Practical 5
19BCE248
2CS501

AIM:- Implement Naive Bayes and KNN classifiers:

Naïve-Bayes's all variants: Multivariate Bernoulli, Multinomial and Gaussian using sklearn.

KNN using sklearn.

About Dataset (Iris):
Iris is a dataset about flowers which is most popular for classification problems. It consists of four columns namely [sepal length (cm), sepal width (cm),petal length (cm),petal width (cm)]. The target value is species of flower which are of three types:
1.Setosa
2.Versicolor
3.Virginica

Rows: 150
Columns: 4
The target column describes the target variable i.e. species of flower according to parameters.

Preprocessing:
● Firstly dividing the entire dataset into Training and Testing dataset.
● As the input variables are continuous we need to categorize it which varies from model to model.

Applying three main variants of Naive Bayes and analyzing accuracy on the same dataset:

Multinomial:

```
Accuracy 1.0
Report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         7
           1       1.00      1.00      1.00        12
           2       1.00      1.00      1.00        11

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30
```

Bernoulli:

```
Accuracy 0.6
Report
              precision    recall  f1-score   support

           0       0.58      1.00      0.74         7
           1       0.00      0.00      0.00        12
           2       0.61      1.00      0.76        11

    accuracy                           0.60        30
   macro avg       0.40      0.67      0.50        30
weighted avg       0.36      0.60      0.45        30
```

Gaussian:

```
Accuracy 1.0
Report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         7
           1       1.00      1.00      1.00        12
           2       1.00      1.00      1.00        11

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30
```

Also the dataset from the textbook was taken into consideration and performed bayesian classification on that too.

	age	income	student	credit_rating
0	0	2	0	0
1	0	2	0	1
2	1	2	0	0
3	2	1	0	0
4	2	0	1	0
5	2	0	1	1
6	1	0	1	1
7	0	1	0	0
8	0	0	1	0
9	2	1	1	0
10	0	1	1	1
11	1	1	0	1
12	1	2	1	0
13	2	1	0	1

Spam/Non-Spam Messages(Extra Part):

```
Accuracy 0.9883408071748879
Report
              precision    recall  f1-score   support

         ham       0.99      1.00      0.99       970
        spam       0.98      0.93      0.95       145

    accuracy                           0.99      1115
   macro avg       0.98      0.96      0.97      1115
weighted avg       0.99      0.99      0.99      1115
```

Used multinomial Naive bayes which gave the highest accuracy among all. Used CountVectorizer for preprocessing dataset.

KNN

Dataset taken is the same as the above Iris one.

Simple KNN:

```
0.9666666666666667
```

```
print(cr)
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         9
           1       1.00      0.91      0.95        11
           2       0.91      1.00      0.95        10

    accuracy                           0.97        30
   macro avg       0.97      0.97      0.97        30
weighted avg       0.97      0.97      0.97        30
```

KNN with GridSearchCV:

```
ac
```

```
1.0
```

```
print(cr)
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         9
           1       1.00      1.00      1.00        11
           2       1.00      1.00      1.00        10

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30
```

```
gsv.best_params_
```

```
{'n_neighbors': 10, 'weights': 'uniform'}
```

Conclusion:

From both types of classifier we learned which datasets suit which ML model. Also various techniques to increase accuracy such as preprocessing through sklearn library and GridSearchCV were explored to great extent during these practicals. And also one thing can be concluded from classification problems is that accuracy must remain very high as a single misclassification leads to bad inference unlike regression problems.