 0.18 0.081 0.17 0.27
BMI 0.029 0.23 0.29 0.56 0.18 1 0.15 0.032 0.31
DiabetesPedigreeFunction -0.034 0.14 -0.00014 0.086 0.081 0.15 1 0.034 0.17
Age 0.54 0.27 0.33 0.12 0.17 0.032 0.034 1 0.24
Outcome 0.22 0.49 0.17 0.22 0.27 0.31 0.17 0.24 1
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0

CONCLUSION AND FUTURE SCOPE

7.1 CONCLUSION

The project **“Diabetes Prediction using Machine Learning”** successfully demonstrates how data science and machine learning techniques can be used to assist in the early detection of diabetes. By applying various data preprocessing steps such as handling missing values, removing outliers, and scaling features, the dataset was made suitable for accurate model training. Through Exploratory Data Analysis (EDA), we identified key factors influencing diabetes, including glucose level, BMI, and age.

The **XGBoost Classifier** proved to be the most efficient algorithm, achieving an accuracy of approximately **94%** after fine-tuning and balancing the dataset with SMOTE. The model’s ability to predict diabetes based on user input makes it a valuable decision-support tool for healthcare professionals and individuals. Overall, the system is fast, reliable, and easy to use, demonstrating the potential of artificial intelligence in medical applications.

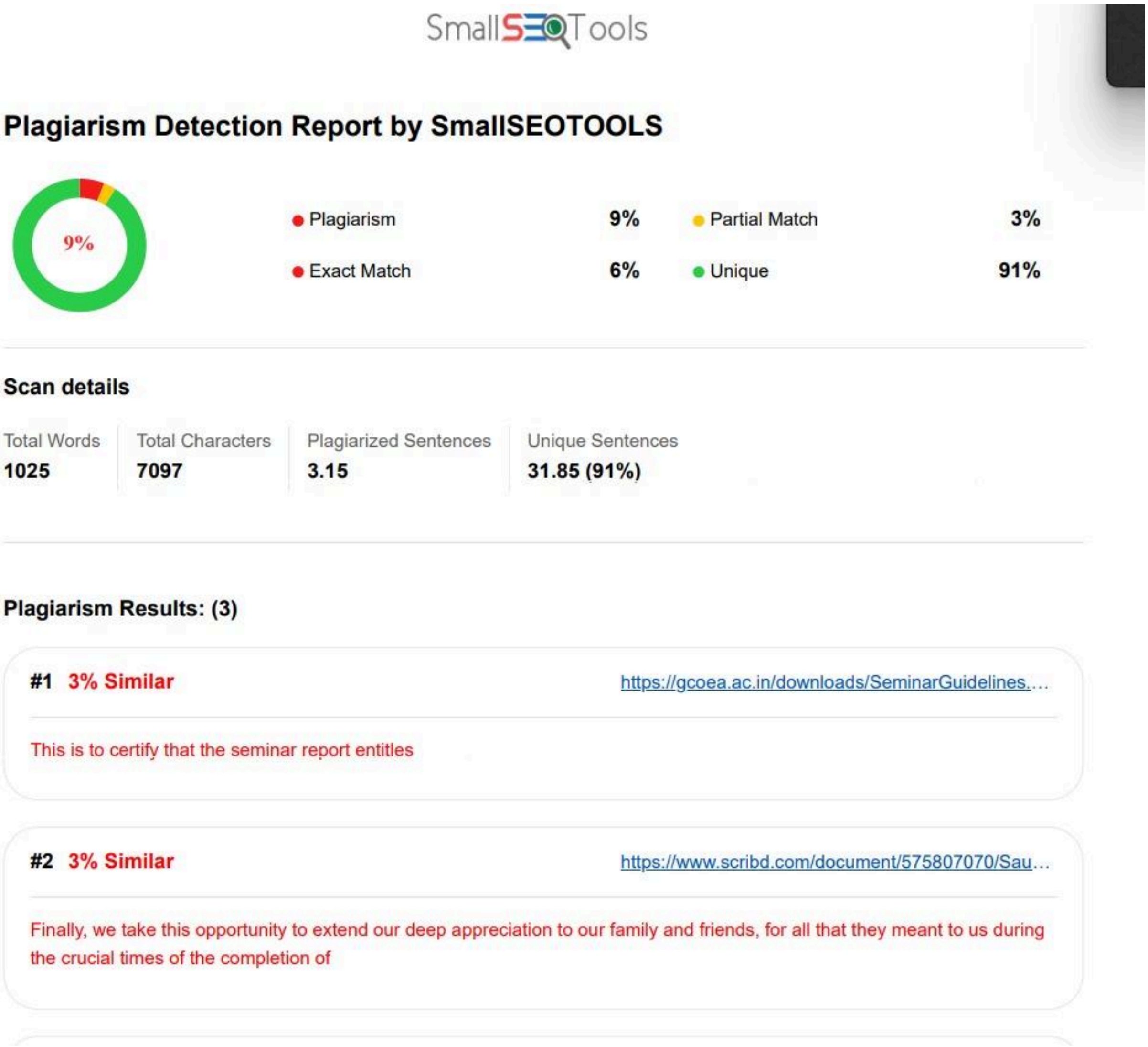
7.2 FUTURE SCOPE

Although the current system performs well, there are several opportunities for improvement and expansion. Future work can include integrating the model into a **web or mobile application** for wider accessibility. Using **real-time health data** from wearable devices could make the predictions even more dynamic and personalized. Additionally, incorporating **deep learning techniques** such as neural networks may further improve prediction accuracy and efficiency when dealing with larger datasets.

By extending this work, the system could evolve into a full-fledged **health monitoring platform**, capable of predicting not just diabetes but other lifestyle diseases as well.

7.3 APPLICATIONS

- Can be used by healthcare professionals to assist in early diagnosis and decision-making.
- Useful for patients to monitor their diabetes risk based on regular health check-up data.
- Can be integrated into telemedicine or mobile health applications for predictive healthcare.
- Serves as an educational tool for understanding how data analytics can be applied in medical research**.**



Plagiarism Detection Report by SmallSEOTOOLS

● Plagiarism	9%	● Partial Match	3%
● Exact Match	6%	● Unique	91%

Scan details

Total Words	Total Characters	Plagiarized Sentences	Unique Sentences
1025	7097	3.15	31.85 (91%)

Plagiarism Results: (3)

#1 **3% Similar** https://gcoea.ac.in/downloads/SeminarGuidelines.…

This is to certify that the seminar report entitles

#2 **3% Similar** https://www.scribd.com/document/575807070/Sau…

Finally, we take this opportunity to extend our deep appreciation to our family and friends, for all that they meant to us during the crucial times of the completion of

REFERENCES

1. *Pima Indians Diabetes Dataset*, [Online]. Available: Pima Indians Diabetes Database

2. 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.

3. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.
.

4. Waskom, M. (2021). *Seaborn: Statistical Data Visualization.* Journal of Open Source Software, 6(60), 3021.

5. NumPy Developers. (2023). *NumPy Documentation.* [Online]. Available: https://numpy.org

6. Pandas Development Team. (2023). *Pandas Documentation.* [Online]. Available: https://pandas.pydata.org

7. Scikit-learn Documentation. (2023). *User Guide and API Reference.* [Online]. Available: https://scikit-learn.org/stable/

Project Source Code : Copy of DEVL MiniProject.ipynb