**Diabetics Prediction**

**Submitted for**

**Statistical Machine Learning CSET211**

Submitted by:

**E23CSEU1894 RAMANSHU SAHANI**

**E23CSEU1869 SHREYANSH**

Submitted to

**DR. SHAKSHI SHARMA**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

Diabetes Prediction Using Machine Learning

# 1. Abstract

Diabetes, a prevalent chronic illness, affects millions of people worldwide and demands early detection to minimize complications. This project applies machine learning techniques to predict the likelihood of diabetes in individuals based on clinical health data. Utilizing the PIMA Indian Diabetes Dataset, we employed four machine learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The models were trained to analyze features such as glucose levels, BMI, and insulin, which are crucial indicators of diabetes. After preprocessing the data to handle missing values and scaling features, we evaluated the models using metrics such as accuracy, precision, and recall. Logistic Regression proved to be the most interpretable and practical model, achieving reliable performance while maintaining simplicity. The project highlights the potential of machine learning in healthcare applications and establishes a foundation for future enhancements in diabetes prediction systems.

---

# 2. Introduction

Diabetes is a metabolic disorder that significantly impacts global health, often leading to severe complications if undetected. Early prediction and timely intervention are crucial to managing this condition effectively. In recent years, machine learning has emerged as a powerful tool to analyze medical data and predict outcomes with high accuracy. This project aims to develop a machine learning-based system that predicts diabetes using health-related attributes from the PIMA Indian Diabetes Dataset. The dataset includes essential features like glucose levels, insulin levels, BMI, and age, which are strong predictors of diabetes risk. By training various machine learning models and comparing their performance, this project seeks to identify the best-suited approach for predicting diabetes. The broader goal is to demonstrate how technology can support healthcare systems in making data-driven decisions and improving patient outcomes.

---

# 3. Related Work

Numerous studies have explored the use of machine learning for diabetes prediction. Researchers have utilized algorithms like Decision Trees, Logistic Regression, and Neural Networks to classify patients based on their likelihood of having diabetes. For instance, Logistic Regression is widely recognized for its simplicity and effectiveness in binary classification tasks. Similarly, Random Forest is known for its ability to handle complex datasets and improve accuracy through ensemble learning. Previous work also highlights the importance of data preprocessing, including handling missing values and feature scaling, to improve model performance. While deep learning models have shown promise in similar tasks, they often require larger datasets and higher computational resources. This project builds on these concepts, focusing on interpretability and practical application, which are essential in healthcare. By using a combination of classical machine learning algorithms, we aim to create an efficient and interpretable diabetes prediction model.

---

# 4. Methodology

The methodology for this project involves several systematic steps to preprocess, train, and evaluate machine learning models for diabetes prediction:

## 1. \*\*Dataset Preparation:\*\*

- The PIMA Indian Diabetes Dataset was utilized, containing health indicators like glucose, BMI, and insulin.

- Missing values were identified and imputed using the mean for continuous features.

## 2. \*\*Data Preprocessing:\*\*

- Features were standardized using `StandardScaler` to ensure uniformity and improve model performance.

## 3. \*\*Model Training:\*\*

- Four machine learning models were trained: Logistic Regression, Random Forest, SVM, and KNN.

- Grid Search CV was applied to tune hyperparameters for Logistic Regression, optimizing performance.

## 4. \*\*Model Evaluation:\*\*

- Models were evaluated using metrics like accuracy, precision, recall, and F1-score.

- Confusion matrices and classification reports provided insights into performance.

## 5. \*\*Visualization:\*\*

- Scatterplots and heatmaps were used to understand feature distributions and correlations.

---

# 5. Hardware/Software Required

## \*\*Hardware:\*\*

- A laptop or desktop with the following specifications:

- Processor: Intel i5 or equivalent

- RAM: 8GB or higher

- Storage: 20GB of free disk space for datasets and dependencies

## \*\*Software:\*\*

- \*\*Programming Language:\*\* Python 3.9 or later

- \*\*Libraries and Frameworks:\*\*

- `pandas` for data manipulation

- `numpy` for numerical operations

- `matplotlib` and `seaborn` for data visualization

- `scikit-learn` for machine learning algorithms

- \*\*IDE/Environment:\*\* Google Colab or Jupyter Notebook for coding and execution

- \*\*Additional Tools:\*\*

- GitHub for version control and project sharing

---

# 6. Experimental Results

The experimental results demonstrate the effectiveness of machine learning models in predicting diabetes.

## 1. \*\*Accuracy Comparison:\*\*

- Logistic Regression: 73%

- Random Forest: 74%

- K-Nearest Neighbors (KNN): 73%

## 2. \*\*Confusion Matrix Analysis:\*\*

- Logistic Regression showed a balance between true positives and true negatives, achieving reliable predictions.

## 3. \*\*Visualization Insights:\*\*

- Scatterplots revealed strong correlations between glucose levels and diabetes prevalence.

- Heatmaps illustrated the interdependence of features, providing insights for feature selection.

The results validate the practicality of using Logistic Regression as a predictive tool due to its simplicity and interpretability.

---

DATA VIZUALIZATION

A screenshot of a graph

Description automatically generated

A diagram of a diagram

Description automatically generated with medium confidenceA chart of different colored dots

Description automatically generated with medium confidence

A group of graphs showing different types of data

Description automatically generated with medium confidence

# A chart of different colored dots Description automatically generated with medium confidence

# 7. Conclusion

This project emphasizes the potential of machine learning in predicting diabetes, a critical health condition. By analyzing patient data from the PIMA Indian Diabetes Dataset, we trained multiple models to classify individuals as diabetic or non-diabetic. Logistic Regression stood out as the most effective and interpretable model, balancing simplicity and performance. The project demonstrates the value of using data-driven approaches in healthcare, enabling early detection and improved patient outcomes. By offering a foundation for future developments, this work underscores the role of technology in advancing healthcare solutions and tackling complex medical challenges.

---

# 8. Future Scope

The scope of this project can be expanded in several ways:

## 1. \*\*Feature Enhancement:\*\*

- Include additional health indicators like cholesterol levels and blood pressure for improved prediction accuracy.

## 2. \*\*Deep Learning Integration:\*\*

- Explore neural networks for uncovering complex patterns in larger datasets.

## 3. \*\*Real-World Deployment:\*\*

- Develop a web-based or mobile application to make the prediction model accessible to healthcare providers.

## 4. \*\*Dataset Diversification:\*\*

- Validate the model on datasets from different demographics to improve generalization.

## 5. \*\*Explainability:\*\*

- Integrate explainable AI techniques to provide insights into model predictions, aiding medical professionals in decision-making.

---

# 9. GitHub Link of Your Complete Project

https://github.com/Shreyansh97531/DIABATESE\_PREDICTION

The repository includes:

- Python scripts for data preprocessing, model training, and evaluation

- The dataset (if permissible to share)

- A README file detailing the project structure and execution steps

- Presentation slides summarizing the project

---