**Experiment 6**

**Shashwat Tripathi**

**D15A Batch C  
RollNo: 64**

**AIM : Classification modelling**

**THEORY :**

1. K-Nearest Neighbors (KNN):

Concept: KNN is a non-parametric, instance-based learning method. It classifies data points based on the majority vote of their k nearest neighbors in the training data.

Theory:

The algorithm assumes similar data points belong to the same class.

During prediction, a new data point is compared to its k nearest neighbors in the training data based on a distance metric (e.g., Euclidean distance).

The class label of the majority of these k neighbors is assigned to the new data point.

Advantages:

Simple to understand and implement.

Effective for both classification and regression problems.

No assumptions about the underlying data distribution.

Disadvantages:

Computationally expensive for large datasets due to distance calculations for each new data point.

Sensitive to irrelevant features and the choice of distance metric.

Performance can suffer from the "curse of dimensionality" in high-dimensional data.

2. Naive Bayes:

Concept: Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes independence between features, which simplifies calculations.

Theory:

The algorithm calculates the probability of a data point belonging to a particular class based on the individual probabilities of each feature value given that class.

Bayes' theorem is used to combine these individual probabilities into a posterior probability for each class.

The class with the highest posterior probability is assigned to the data point.

Advantages:

Efficient for classification with a large number of features.

Performs well even with limited training data.

Disadvantages:

The assumption of independence between features can be unrealistic for many datasets, potentially impacting accuracy.

Naive Bayes may struggle with rare classes or continuous features.

3. Support Vector Machine (SVM):

Concept: SVM is a discriminative classifier that finds the optimal hyperplane separating data points of different classes. The hyperplane maximizes the margin between the classes.

Theory:

SVMs map the data points to a higher-dimensional space if necessary using kernel functions.

The algorithm identifies a hyperplane in this high-dimensional space that maximizes the margin between the closest data points of each class (support vectors).

New data points are classified based on which side of the hyperplane they fall on.

Advantages:

Effective for high-dimensional data with kernel functions.

Good at handling noisy data due to focus on the margin.

Disadvantages:

Can be computationally expensive for large datasets.

Choosing the right kernel function can be crucial for performance.

SVM may not be well-suited for problems with many classes.

4. Decision Tree:

Concept: Decision Tree is a tree-like structure where each internal node represents a feature and each branch represents a possible outcome of a test on that feature. The leaf nodes represent the final class labels.

Theory:

The algorithm iteratively selects the best feature at each node to split the data based on a splitting criterion (e.g., information gain).

The splitting process continues until a stopping criterion is met (e.g., reaching a pure leaf node with all data points belonging to the same class).

New data points are classified by traversing the tree based on the feature values, reaching a leaf node that represents the predicted class.

Advantages:

Easy to interpret and visualize the decision-making process.

Can handle both categorical and continuous features.

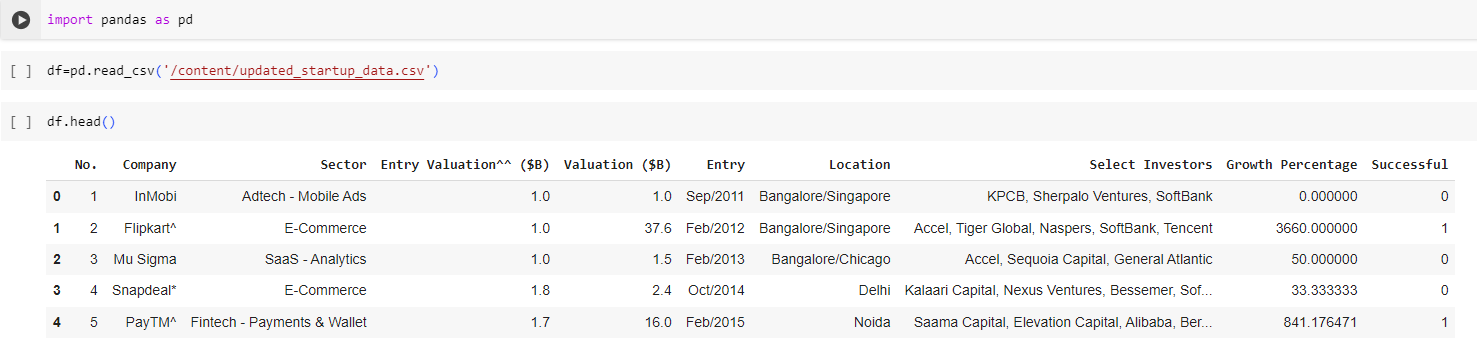
Disadvantages:

Decision trees can be prone to overfitting if not properly pruned.

