Disease Diagnosis Prediction Project Report

# Objective:

To build a predictive model for early diagnosis of diabetes using medical data, aiding healthcare professionals in identifying high-risk patients and providing timely interventions.

# Dataset:

PIMA Indians Diabetes Dataset  
• Source: UCI Machine Learning Repository  
• Features: 8 input medical measurements (e.g., glucose, BMI, age, etc.)  
• Target: Outcome (1 = diabetic, 0 = non-diabetic)

# Exploratory Data Analysis (EDA):

- Detected that some features like Glucose, BloodPressure, BMI had 0 values, which are medically implausible.  
- Found that Glucose, BMI, and Age were highly correlated with diabetes.  
- Outcome distribution showed imbalanced classes: more non-diabetics than diabetics.

# Problems Faced and How I Solved Them:

|  |  |  |
| --- | --- | --- |
| Problem | Description | Action Taken |
| Invalid Zero Values | Many entries had 0 for Glucose, BMI, etc. | Replaced zeros with median values of each feature. |
| Imbalanced Dataset | Class 0 outnumbered class 1. | Used F1 Score and AUC instead of accuracy to evaluate models. |
| Overfitting in Neural Network | MLPClassifier overfit small training data. | Reduced hidden layers and added early stopping criteria. |
| Interpretability of Models | Tree-based models can be black-boxes. | Used SHAP to visualize feature importance and patient-specific risk. |

# Models Used:

• Gradient Boosting – Strong performance on tabular data  
• Support Vector Machine (SVM) – Good for small, clean datasets  
• Neural Network (MLP) – Captures complex nonlinear patterns

# Evaluation Metrics:

• F1 Score to balance precision and recall.  
• AUC-ROC to measure discrimination between classes.  
• Confusion Matrix to visualize TP, FP, FN, TN.

# Results Summary:

• Gradient Boosting: F1 Score = 0.76, AUC = 0.84  
• SVM: F1 Score = 0.73, AUC = 0.82  
• Neural Network (MLP): F1 Score = 0.71, AUC = 0.80

# Insights for Healthcare Professionals:

1. Glucose, BMI, and Age are the most impactful indicators for diabetes risk.  
2. Patients with high predicted probabilities can be flagged for early intervention.  
3. The model is interpretable using SHAP values, helping doctors understand why a prediction was made.  
4. This tool can be integrated into EHR systems to assist doctors in routine screenings.

# Personal Learnings:

• Gained deep understanding of model interpretability using SHAP.  
• Understood the importance of preprocessing medical data (e.g., imputing zeros).  
• Learned to evaluate imbalanced classification models with proper metrics.  
• Became familiar with deploying models that provide clinically relevant insights.

# Conclusion:

This project demonstrates how machine learning can effectively assist in early disease detection using routine medical data. With careful preprocessing, model tuning, and interpretability tools, such models can provide actionable insights that empower healthcare professionals to take data-driven decisions.