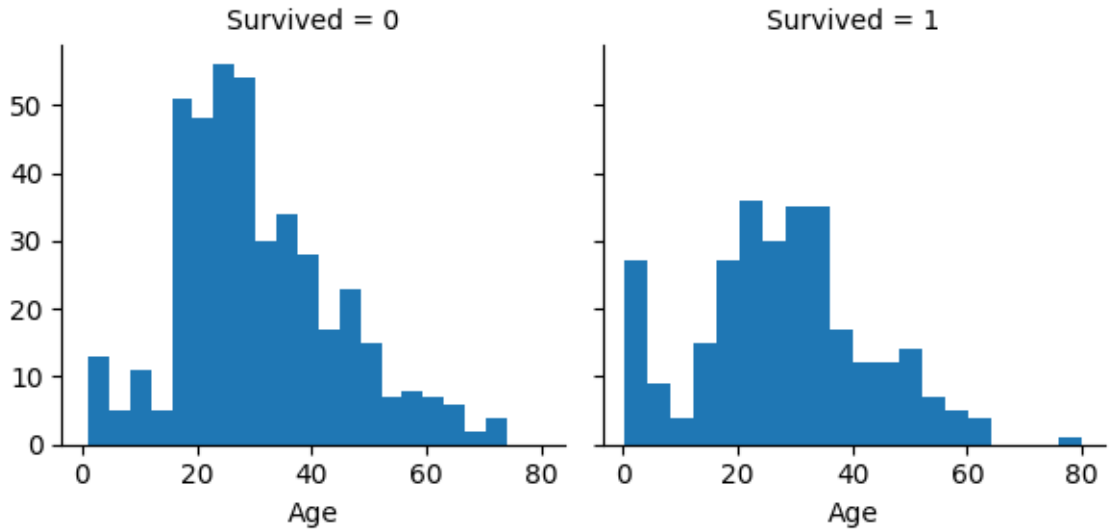
Assignment 3 Documentation

Kavyasri

700728990

1.Titanic dataset

- ✓ Read train and test dataset using pandas.
- ✓ Check head to see columns and type of data.
- ✓ Check for data imbalance using value_counts on y variable.
- ✓ Plot histogram for survived vs age.



- ✓ Drop unnecessary columns and separate y variable. Out: X_train, Y_train, X_test
- ✓ Check for null values and fill them with mean or median accordingly.
- ✓ Convert categorical columns to numerical with label encoding.
- ✓ Sample data after preprocessing is as below.

	Pclass	Sex	Age	Fare	Embarked
0	3	male	34.5	7.8292	Q
1	3	female	47.0	7.0000	S
2	2	male	62.0	9.6875	Q
3	3	male	27.0	8.6625	S
4	3	female	22.0	12.2875	S

- ✓ Fit 4 naïve bayes models(gaussian, multinomial, Bernoulli, complement) on X_train and Y_train. Use scikit-learn for the same.

- ✓ Predict on X_test.
- ✓ Since X_test original labels are not available we calculate accuracy on train data itself.
- ✓ Find classification_report, confusion_matrix, accuracy_score for each of the 4 models using scikit-learn
- ✓ Gaussian:

```
              precision    recall  f1-score   support

           0       0.82      0.81      0.82       549
           1       0.70      0.72      0.71       342

    accuracy                           0.78       891
   macro avg       0.76      0.77      0.77       891
weighted avg       0.78      0.78      0.78       891

[[445 104]
 [ 95 247]]
accuracy is 0.77665544332211
```

- ✓ Multinomial:

```
              precision    recall  f1-score   support

           0       0.72      0.83      0.77       549
           1       0.64      0.48      0.55       342

    accuracy                           0.70       891
   macro avg       0.68      0.65      0.66       891
weighted avg       0.69      0.70      0.68       891

[[457  92]
 [179 163]]
accuracy is 0.6958473625140292
```

- ✓ Bernoulli:

```
              precision    recall  f1-score   support

           0       0.81      0.85      0.83       549
           1       0.74      0.68      0.71       342

    accuracy                           0.79       891
   macro avg       0.78      0.77      0.77       891
weighted avg       0.78      0.79      0.78       891

[[468  81]
 [109 233]]
accuracy is 0.7867564534231201
```

