WEEK 3 :

DIABETES PREDICTION USING MACHINE LEARNING

ML_Internship_MIT_2025/WEEK-3/ML_Diabetes_prediction.ipynb at main · MedhaChawla5/ML_Internship_MIT_2025

Name : Medha Chawla
Reg No: 220962340
CSE(AI-ML)
medha.mitmpl2022@learner.manipal.edu

CONTENT

Dataset Overview

Problem statement

Objectives

Exploratory Data Analysis

Data Preparation

ML Model Used

Results and Conclusion

DATASET OVERVIEW (DIABETES)

The dataset used in this project consists of 2,000 records and 9 features, all related to medical parameters and patient information that could influence diabetes diagnosis. Below is a brief description of each feature:

Column Name	Description
Pregnancies	Number of times the patient has been pregnant.
Glucose	Plasma glucose concentration (mg/dL) after a 2-hour oral glucose tolerance test.
BloodPressure	Diastolic blood pressure (mm Hg).
SkinThickness	Thickness of the triceps skin fold (mm).
Insulin	2-hour serum insulin level (µU/mL).
BMI	Body Mass Index, calculated as weight in kg / (height in m)^2.
DiabetesPedigreeFunction	A function that scores likelihood of diabetes based on family history.
Age	Age of the patient in years.
Outcome	Binary classification (0 = No Diabetes, 1 = Has Diabetes).

PROBLEM STATEMENT

- **Diabetes Mellitus** is a metabolic disorder characterized by high blood sugar levels over a prolonged period. It affects millions globally and is a leading cause of heart disease, kidney failure, and other complications.

- Many individuals remain **undiagnosed** due to lack of symptoms in the early stages, which delays treatment and increases health risks.

- **Timely prediction and diagnosis** of diabetes can help initiate early interventions, enabling better management and prevention of complications.

- In this project, we aim to develop a **supervised machine learning model** that can accurately predict the presence or absence of diabetes in patients based on various **medical and physiological features** (e.g., glucose level, BMI, insulin level, etc.).

- The goal is to assist healthcare professionals by providing a **data-driven decision support system** that can complement clinical diagnoses and improve patient outcomes.

OBJECTIVES

- **Understand and explore the dataset**
  - Identify feature types and data distribution
  - Spot anomalies and inconsistencies (e.g., zero values in medical features)

- **Handle missing and invalid data**
  - Replace biologically impossible values (like 0 in Glucose , BMI etc.) with nan.
  - Apply mean imputation to fill missing values

- **Prepare features for modeling**
  - Rename and format columns for clarity
  - Analyze feature relationships and relevance

- **Build a machine learning model**
  - Use Logistic Regression as a baseline classifier
  - Use Decision Trees , SVM and Random Forest classifier .
  - Train the model to predict diabetes presence based on input features

- **Evaluate model performance**
  - Measure predictive accuracy on test data
  - Analyze strengths and limitations of the baseline model

EXPLORATORY DATA ANALYSIS

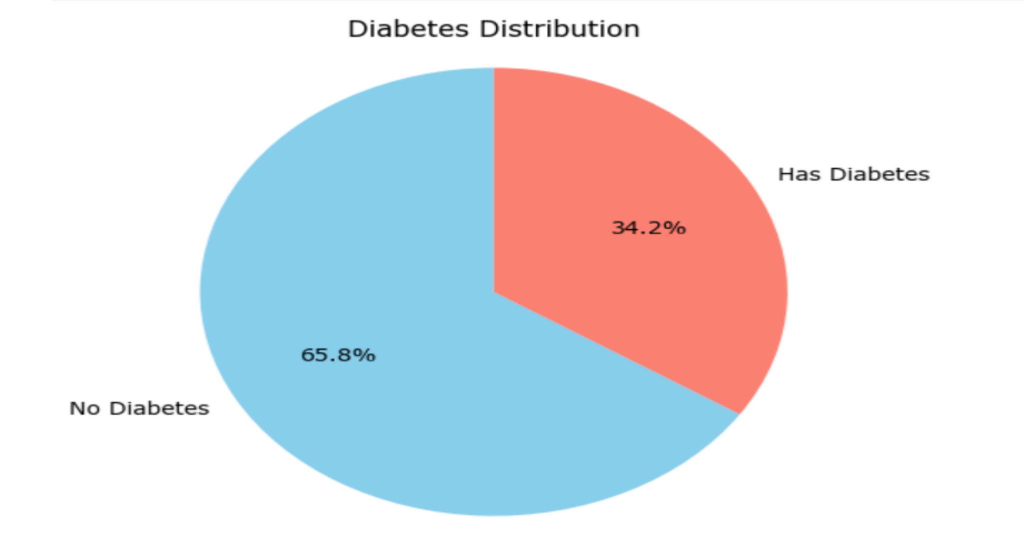
- **Dataset Shape:** 2,000 rows × 9 columns
- **Features:** 8 input variables + 1 target variable (Outcome)

