Predicting Diabetes Using Machine Learning Techniques

This project aims to leverage **machine learning (ML) techniques** to predict **diabetes onset** based on diagnostic health data. Diabetes is a **chronic disease** that affects millions worldwide, leading to severe complications like cardiovascular diseases, kidney failure, and blindness. The goal is to develop a **predictive model** that can help in early detection, ultimately improving patient outcomes and reducing healthcare costs.

Why This Project?

1. **Healthcare Impact** – Early detection of diabetes can significantly improve patient care and reduce medical expenses.
2. **Data-Driven Insights** – Traditional methods rely on static thresholds (e.g., glucose levels), but ML can capture complex relationships between various risk factors.
3. **Public Health Importance** – Identifying high-risk individuals allows for **targeted interventions** such as diet modifications, exercise plans, and lifestyle changes.

What Was Used?

- **Dataset:** The **Pima Indians Diabetes Dataset** (from the UCI ML Repository) containing **768 records** of diagnostic health data.
- **Data Preprocessing:**

  ✓ **Handling Missing Values:** Imputation techniques (mean/median) were used for variables like **Glucose, Blood Pressure, Skin Thickness, and Insulin**.

  ✓ **Class Imbalance Handling: SMOTE (Synthetic Minority Oversampling Technique)** was applied to balance the dataset.

- **Exploratory Data Analysis (EDA):** Boxplots, correlation heatmaps, and **Principal Component Analysis (PCA)** were used to understand the data.

What Was Done?

1. Machine Learning Models Applied:

Four models were used to predict whether a patient has diabetes (**binary classification problem**):

1. **Logistic Regression** – A linear model that achieved **76% accuracy**, identifying **glucose and BMI** as the most significant predictors.
2. **Decision Tree** – A non-linear model with **78% accuracy**, offering **intuitive decision rules** for easy interpretation.
3. **Random Forest** – An **ensemble method** that reduced overfitting and provided reliable predictions with feature importance rankings.
4. **Ensemble Voting Classifier** – A combination of the above models, achieving the **highest accuracy and robustness**.

2. Key Findings:

- **Glucose, BMI, and Insulin** are the most critical factors in diabetes prediction.
- **Patients with glucose levels above 130 mg/dL** have a significantly higher risk of developing diabetes.
- **Machine learning models outperform traditional diagnostic methods** by considering multiple risk factors simultaneously.

3. Performance Metrics Evaluated:

- **Accuracy, Precision, Recall, F1-score, and ROC-AUC** were used to measure model performance.
- **Ensemble Voting Classifier** performed the best, as it combined the strengths of multiple models.

Why These Methods?

- **Logistic Regression** is simple and interpretable, making it useful in clinical settings.
- **Decision Trees** provide clear decision rules but can overfit.
- **Random Forest** improves robustness by reducing overfitting.
- **Ensemble Voting Classifier** combines all models, offering the best performance.

Implications & Future Work

- **Integration into Healthcare Systems** – Predictive models can be used for **early screening and risk assessment**.
- **Preventive Measures** – Targeted interventions (e.g., **lifestyle changes for at-risk individuals**) can reduce diabetes cases.
- **Expanding the Dataset** – Including **more diverse populations and additional risk factors (e.g., diet, exercise, and genetics)** can improve prediction accuracy.

This project **demonstrates how machine learning can transform healthcare** by providing **early warnings and actionable insights** for diabetes prevention and management.