K-Nearest Neighbours Classifier

**Ex no: 5**

**RUBAN S**

**Aim:**

To build a K-Nearest Neighbour classifier to predict drug types using the provided dataset.

**What is K-Nearest Neighbour (KNN)?**

K-Nearest Neighbour (KNN) is a simple, non-parametric, and lazy learning algorithm used for both classification and regression tasks. It classifies a data point based on how its Neighbour are classified. The KNN algorithm stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

**Procedure:**

1. **Import necessary libraries and load the dataset:**
   * Import libraries such as pandas, numpy, scikit-learn, matplotlib, and seaborn.
   * Load the dataset into a DataFrame.
2. **Encode categorical data:**
   * Use LabelEncoder to transform categorical columns (Sex, BP, and Cholesterol) into numerical values.
3. **Prepare the data:**
   * Select features (Age, Sex\_encode, BP\_encode, Chol\_encode, Na\_to\_K) and the target variable (Drug).
   * Split the data into training and testing sets using train\_test\_split.
4. **Define models:**
   * Include KNeighbourClassifier for the KNN model.
   * Use GridSearchCV for hyperparameter tuning of KNN to find the best parameters.
5. **Train the KNN model with GridSearchCV:**
   * Define a parameter grid for KNN, including n\_Neighbour, weights, and metric.
   * Perform grid search to find the best parameters.
   * Print the best parameters found by GridSearchCV.
6. **Evaluate the model:**
   * Train the KNN model with the best parameters on the training data.
   * Make predictions on the test data.
   * Evaluate the model using accuracy score, classification report, and confusion matrix.
7. **Plot feature distributions:**
   * Use seaborn to plot the distributions of features such as Age, Sex\_encode, BP\_encode, Chol\_encode, Na\_to\_K.
8. **Plot confusion matrices:**
   * Plot the confusion matrices for the KNN model to visualize the classification performance.

