Diabetes Prediction Using Machine Learning

Author: Aditi Reddy Doma

Dataset Information

The dataset used in this project is the PIMA Indian Diabetes Dataset, publicly available on Kaggle.

Link to dataset: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

This dataset consists of 768 records with 8 input features and 1 binary output (diabetes diagnosis).

Exploratory Data Analysis (EDA)

- Checked for missing values and zero entries in key medical columns (Glucose, Insulin, BMI).
- Plotted histograms and countplots to understand distribution of outcome and features.
- Used a heatmap to visualize correlations among features.
- Observed class imbalance with more non-diabetic cases.
- Identified strong positive correlation between Glucose and Outcome.
- Boxplots revealed outliers in Insulin, Skin Thickness, and BMI.

Preprocessing Steps

- Replaced zero values in Glucose, Insulin, Skin Thickness, and BMI with median values.
- Scaled all features using StandardScaler.
- Split data into 80% training and 20% testing.

Machine Learning Models Used

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost Classifier

Diabetes Prediction Using Machine Learning

Author: Aditi Reddy Doma

Model Evaluation

- Evaluated all models using Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

- Visualized ROC curves for top models.

Model Comparison

Model Accuracy Summary:

- Gradient Boosting: 91.45%

- SVM: 90.79%

- Logistic Regression, Decision Tree, Random Forest: 89.47%

- KNN, XGBoost: 88.16%

Top Performer: Gradient Boosting Classifier

Tools & Technologies

- Python

- Pandas, NumPy

- Scikit-learn, XGBoost

- Matplotlib, Seaborn

- Jupyter Notebook

Future Scope

- Add GridSearchCV for hyperparameter tuning

- Integrate SHAP for explainability

- Deploy model with Streamlit