Project-1

Experiment-4:

Diabetes data:

Context: The diabetes dataset is about finding the presence of diabetes in a person depending upon the attributes such as {Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedegreeFunction, Age}.

Content: The input dataset has 768 records with 9 columns where first 8 columns are the features of the dataset and the last column has the label. The output label is a binary datatype and the dataset has 500 patients with no diabetes and 268 patients with diabetes. The features {Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, Age} are discrete data type and {BMI, DiabetesPedigreeFunction} are continuous data type.

Missing values or null data points: There is no missing or null data in the given dataset.

Outliers in the dataset: Few attributes with zero value are outliers here and the following are those attributes. Blood Pressure has 35 zeros with 19 having no diabetes and 16 having diabetes; Glucose has 5 zeros with 3 having no diabetes and 2 having diabetes; SkinThickness has 227 zeros with 139 having no diabetes and 88 having diabetes; BMI has 11 zeros with 9 having no diabetes and 2 having diabetes; Insulin has 374 zeros with 236 having no diabetes and 138 having diabetes. There 724 records where BloodPressure, BMI and Glucose are not zero.

Feature: From the histogram data grouped between individual feature and outcome, features like gluose, skinthickness, age, diabetes pedigree function and insulin have variations in their plots compared to other features, hence, these features have significant impact while training the model.

Testing Strategy: Stratified K Fold with 10 splits is the testing strategy deployed.

Classification accuracy of algorithms:

Using Train/ Test split:

Model	**Accuracy**
KNN	0.729282
DB	0.745856
GNB	0.734807
BNB	0.657459

Using K-Fold cross validation:

Model	**Accuracy**
KNN	0.714136
DB	0.696328
GNB	0.754205
BNB	0.656069

Though DB has highest accuracy in train/test split it has lesser accuracy in the K-Fold cross validation because the model is overfitting due to low variance in the train/ test split compared to k-Fold cross validation.

GNB has consistent accuracy in both Train/ test split and K-means cross validation. Hence GNB would be a better model for the diabetes data.