Disease Diagnosis Prediction Report

# 1. Dataset Description and Preprocessing

Dataset: PIMA Indian Diabetes Dataset  
  
Target Variable: Outcome (1 = Diabetes, 0 = No Diabetes)  
  
Features: 8 features including Glucose, BMI, Age, BloodPressure, etc.  
  
Preprocessing:  
- Checked for nulls and data types.  
- Used SelectKBest with ANOVA F-score to select top 6 features.  
- Applied StandardScaler to normalize feature values.

# 2. Models Implemented

Below are the models and reasons for selection:

|  |  |
| --- | --- |
| Model | Reason for Selection |
| Gradient Boosting | Handles non-linearity, robust performance |
| SVM | Good for binary classification with margin optimization |
| Neural Network | Captures complex feature interactions |

# 3. Evaluation & Key Visualizations

Metrics: F1 Score, ROC-AUC Score  
  
Best-performing model: Typically Gradient Boosting or Neural Network depending on data.  
  
ROC curves plotted to visually assess model discrimination power.

# 4. Insights for Healthcare Professionals

- High glucose and BMI levels are most predictive of diabetes.  
- Models can aid early detection and targeted intervention.  
- Visual tools like ROC curves can help assess model reliability for decision support.

# 5. Challenges and Solutions

|  |  |
| --- | --- |
| Challenge | Solution |
| Feature relevance | Used statistical method (SelectKBest) |
| Class imbalance (minor issue) | Used F1 and ROC-AUC metrics |
| Model tuning complexity | Started with defaults, increased epochs for MLP |