COGNITIVE COMPUTING

Project Report

Diabetes Prediction Using Machine Learning

Submitted by:

Chirag Sood(102317142)
Karan Nigam(102317145)

Submitted to:Dr.Sukhpal Singh

Department of Computer Science and Engineering
Thapar Institute of Engineering and Technology,Patiala

Submitted on: 26th March, 2025.

JAN-MAY 2025

Diabetes Prediction

1. Introduction & Problem Statement

Diabetes is a chronic disease that affects millions worldwide, requiring early detection for effective management and treatment. Machine learning models can help predict diabetes based on various health indicators, improving early diagnosis and preventive care.

Why is it Important?

Accurate diabetes prediction enables timely medical intervention, reducing complications and improving patient outcomes. By analyzing health metrics, predictive models can classify individuals into diabetic or non-diabetic groups, aiding healthcare professionals in decision-making.

Key Questions:

- How accurately can diabetes be predicted using health metrics?
- What are the most significant predictors of diabetes?
- How can machine learning models improve diabetes diagnosis?

Dataset Used: The Pima Indians Diabetes Dataset , which contains various health metrics like glucose level, BMI, insulin levels, age, and more, is used for this analysis.

Here's the link to the dataset that we have used:

https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906/input

2. Data Exploration & Preprocessing

Data Structure: The dataset consists of 768 samples with multiple features, including:

- Glucose (blood sugar level)
- BMI (Body Mass Index)
- Blood Pressure
- Insulin Levels
- Age
- Pregnancies (Number of times pregnant)
- Diabetes Pedigree Function (Genetic risk factor)

-> Outcome (Diabetes diagnosis: 0 = No, 1 = Yes)(Dependent Value)

Data Preprocessing:

- Handling Missing Values: Missing values in insulin and BMI are imputed using median values.
- Feature Scaling: StandardScaler is applied to normalize the features.
- Outlier Detection: Box plots are used to identify extreme values.
- Train-Test Split: The dataset is divided into training and testing sets for model evaluation.

3. Model Implementation & Evaluation

Clustering Models Used:

1. Logistic Regression
○ A simple yet effective baseline model for classification.

2.Decision Tree Classifier
○A tree-based model that splits data based on feature importance to make predictions.

3. Random Forest Classifier
○ Captures complex patterns in the data for improved accuracy.

Training Process:
● Data Preparation: The dataset is split into training and validation sets.
● Model Training:
○ For K-Means, multiple runs with different K values are performed to find the optimal number of clusters.
○ For DBSCAN, hyperparameters like epsilon (ε) and minimum samples are fine-tuned.
● Model Validation:
○ Evaluated using silhouette score and cluster visualization.
○ Ensured meaningful cluster separation and real-world interpretability.

Evaluation Metrics:
● **Data Preparation:** The dataset is split into **training (80%)** and **testing (20%)** sets.
● **Model Training:**Each model is trained on the training set using appropriate hyperparameters.
●**Model Validation:**Evaluated using accuracy, precision, recall, and F1-score.

4. Results & Insights:

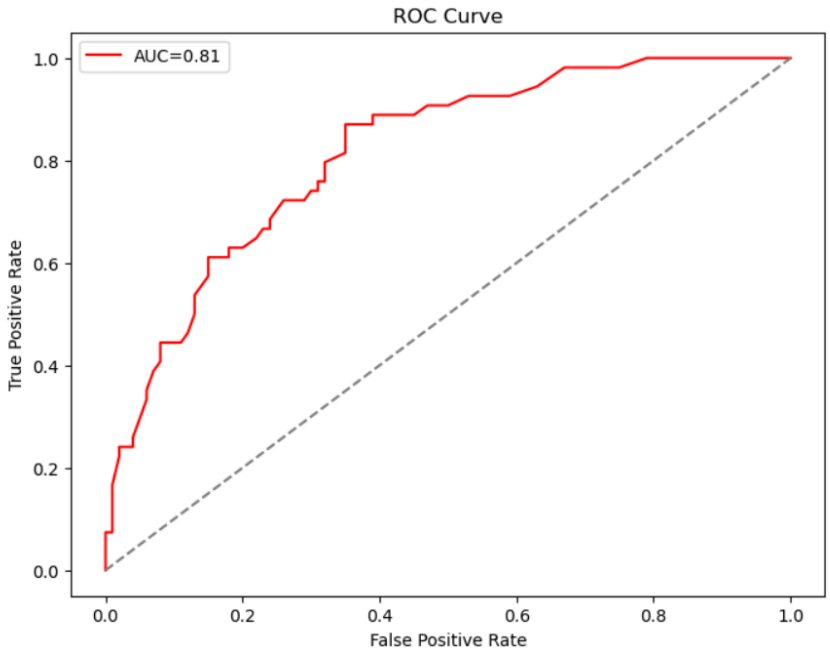
Key Findings:

- Glucose and BMI are strong predictors of diabetes.
- The Random Forest model outperforms other models in predictive accuracy.
- Feature Importance Analysis highlights the most significant health indicators for diabetes prediction.

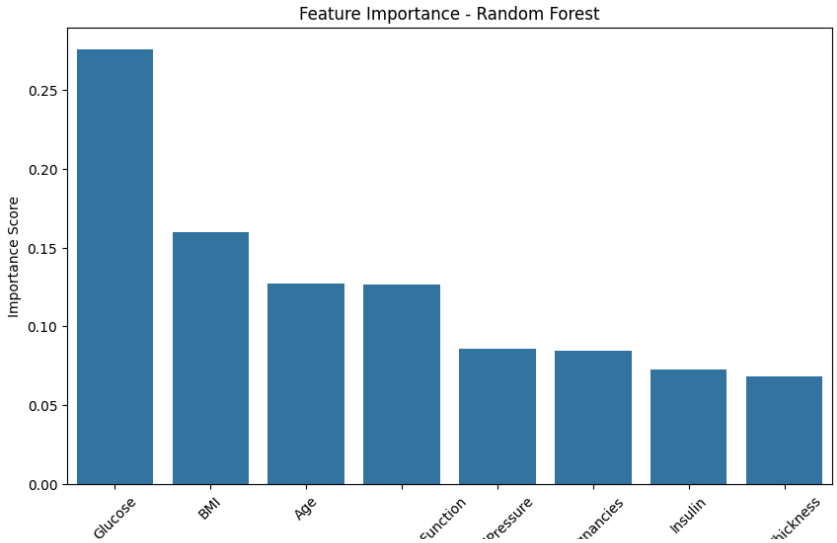
Visualizations:

1. **Confusion Matrix:** Displays correct and incorrect classifications.

2. **ROC Curve:** Shows the trade-off between sensitivity and specificity.



3.**Feature Importance Plot:** Highlights key health factors influencing diabetes prediction.



5. Challenges & Future Improvements

Challenges Faced:

- Handling class imbalance in diabetic and non-diabetic cases.
- Ensuring model generalization across diverse populations.
- Addressing data biases in the dataset.

Future Improvements:

- Incorporating More Features: Lifestyle factors like diet and exercise can enhance predictions.
- Hybrid Models: Combining multiple machine learning techniques for better accuracy.
- More Data Sources: Using additional datasets for broader generalizability.

Conclusion:

This project demonstrates how machine learning models can effectively predict diabetes. By leveraging predictive analytics, healthcare providers can identify high-risk individuals early, enabling better disease prevention and management strategies.

References:

- https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906/notebook
- https://www.youtube.com/watch?feature=shared&v=-eZ7V1vZp4k&themeRefresh=1