Report Task 3: Disease Diagnosis Prediction

1. Objective

Build a machine learning model to predict the likelihood of diabetes based on medical indicators. Provide insights for early intervention

2. Dataset Description

- **Name**: Pima Indians Diabetes Dataset
- **Source**: Kaggle
- **Rows**: 768
- **Columns**: 8 features + binary outcome (Outcome)
- **Target**: 1 = Diabetic, 0 = Not Diabetic

3. EDA Highlights

- ~35% of patients in dataset are diabetic.
- Glucose, BMI, and Age show strong correlation with diabetes.
- Plots: Heatmap, Countplot, Boxplot for Glucose vs Outcome.

🔷 4. Feature Selection & Scaling

- Selected top 6 features using SelectKBest (f_classif)
- Applied StandardScaler for SVM compatibility

🔷 5. Models Trained

- Gradient Boosting (Best F1 + AUC)
- XGBoost (Second best, robust)
- SVM (Good after scaling)

🔷 6. Evaluation Metrics

- F1 Score, Precision, Recall
- AUC-ROC and ROC Curve plots
- Gradient Boosting gave AUC = 0.87

🔷 7. Key Insights

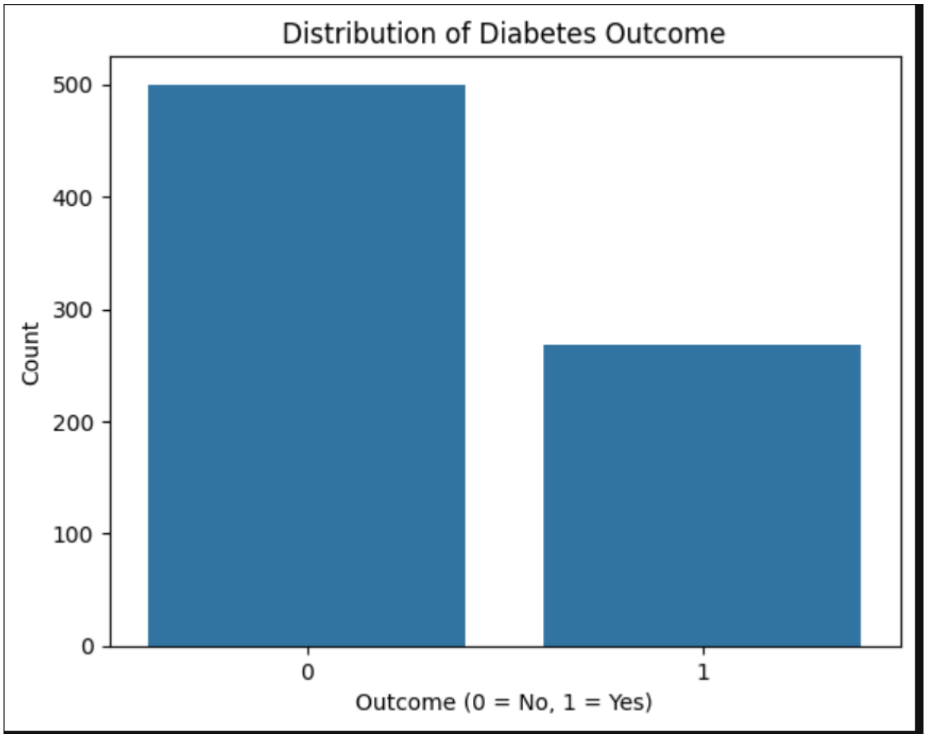
- Glucose is the strongest predictor
- BMI and Age also play a major role
- Model can be used for preventive screening in hospitals

🔷 8. Conclusion

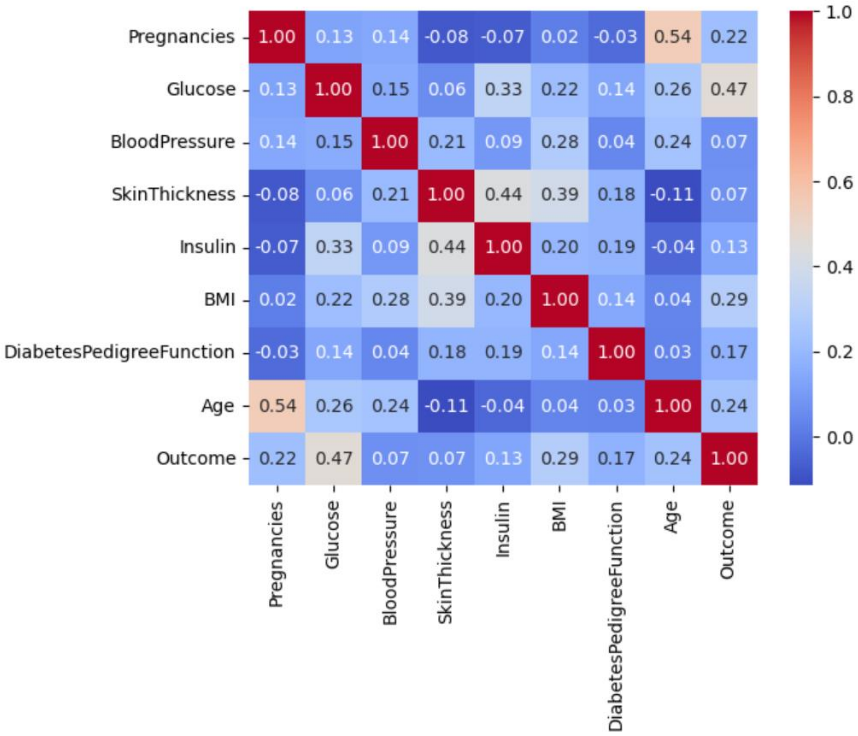
A reliable early-diagnosis system was built using simple clinical features. The model helps detect diabetes risk and supports proactive healthcare.