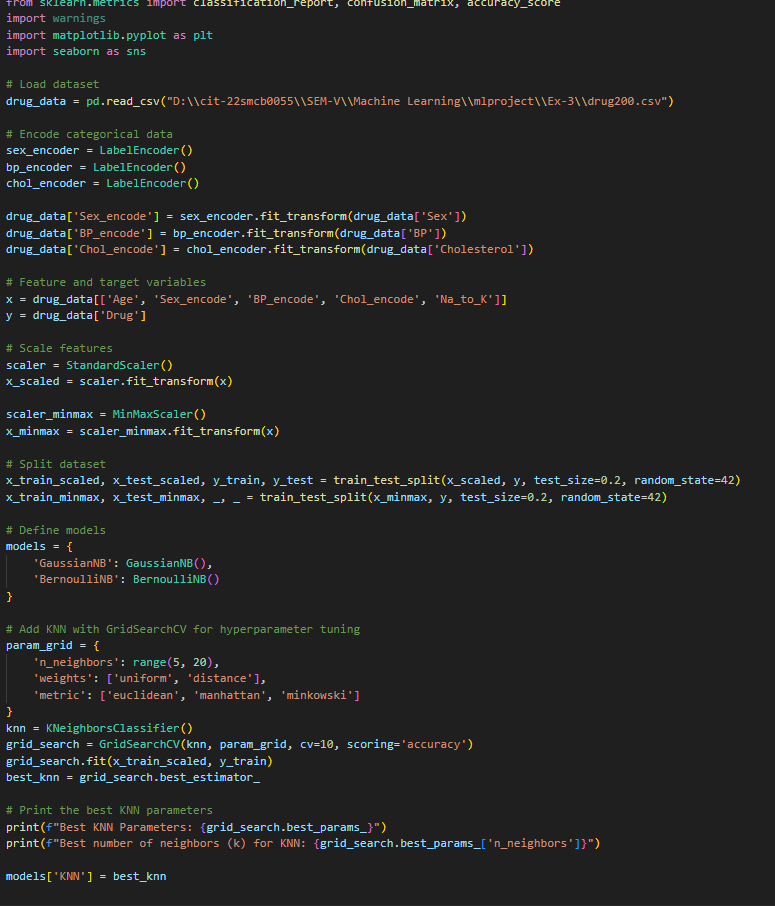
**Choosing Features in KNN:**

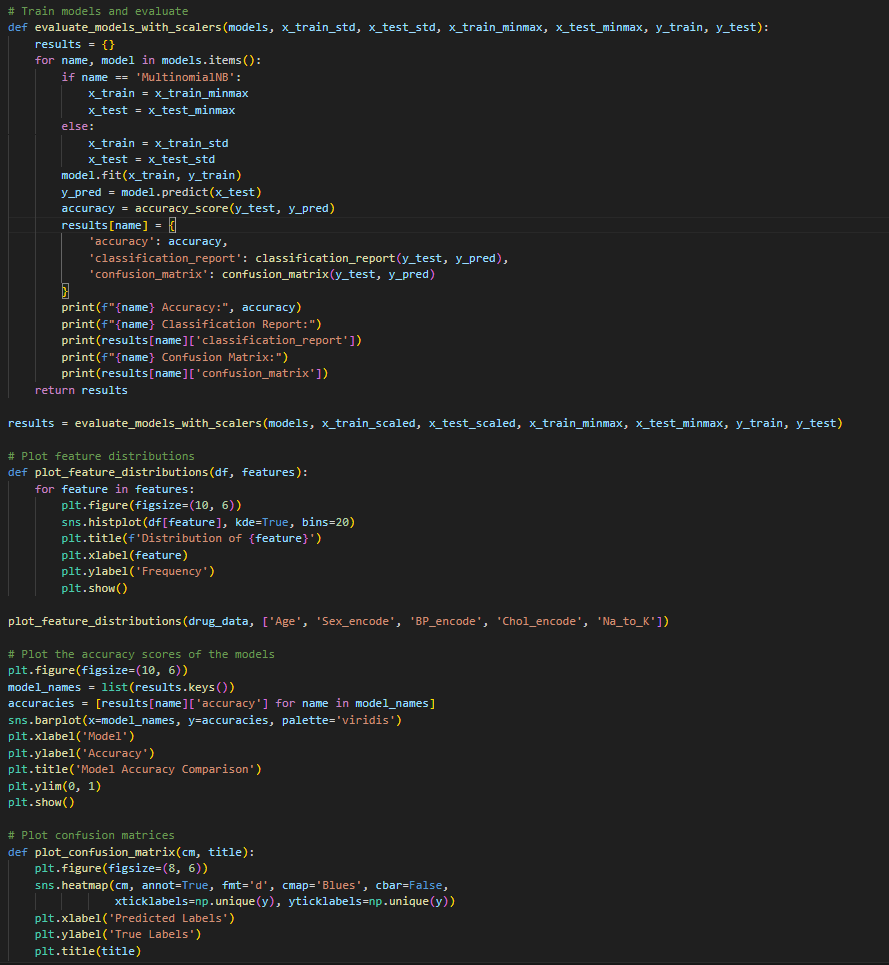
Features are selected based on their relevance to the target variable. In this exercise, we choose features like Age, Sex\_encode, BP\_encode, Chol\_encode, and Na\_to\_K based on their correlation with the target variable Drug.

**When Can Overfitting Occur?**

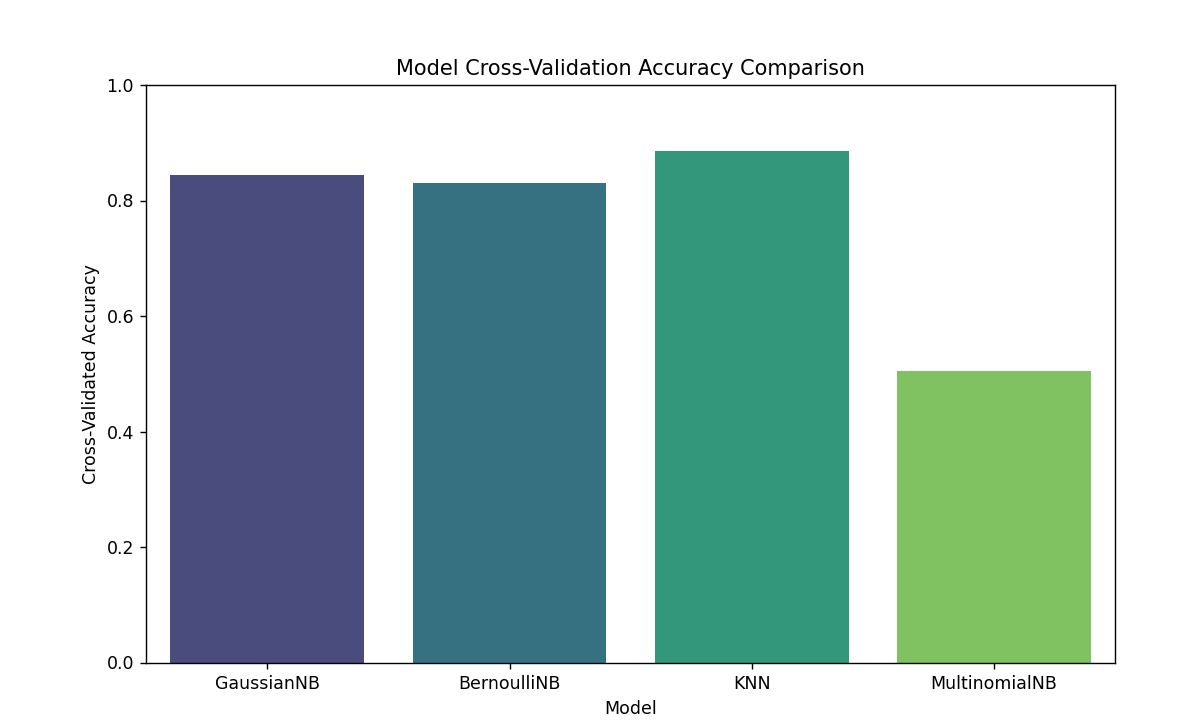
Overfitting can occur when the model is too complex and captures noise in the training data. In KNN, overfitting can happen if the number of Neighbour (k) is too small, leading the model to become overly sensitive to the training data. Grid search helps to find the optimal value of k to avoid overfitting.

