PIMA Indians Diabetes Predictions using SVM - Project Documentation

# 1. Introduction

This project focuses on predicting diabetes occurrence among PIMA Indian women using Support Vector Machines (SVM). Machine learning, particularly classification techniques like SVM, plays a vital role in identifying and diagnosing diseases based on historical patient data. The dataset includes various medical attributes such as glucose levels, BMI, insulin levels, and more. The objective is to use these features to develop a predictive model that can accurately classify whether or not a patient has diabetes.

# 2. Dataset Description

The dataset used is the PIMA Indians Diabetes dataset which contains the following features:  
- Pregnancies  
- Glucose  
- BloodPressure  
- SkinThickness  
- Insulin  
- BMI  
- DiabetesPedigreeFunction  
- Age  
- Outcome (target variable: 1 for diabetic, 0 for non-diabetic)  
  
It consists of 768 records and is widely used in binary classification problems in the medical field.

# 3. Data Preprocessing

To ensure model efficiency, preprocessing steps include:  
- Handling missing or zero values by imputation.  
- Normalizing features using StandardScaler.  
- Splitting the dataset into training and testing sets.  
- Encoding target labels if required.

# 4. Model Building Using SVM

Support Vector Machine (SVM) is a supervised learning algorithm that is effective for binary classification tasks. Different kernel functions (linear, polynomial, RBF) were tested. The SVM model aims to find the optimal hyperplane that separates the classes with maximum margin.

# 5. Hyperparameter Tuning

To improve the model’s accuracy, GridSearchCV was used to tune hyperparameters such as:  
- C (Regularization parameter)  
- Kernel (linear, poly, rbf)  
- Gamma (Kernel coefficient)  
The best combination was selected based on cross-validation accuracy.

# 6. Evaluation Metrics

The model performance was evaluated using the following metrics:  
- Accuracy Score  
- Confusion Matrix  
- Precision, Recall, F1-Score  
- ROC Curve and AUC Score  
  
These metrics provide a detailed insight into how well the model performs on unseen data.

# 7. Results and Analysis

The tuned SVM model achieved a good accuracy, and the ROC-AUC curve showed high discriminative power. It was found that features such as glucose and BMI had high impact on the prediction of diabetes. Visualizations such as pair plots, heatmaps, and confusion matrices helped interpret the model’s decision boundary and performance effectively.

# 8. Conclusion

This project demonstrates how machine learning can effectively assist in medical diagnosis. SVM, when combined with proper preprocessing and hyperparameter tuning, offers reliable results for binary classification problems such as diabetes prediction. Future work may include incorporating more medical records, trying deep learning models, or using ensemble methods for improved results.