- **Initial Checks**
  - Verified data types and column names
  - Checked for missing or invalid values (0)

- **Target Variable**
  - **Pie Chart**: Shows diabetes distribution (0 = No Diabetes, 1 = Has Diabetes)
  - Moderate class imbalance observed

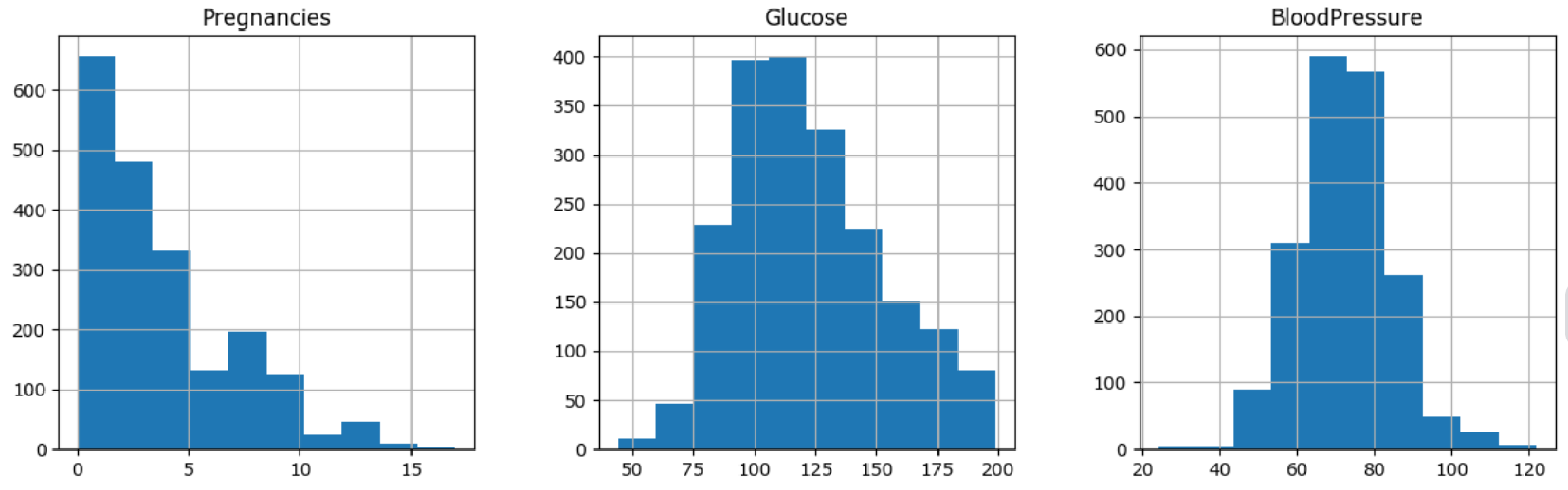
- **Data Cleaning Insight**
  - Zeros (0) In medical features like Glucose ,BMI replaced with nan
  - Imputation planned using mean values and median values

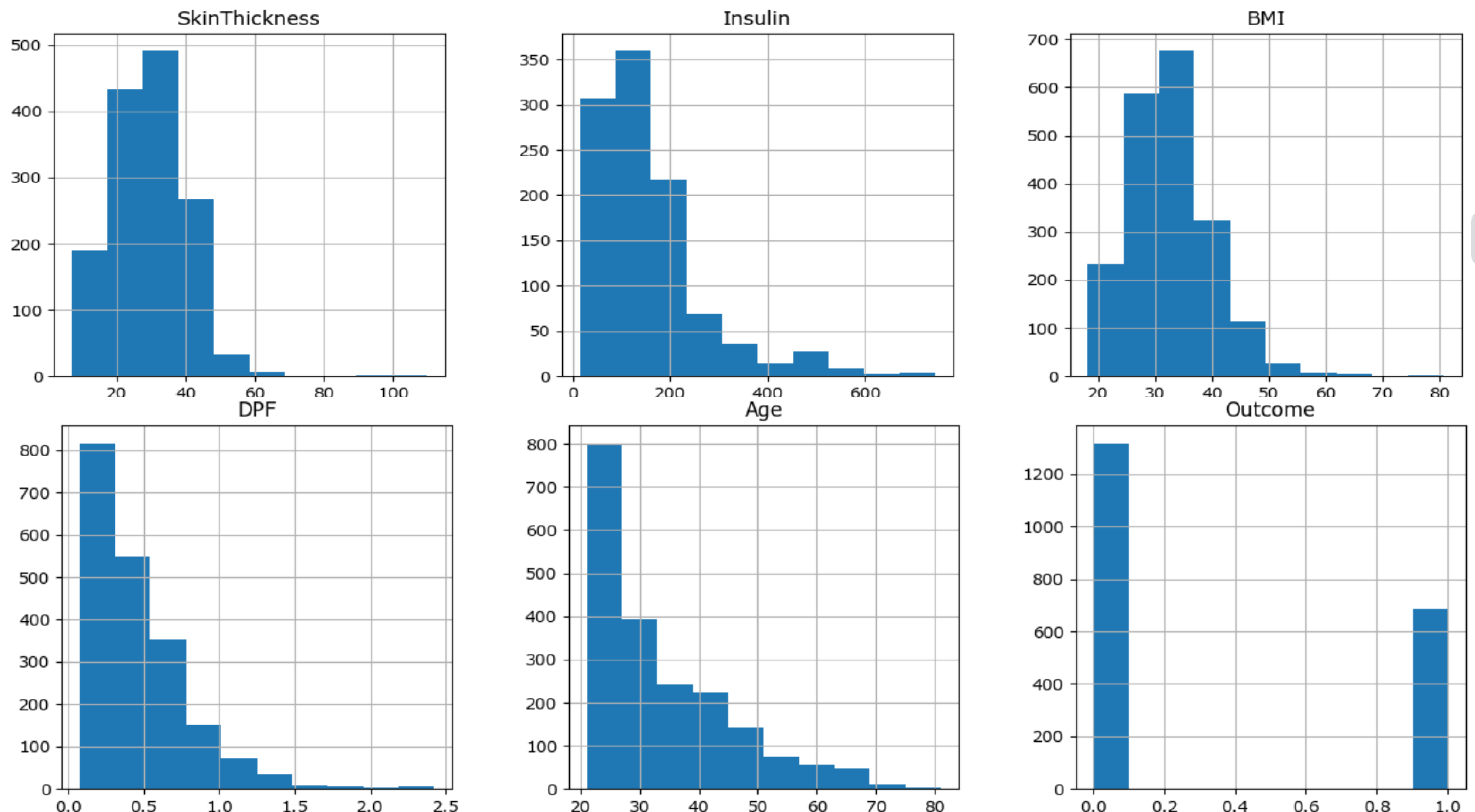
- **Feature Distributions**
  - **Histograms** plotted for all numeric features
  - Revealed skewness and possible outliers in columns like insulin and SkinThickness



