Out[104]: Click here to toggle on/off the raw code.

Question 2

Abalone Datasets

Loading the dataset and normalizing it

Out[78]:
```
array([[0.51351351, 0.5210084 , 0.0840708 , ..., 0.15030262, 0.1323239 ,
        0.14798206],
       [0.37162162, 0.35294118, 0.07964602, ..., 0.06624075, 0.06319947,
        0.06826109],
       [0.61486486, 0.61344538, 0.11946903, ..., 0.17182246, 0.18564845,
        0.2077728 ],
       ...,
       [0.70945946, 0.70588235, 0.18141593, ..., 0.3527236 , 0.37788018,
        0.30543099],
       [0.74324324, 0.72268908, 0.13274336, ..., 0.35642233, 0.34298881,
        0.29347285],
       [0.85810811, 0.84033613, 0.17256637, ..., 0.63517149, 0.49506254,
        0.49177877]])
```

Applying PCA pre-processing on Abalone dataset and select first 3 principal components

Out[79]:

	PC1	PC2	PC3
0	-0.230816	-0.026563	-0.006786
1	-0.497671	0.043791	0.003049
2	-0.068857	-0.081454	0.011720
3	-0.230997	-0.012962	0.004214
4	-0.532797	0.057362	-0.000513
...	...	...	...
4172	0.100632	-0.034549	-0.011468
4173	0.128141	-0.023082	-0.028686
4174	0.273938	0.019037	-0.025086
4175	0.262282	-0.027659	-0.045737
4176	0.739028	0.130322	-0.046922

4177 rows × 3 columns

Applying LDA pre-processing on Abalone dataset and select 3 linear descriptors

Out[80]:

	0	1	2
Rings			
15	-0.791003	-0.235208	0.359351
7	-2.355522	0.336978	0.214024
9	0.766719	-0.246564	1.129422
10	-0.611434	0.098075	0.230542
7	-2.674301	0.527509	0.102575
...	...	...	...
11	0.921330	-0.612381	-0.272399
10	0.425796	-0.894428	-0.034727
9	1.064523	-0.385654	-0.787231
10	0.840757	-1.513723	-0.864217
12	0.843580	0.352389	-2.262564

4177 rows × 3 columns

Wine Datasets

Out[81]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcoho
0	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.
1	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.
2	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.
3	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.
4	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.
...	...	...	...	...	...	...	...	...	...	...	.
95	7.1	0.260	0.29	12.4	0.044	62.0	240.0	0.9969	3.04	0.42	9.
96	6.0	0.340	0.66	15.9	0.046	26.0	164.0	0.9979	3.14	0.50	8.
97	8.6	0.265	0.36	1.2	0.034	15.0	80.0	0.9913	2.95	0.36	11.
98	9.8	0.360	0.46	10.5	0.038	4.0	83.0	0.9956	2.89	0.30	10.
99	6.0	0.340	0.66	15.9	0.046	26.0	164.0	0.9979	3.14	0.50	8.

100 rows × 13 columns

Out[82]: ((6497, 12), (6497,))

Applying PCA pre-processing on Wine dataset and select first 3 principal components

Out[83]:

	PC1	PC2
0	-0.298897	-0.337622
1	-0.241913	-0.084556
2	-0.225052	-0.036821
3	-0.290807	-0.123910
4	-0.290807	-0.123910
...	...	...
6492	0.793106	0.054298
6493	0.789921	0.151012
6494	0.790787	0.111938
6495	0.808678	0.031407
6496	0.751111	0.102306

6497 rows × 2 columns

Applying LDA pre-processing on Abalone dataset and select first 3 linear descriptors

```
(6497, 2)
```

Out[84]:

quality	0	1
6	0.752078	-1.466209
6	1.445150	0.392049
6	-0.123015	0.911451
6	0.288961	-0.721769
6	0.288961	-0.721769
...	...	...
5	0.512278	-0.224430
6	-0.514707	-0.597340
6	-0.231160	-0.831907
5	0.630811	0.158871
6	-0.668993	-2.296580

6497 rows × 2 columns

Abalone - raw dataset - Multimonial naive bayes:

If we apply Standardisation to the Abalone dataset, values become negative and that is not acceptable as a values to Naive Bayes classifiers. Hence, we need to use MinMaxScaler (Normalization) to scale down values only within 0 and 1. However, this will decrease the accuracy of the model.

The accuracy of a model on the Raw abalone dataset has significantly reduced from 26% to 16.5% with Naive Bayes compared to KNN using 10 neighbors measured in the previous assignment. While it's likely that neither algorithm is adequate for predicting the abalone age, the KNN model is more accurate so far

Wine - Raw dataset - Multinomial Naive Bayes: Mean accuracy

Out[86]: `0.46159223071001365`

Out[88]: [0.16495086382259405, 0.414954106709303]

KNN Algorithm has worked slightly better on the Wine (Raw) dataset compared to Multinomial Naive Bayes as the accuracy has gone down from 46.15% to an average of 41.5% accross 5-folds. A combination of Standardisation and then KNN has no significant effect on the accuracy improvement.