- ✓ Complement:

```
              precision    recall  f1-score   support

           0       0.72      0.83      0.77       549
           1       0.64      0.48      0.55       342

    accuracy                           0.70       891
   macro avg       0.68      0.66      0.66       891
weighted avg       0.69      0.70      0.69       891
```

```
[[455  94]
 [177 165]]
accuracy is 0.6958473625140292
```

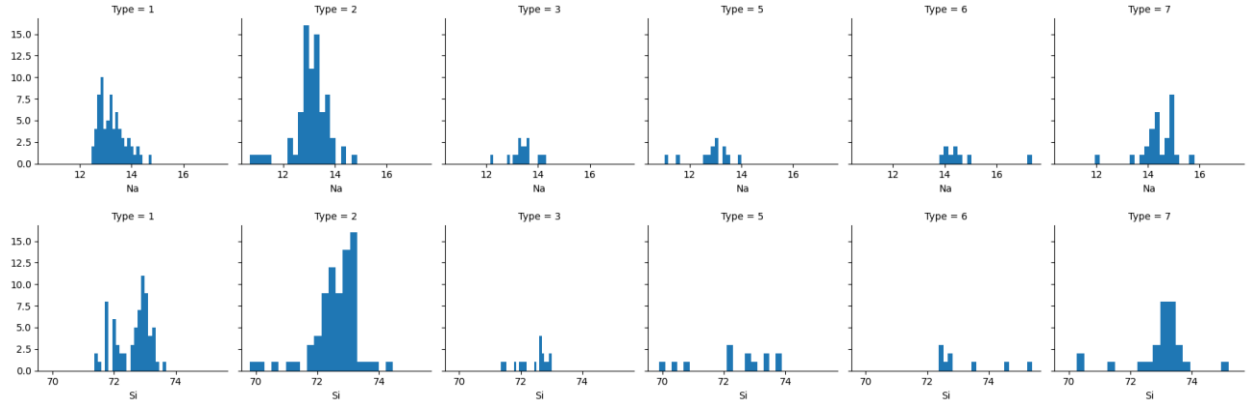
- ✓ Best fit model is Bernoulli naïve as it works well on binary dataset.

2.Glass dataset

- ✓ Read glass dataset using pandas.
- ✓ Check head to see columns and type of data.

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type		
0	1.52101		13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.0	1
1	1.51761		13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.0	1
2	1.51618		13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.0	1
3	1.51766		13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.0	1
4	1.51742		13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.0	1

- ✓ Check for data imbalance using value_counts on y variable. This dataset is highly **imbalanced**.
- ✓ Plot histogram for Type vs Na and Type vs Si to observe the relation of dependant variables with independent variable.



- ✓ Separate X and Y data.
- ✓ Split data for train and test with 20% test using train_test_split. Set random_set to get same split on repetition. Out: X_train, Y_train, X_test, Y_test.
- ✓ Fit 4 naïve bayes models(gaussian, multinomial, Bernoulli, complement) , SVC and Linear SVC on X_train and Y_train. Use scikit-learn for the same.
- ✓ Predict on X_test.
- ✓ Calculate accuracy on test data with Y_test and Y_pred_test.
- ✓ Find classification_report, confusion_matrix, accuracy_score for each of the 4 models using scikit-learn.
- ✓ Gaussian:

```
precision    recall  f1-score   support

           1       0.19      0.44      0.27         9
           2       0.33      0.16      0.21        19
           3       0.33      0.20      0.25         5
```

```
              5       0.00      0.00      0.00         2
              6       0.67      1.00      0.80         2
              7       1.00      1.00      1.00         6

       accuracy                           0.37        43
      macro avg       0.42      0.47      0.42        43
   weighted avg       0.40      0.37      0.36        43

   [[ 4  3  1  0  1  0]
    [14  3  1  1  0  0]
    [ 3  1  1  0  0  0]
    [ 0  2  0  0  0  0]
    [ 0  0  0  0  2  0]
    [ 0  0  0  0  0  6]]
   accuracy is 0.37209302325581395
```