Diabetes distribution :
65.8% : No Diabetes
34.2% : Has Diabetes

Histogram Distribution of Features





SkinThickness
Insulin
BMI
DPF
Age
Outcome

DATA PREPARATION

- Renamed the column **DiabetesPedigreeFunction** to **DPF** for easier reference.
- Identified zero values in the following features: **Glucose, BloodPressure, SkinThickness, Insulin, and BMI.**
- Since zero values are not medically valid for these features, they were treated as missing and replaced with **NaN**.
- Applied **mean/median imputation** to fill missing values in the affected columns.
- Mean imputation helped retain the statistical distribution of each feature while ensuring no loss of data.

```
df_copy['Glucose'].fillna(df_copy['Glucose'].mean(), inplace=True)
df_copy['BloodPressure'].fillna(df_copy['BloodPressure'].mean(), inplace=True)
df_copy['SkinThickness'].fillna(df_copy['SkinThickness'].median(), inplace=True)
df_copy['Insulin'].fillna(df_copy['Insulin'].median(), inplace=True)
df_copy['BMI'].fillna(df_copy['BMI'].median(), inplace=True)
```

ML MODELS USED

ML_Internship_MIT_2025/WEEK-3/ML_Diabetes_prediction.ipynb at main · MedhaChawla5/ML_Internship_MIT_2025

•Evaluated multiple machine learning classification models to predict diabetes based on medical features.

•Used **GridSearchCV** with cross-validation (Shuffle Split) to find best hyperparameters for each model.

To identify the most suitable machine learning model for predicting diabetes, multiple classification algorithms were implemented and evaluated. These included **Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM).**

The goal was to compare their performance and determine the best-performing model based on accuracy.

	model	best_parameters	score
0	logistic_regression	{'C': 5}	0.763125
1	decision_tree	{'criterion': 'gini', 'max_depth': 10}	0.905000
2	random_forest	{'n_estimators': 200}	0.952500
3	svm	{'C': 20, 'kernel': 'rbf'}	0.869375

RESULTS AND CONCLUSION

- After training and tuning four different machine learning models using GridSearchCV with cross-validation, the performance of each model was compared based on their accuracy scores.
- The logistic regression model, used as a baseline, achieved an accuracy of **76.31%** with C=5 as its best parameter.
- The decision tree classifier improved significantly over the baseline with an accuracy of **90.50%**, using the gini criterion and a maximum depth of 10.
- The random forest classifier outperformed all other models, achieving the highest accuracy of **95.25%** with n_estimators=200, indicating its strong capability to generalize on unseen data by leveraging ensemble learning.
- The SVM model also showed solid performance with an accuracy of **86.93%**, particularly when using the rbf kernel and C=20.
- In conclusion, the **random forest classifier emerged as the best model for this dataset**, providing the highest predictive accuracy and demonstrating robustness across varying hyperparameter settings.

THANKYOU

Name : Medha Chawla
Reg No : 220962340
CSE(AI-ML)
medha.mitmpl2022@learner.manipal.edu