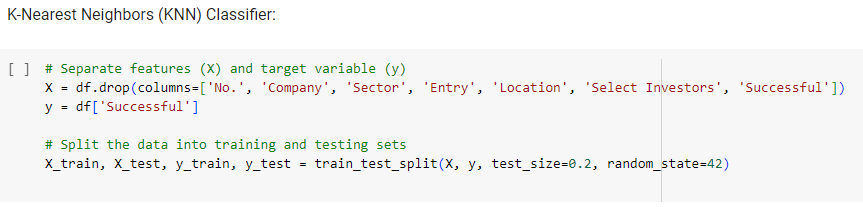
The performance can be sensitive to the order in which features are considered during splitting.

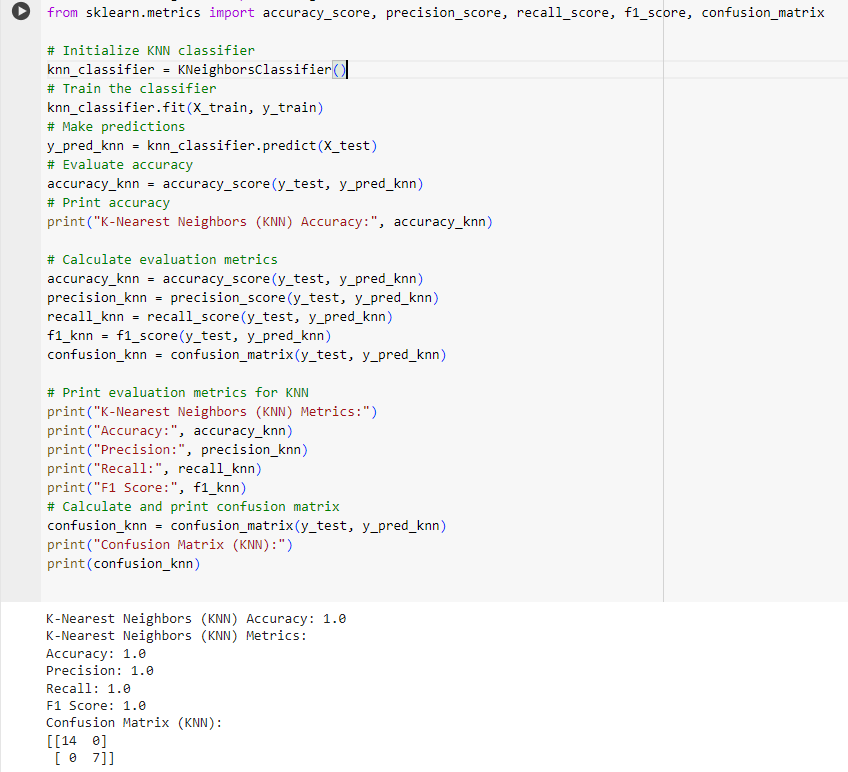
**OUTPUT :**

**Importing the Dataset** [https://www.kaggle.com/datasets/niekvanderzwaag/unicorn-startups](https://www.kaggle.com/datasets/niekvanderzwaag/unicorn-startups-cleaned)

