Cara Evangeline Chandra Mohan
190539421

Data Mining - Lab Assignment 4

1.Now starting from this 4-attribute subset, find the best 3, 2, 1 attribute subset, filling in the table below. Which sized subset, and which set of attributes yields the best accuracy?

Subset size	Attributes Selected	Accuracy	Attributes Removed
5	All: W, P, Hol, Vac, Health	85.9649	None
4	W, P, Hol, Health	89.4737	Vac
3	W, P, Hol	91.2281	Vac, Health
2	W, P	85.9649	Vac, Health, Hol
1	P	80.7018	Vac, Health, Hol, W

Subset size of 3 {Wage, Pension, Holidays} gives the best accuracy.

2.How many feature combinations did you try? How many combinations of features are there in total? Give an example of a combination of features that you did NOT try when doing backward selection.

Subset size = 4
Number of feature combinations tried = 5
Number of feature combinations in total = 5

Subset size = 3
Number of feature combinations tried = 4
Number of feature combinations in total = 10
Combination of feature not tried = {W, P, Hol, Vac}

Subset size = 2
Number of feature combinations tried = 3
Number of feature combinations in total = 10
Combination of feature not tried = {W, P, Health}

Subset size = 1
Number of feature combinations tried = 2
Number of feature combinations in total = 5
Combination of feature not tried = {P, Hol}

Cara Evangeline Chandra Mohan
190539421

3.How many and which attributes are selected? Do they match the results from Section 2?

Three attributes are selected, ie.Wage, Pension and holidays.

```
Selected attributes: 1,2,3 : 3
                     wage-increase-first-year
                     pension
                     statutory-holidays
```

And the results are same as observed in Section 2.

4.Which attributes does it pick (and hence which ones are discarded?)

Attributes selected by WrapperSubsetEval for Naive Bayes are Petal Length and Petal width. The attributes discarded are Sepal Length, Sepal width and the three copies of Sepal Length and Sepal Width. So in total 8 attributes are discarded.

```
Selected attributes: 3,4 : 2
                     petallength
                     petalwidth
```

5.Use the data used to produce the above plot to find out what number of PCs is required to explain 99% of the data variance (achieve 99% reconstruction accuracy). What # is this and does it match the value from Q8? Provide a short discussion.

```
[0.85279152 0.88362601 0.9070117  0.91696321 0.92614474 0.93360688
 0.93921873 0.94433953 0.9478992  0.95126595 0.9544066  0.95702192
 0.95943594 0.96168741 0.9635197  0.96518962 0.96668353 0.96807027
 0.96937869 0.97062105 0.97185131 0.97295654 0.97395931 0.9748821
 0.97574049 0.97658673 0.9774194  0.97820148 0.97890606 0.97959923
 0.98028185 0.98090676 0.9815193  0.98209686 0.98263086 0.98313697
 0.98363809 0.98411418 0.98456907 0.98500439 0.98542918 0.98582616
 0.98620484 0.98657221 0.98692853 0.98727767 0.98761761 0.98793595
 0.98823261 0.98852422 0.98880657 0.98908575 0.98936109 0.9896329
 0.98989749 0.99015207 0.99039802 0.99063439 0.99086522 0.99109389
 0.99130889 0.99152082 0.99172843 0.99193398 0.9921286  0.99231927
 0.99250404 0.99268844 0.99286478 0.99303563 0.99320476 0.9933715
 0.99353281 0.99368924 0.99384434 0.9939935  0.9941391  0.99428236
 0.99442416 0.99456281 0.99469879 0.99483168 0.99496263 0.99508875
 0.99521348 0.99533766 0.9954589  0.99557672 0.99569218 0.99580495
 0.99591647 0.99602535 0.99613115 0.99623633 0.99633662 0.99643581
 0.9965333  0.99662875 0.99672343 0.99681691 0.99690788 0.99699793
 0.99708506 0.99717148 0.99725732 0.99733978 0.99742061 0.99750014
 0.9975774  0.99765446 0.99772921 0.99780254 0.99787235 0.99794137
 0.99800972 0.99807625 0.99814184 0.99820639 0.99827044 0.99833314
 0.99839486 0.9984547  0.99851259 0.99857013 0.99862668 0.99868043
 0.99873343 0.99878503 0.99883585 0.99888608 0.99893563 0.99898511
 0.99903197 0.99907795 0.99912336 0.99916762 0.9992111  0.99925405
 0.99929625 0.99933675 0.99937667 0.99941555 0.99945314 0.9994903
 0.99952681 0.99956319 0.99959837 0.99963218 0.99966545 0.99969789
 0.99972914 0.99975974 0.99978962 0.99981762 0.99984598 0.99987116
```

At row 10 and column 2 (ie. (9*6 + 2) = 56) we can see that reconstruction accuracy is 99.01%. Hence Number of principal components required is 56 and it is the same as that obtained from Q8. The reconstruction error that we get in Q8 is 0.0098 for nPCA =56.

Cara Evangeline Chandra Mohan
190539421

6.Which number of PCA dimensions gets the maximum face recognition accuracy? Is it better or worse than the accuracy when classifying the raw images? Why? (What factors contribute to this?) Provide a brief discussion.

- PCA dimension of 27 gets the maximum face recognition accuracy.

Accuracy score of 1-NN when nPCA = 27 is 0.707

- For raw images: Number of PCA = 4096

Accuracy score of 1-NN when nPCA = 4096 is 0.671

Using Principal Component Analysis for dimensionality reduction for classification of images gives a better accuracy than classifying raw images. In PCA, the features of low variance is eliminated so that the unnecessary features doesn't deviate from the actual class values whereas while classifying raw images, the redundant values makes the classifier less accurate.