Abalone - Raw - Complement NB

Out[89]: 0.17500329484571525

Wine dataset - raw - complement NB

Out[90]: 0.38971090187718366

Test accuracy of Raw abalone

```
Cross-validation accuracy of Raw abalone using Multinomial Naive Bayes classi
fier: 16.37%
Cross-validation accuracy of Raw abalone using Complement Naive Bayes classif
ier: 18.14%
Test accuracy of Raw abalone using Multinomial Naive Bayes classifier: 16.99%
Test accuracy of Raw abalone using Complement Naive Bayes classifier: 19.14%
```

Test accuracy of PCA processed abalone

```
Cross-validation accuracy of processed abalone using Multinomial Naive Bayes
classifier with PCA: 16.37%
Cross-validation accuracy of processed abalone using Complement Naive Bayes c
lassifier with PCA: 18.26%
Test accuracy of processed abalone using Multinomial Naive Bayes classifier w
ith PCA: 16.99%
Test accuracy of processed abalone using Complement Naive Bayes classifier wi
th PCA: 17.22%
```

Test accuracy of LDA processed abalone

```
Cross-validation accuracy of processed abalone using Multinomial Naive Bayes
classifier with LDA: 16.37%
Cross-validation accuracy of processed abalone using Complement Naive Bayes c
lassifier with LDA: 23.97%
Test accuracy of processed abalone using Multinomial Naive Bayes classifier L
DA: 16.99%
Test accuracy of processed abalone using Complement Naive Bayes classifier LD
A: 21.53%
```

Wine Dataset implementation

Out[94]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcoh
0	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.
1	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.
2	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.
3	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.
4	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.
...	...	...	...	...	...	...	...	...	...	...	.
95	7.1	0.260	0.29	12.4	0.044	62.0	240.0	0.9969	3.04	0.42	9.
96	6.0	0.340	0.66	15.9	0.046	26.0	164.0	0.9979	3.14	0.50	8.
97	8.6	0.265	0.36	1.2	0.034	15.0	80.0	0.9913	2.95	0.36	11.
98	9.8	0.360	0.46	10.5	0.038	4.0	83.0	0.9956	2.89	0.30	10.
99	6.0	0.340	0.66	15.9	0.046	26.0	164.0	0.9979	3.14	0.50	8.

100 rows × 13 columns

```
0       15
1        7
2        9
3       10
4        7
        ..
4172    11
4173    10
4174     9
4175    10
4176    12
Name: Rings, Length: 4177, dtype: int64
The normalized dataset is:
 [[0.51351351 0.5210084  0.0840708  ... 0.15030262 0.1323239  0.14798206]
 [0.37162162 0.35294118 0.07964602 ... 0.06624075 0.06319947 0.06826109]
 [0.61486486 0.61344538 0.11946903 ... 0.17182246 0.18564845 0.2077728 ]
 ...
 [0.70945946 0.70588235 0.18141593 ... 0.3527236  0.37788018 0.30543099]
 [0.74324324 0.72268908 0.13274336 ... 0.35642233 0.34298881 0.29347285]
 [0.85810811 0.84033613 0.17256637 ... 0.63517149 0.49506254 0.49177877]]
```

Test Accuracy of raw wine

```
Cross-validation accuracy of raw wine using Multinomial Naive Bayes classifie
r: 43.54%
Cross-validation accuracy of raw wine using Complement Naive Bayes classifie
r: 47.51%
Test accuracy of raw wine using Multinomial Naive Bayes classifier: 44.62%
Test accuracy of raw wine using Complement Naive Bayes classifier: 48.54%
```

Test accuracy of PCA processed wine

```
Cross-validation accuracy of processed wine using Multinomial Naive Bayes cla
ssifier with PCA: 43.43%
Cross-validation accuracy of processed wine using Complement Naive Bayes clas
sifier with PCA: 45.41%
Test accuracy of processed wine using Multinomial Naive Bayes classifier with
PCA: 44.54%
Test accuracy of processed wine using Complement Naive Bayes classifier with
PCA: 46.92%
```

Test accuracy of LDA processed wine

```
Cross-validation accuracy of processed wine using Multinomial Naive Bayes cla
ssifier with LDA: 43.43%
Cross-validation accuracy of processed wine using Complement Naive Bayes clas
sifier with LDA: 0.54%
Test accuracy of processed wine using Multinomial Naive Bayes classifier with
LDA: 44.54%
Test accuracy of processed wine using Complement Naive Bayes classifier with
LDA: 0.15%
```

Conclusion for Abalone Dataset

Type *Markdown* and LaTeX: α^2

The above accuracies summarizes the test accuracy of two algorithms (Multinomial Naive Bayes and Complement Naive Bayes) on three versions of the abalone dataset: raw, PCA-preprocessed (with 3 principal components), and LDA-preprocessed (with 3 linear discriminants).

For Multinomial Naive Bayes, the test accuracy remains the same (16.99%) across all three versions of the dataset.

For Complement Naive Bayes, the test accuracy is highest on the LDA-preprocessed dataset (21.53%), followed by the raw dataset (19.14%), and lowest on the PCA-preprocessed dataset (17.22%). This suggests that LDA pre-processing is more effective for improving the performance of Complement Naive Bayes on the abalone dataset compared to PCA pre-processing.

Conclusion for Wine Dataset

For the Multinomial Naive Bayes algorithm, there is not much difference in performance between the raw wine dataset and the dataset preprocessed with PCA or LDA. The accuracy remains around 44-45% for all three settings.

For the Complement Naive Bayes algorithm, the performance is significantly better on the raw wine dataset compared to the preprocessed datasets. The accuracy is around 48.5% for the raw dataset, but drops to around 47% for the dataset preprocessed with PCA, and drops even further to 0.15% for the dataset preprocessed with LDA.