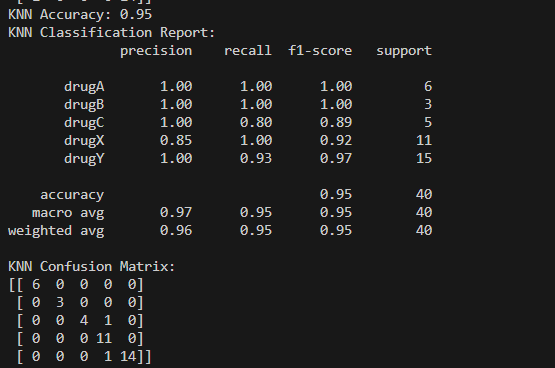
**Code:**

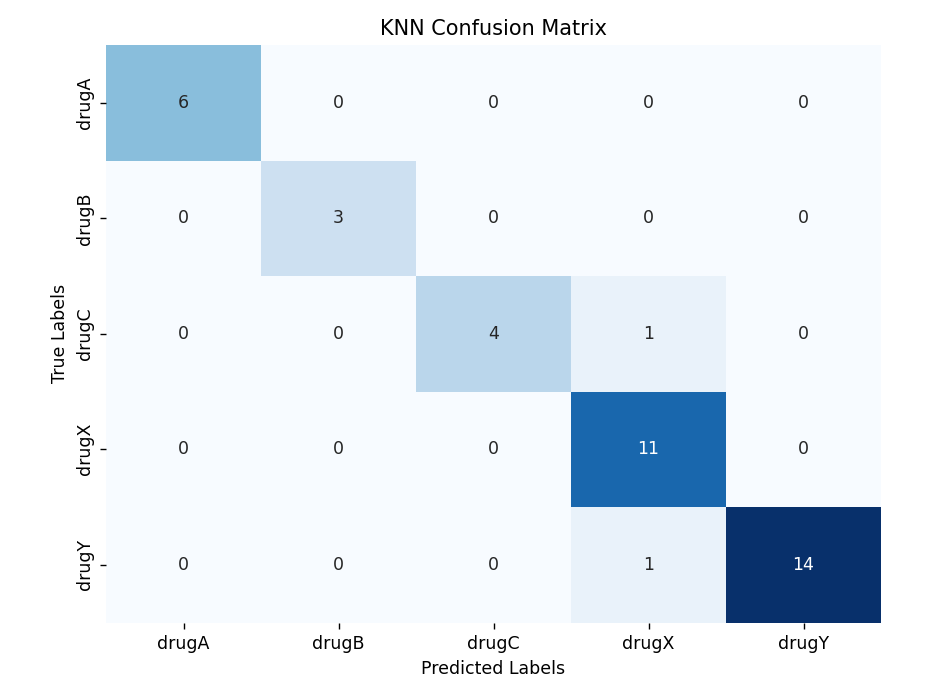
****



**Output:**

****

****

****

**Conclusion:**

K-Nearest Neighbour (KNN) is a straightforward and powerful tool for classification tasks. It

is easy to understand and implement, making it useful for various applications. By tuning

hyperparameters and evaluating the model using cross-validation, we can achieve reliable and

accurate predictions.