Project-1

Experiment-4:

Thyroid data:

Context: The Thyroid dataset is about finding the presence of Thyroid in a person depending upon the attributes such as {T3_resin, Serum_thyroxin, Serum_triiodothyronine, B asal_TSH, Abs_diff_TSH}.

Content: The input dataset has 215 records with 6 columns where first 5 columns are the features of the dataset and the last column has the label. The output label is a categorical datatype with 3 classes (1, 2, 3) and the dataset has 150 patients with Thyroid type '1', 35 patients with Thyroid type '2' and 30 patients with Thyroid type '3'.

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

Feature: From the histogram data grouped between individual feature and outcome, all features have variations in their plots compared to other features, hence, all 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	0.944444
DB	0.888889
GNB	0.944444
BNB	0.796296

Using K-Fold cross validation:

Model	**Accuracy**
KNN	0.925758
DB	0.939394
GNB	0.967749
BNB	0.734848

KNN and GNB have highest accuracy in the train/ test split but GNB has more accuracy in the K-Fold validation. Hence, GNB would be a better model in the case of Thyroid data.