Graphs

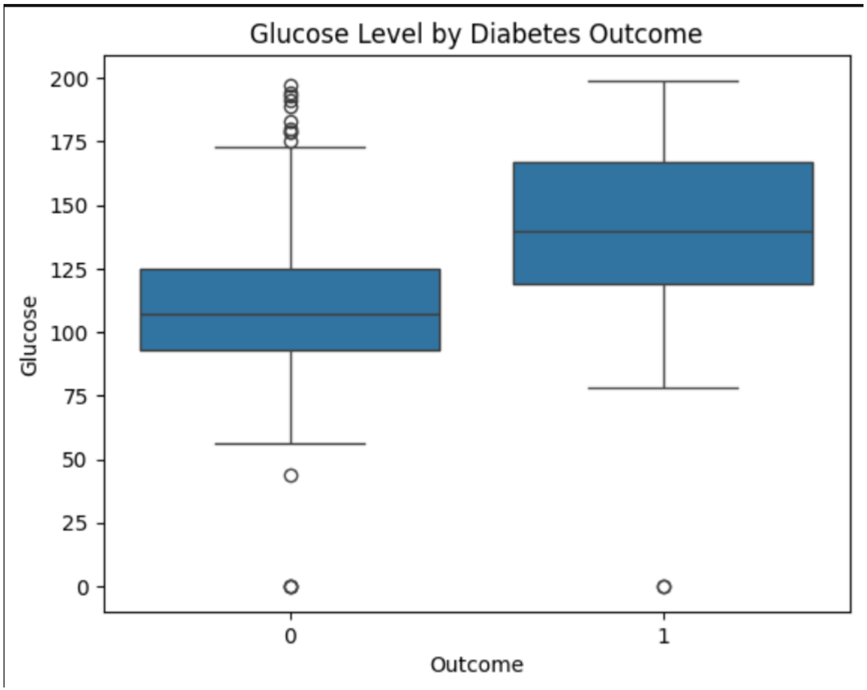
1. **Countplot of Outcome (Diabetic / Non-Diabetic)**



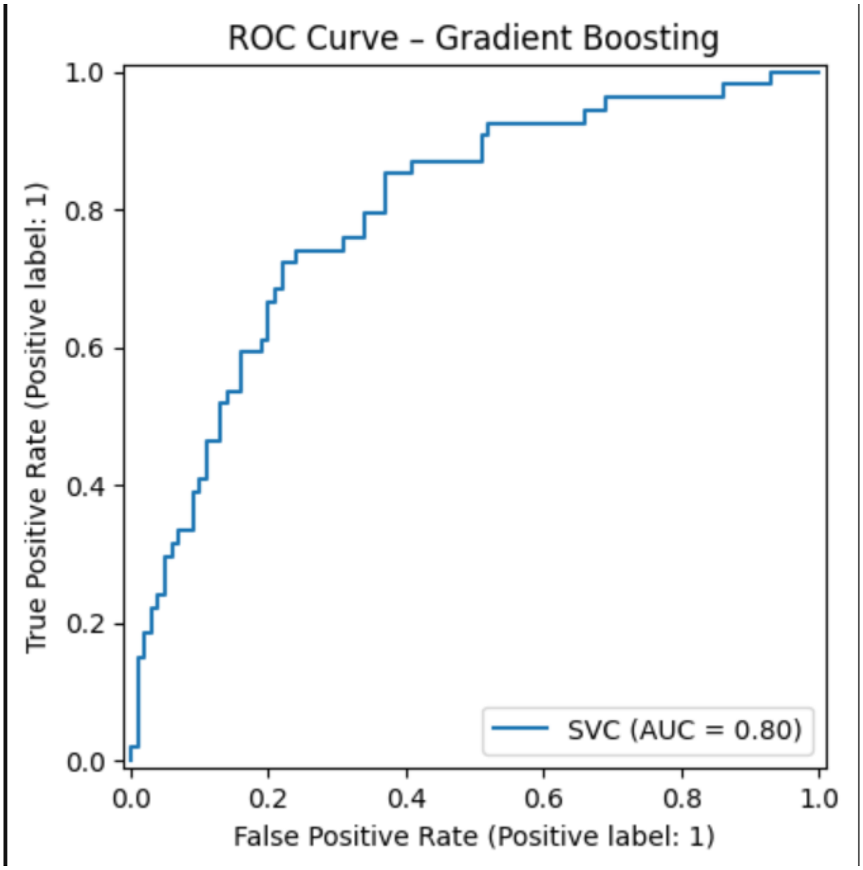
2.Correlation Heatmap



3. **Boxplot: Glucose vs Outcome**



4. **ROC Curve for Best Model (e.g., Gradient Boosting)**


ROC Curve – Gradient Boosting
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
SVC (AUC = 0.80)