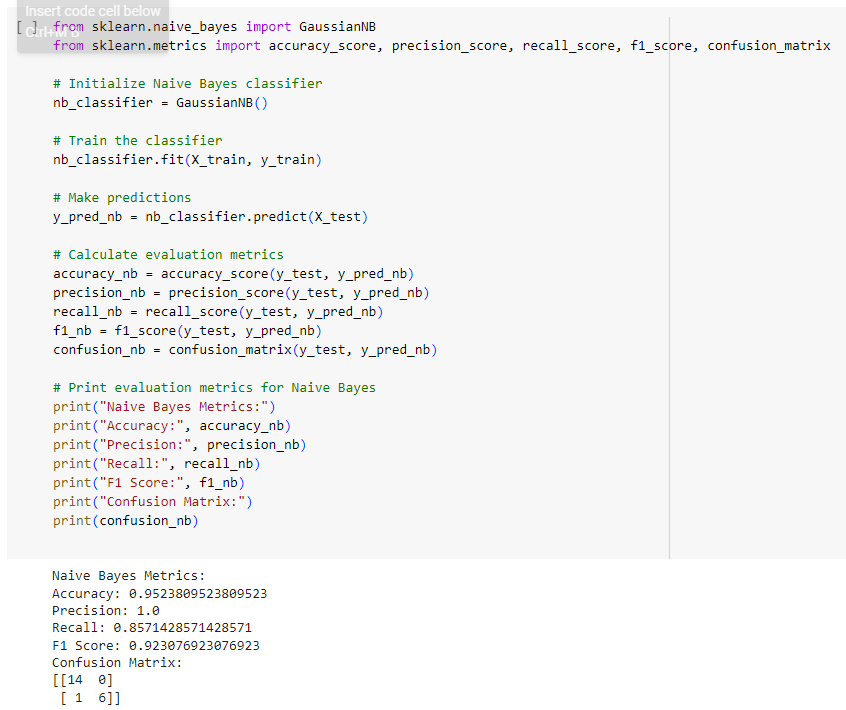


**KNN**

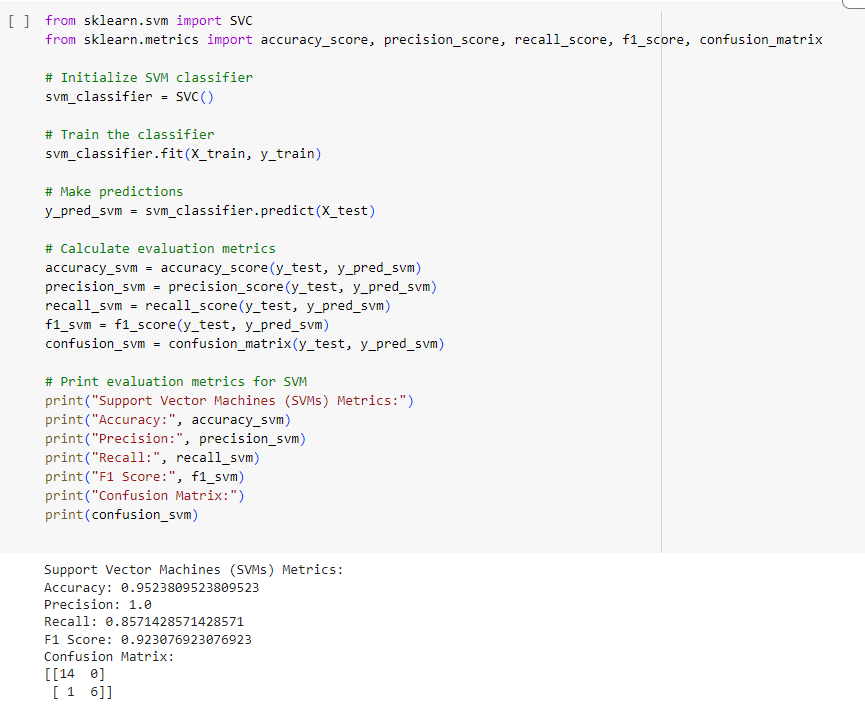




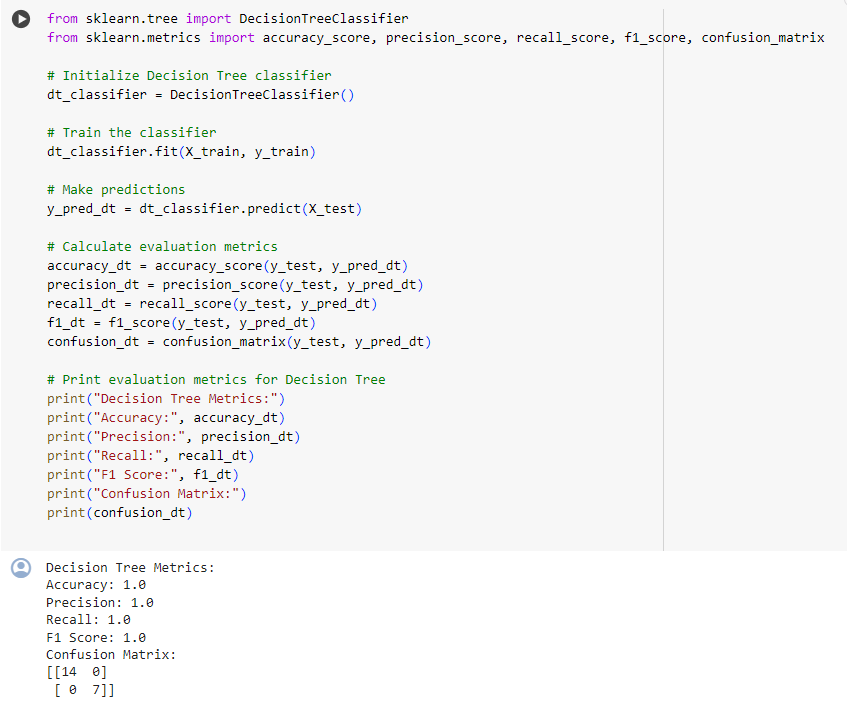
**Naive Bayes**



**SVM**



**Decision Tree Classifier**

****

**CONCLUSION : Thus, we have performed various regression techniques and found useful insights.**