✓ Multinomial:

```
   precision    recall  f1-score   support

              1       0.28      0.89      0.42         9
              2       0.40      0.11      0.17        19
              3       0.00      0.00      0.00         5
              5       0.00      0.00      0.00         2
              6       0.00      0.00      0.00         2
              7       0.67      1.00      0.80         6

       accuracy                           0.37        43
      macro avg       0.22      0.33      0.23        43
   weighted avg       0.33      0.37      0.27        43

   [[ 8  1  0  0  0  0]
    [16  2  0  0  0  1]
    [ 5  0  0  0  0  0]
    [ 0  0  0  0  0  2]
    [ 0  2  0  0  0  0]
    [ 0  0  0  0  0  6]]
   accuracy is 0.37209302325581395
```

✓ Bernoulli:

```
   precision    recall  f1-score   support

              1       0.27      0.89      0.41         9
              2       0.29      0.11      0.15        19
              3       0.00      0.00      0.00         5
              5       0.00      0.00      0.00         2
              6       0.00      0.00      0.00         2
              7       0.83      0.83      0.83         6

       accuracy                           0.35        43
      macro avg       0.23      0.30      0.23        43
   weighted avg       0.30      0.35      0.27        43

   [[ 8  1  0  0  0  0]
```

```
 [16  2  0  0  0  1]
 [ 5  0  0  0  0  0]
 [ 0  2  0  0  0  0]
 [ 0  2  0  0  0  0]
 [ 1  0  0  0  0  5]]
accuracy is 0.3488372093023256
```

- ✓ Complement:

```
precision    recall  f1-score   support

           1       0.28      1.00      0.44         9
           2       0.00      0.00      0.00        19
           3       0.00      0.00      0.00         5
           5       1.00      0.50      0.67         2
           6       0.50      0.50      0.50         2
           7       0.75      1.00      0.86         6

    accuracy                           0.40        43
   macro avg       0.42      0.50      0.41        43
weighted avg       0.23      0.40      0.27        43

[[ 9  0  0  0  0  0]
 [17  0  0  0  1  1]
 [ 5  0  0  0  0  0]
 [ 0  0  0  1  0  1]
 [ 1  0  0  0  1  0]
 [ 0  0  0  0  0  6]]
accuracy is 0.3953488372093023
```

- ✓ SVC:

```
precision    recall  f1-score   support

           1       0.21      1.00      0.35         9
           2       0.00      0.00      0.00        19
           3       0.00      0.00      0.00         5
           5       0.00      0.00      0.00         2
           6       0.00      0.00      0.00         2
           7       0.00      0.00      0.00         6

    accuracy                           0.21        43
   macro avg       0.03      0.17      0.06        43
weighted avg       0.04      0.21      0.07        43

[[ 9  0  0  0  0  0]
 [19  0  0  0  0  0]
 [ 5  0  0  0  0  0]
 [ 2  0  0  0  0  0]
 [ 2  0  0  0  0  0]
 [ 6  0  0  0  0  0]]
accuracy is 0.20930232558139536
```

- ✓ Linear SVC(max_iter=1000):

```
precision    recall  f1-score   support

           1       0.31      1.00      0.47         9
           2       1.00      0.11      0.19        19
           3       0.00      0.00      0.00         5
           5       0.50      0.50      0.50         2
           6       0.50      1.00      0.67         2
           7       1.00      1.00      1.00         6

    accuracy                           0.47        43
   macro avg       0.55      0.60      0.47        43
weighted avg       0.69      0.47      0.38        43

[[ 9  0  0  0  0  0]
 [15  2  0  1  1  0]
 [ 5  0  0  0  0  0]
 [ 0  0  0  1  1  0]
 [ 0  0  0  0  2  0]
 [ 0  0  0  0  0  6]]
accuracy is 0.46511627906976744
```

- ✓ Complement naïve bayes works well for imbalanced dataset. Overall Linear SVC has better accuracy because data may be linearly related.