LUT
Lappeenranta
University of Technology

A220A0752 Analytics for Business

Individual Assignment

Dilhara Liyanaaratchi

000286947

Contents

Task 1

Initially the summary was checked for the data set and as there were no missing values, proceeded with the tasks.

Task 1.1

In order to observe the relationship between air temperature(T_mu) and atmospheric pressure (P_mu) based on the obersed groups a scatter plot was created.

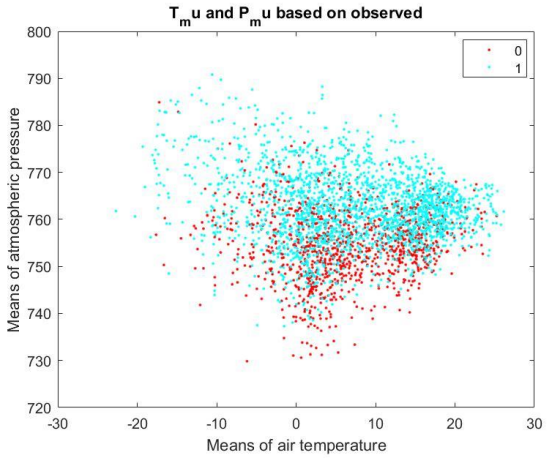

Figure 1: Scatter plot

Task 1.2

Whereas, the distribution of the minimum air temperature was plotted in a histogram and as per the following figure, it is left skewed.

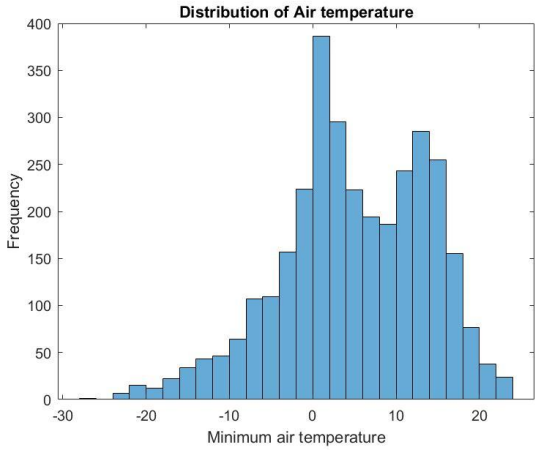

Figure 2: Minimum air temperature

Task 1.3

The maximum values of the variable dewpoint temperature (Td_mu) was calculated based on the two groups of observed which are dry (1) and not dry(0).

- Max value of Td_mu of the group dry: 21.4625
- Max value of Td_mu of the group not dry: 20.3

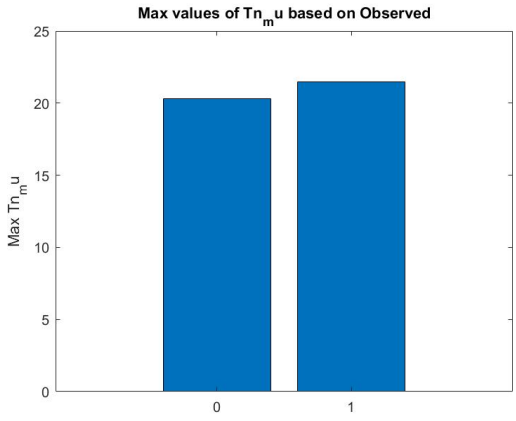

Figure 3: Max value of Td_mu

Task 1.4

On the other hand, the mean value of P_var variable which indicate the atmospheric pressure was calculated for the two groups of observed attributes.

- Mean value of P_var of group dry: 3.0735
- Mean value of P_var of group not dry: 5.4376

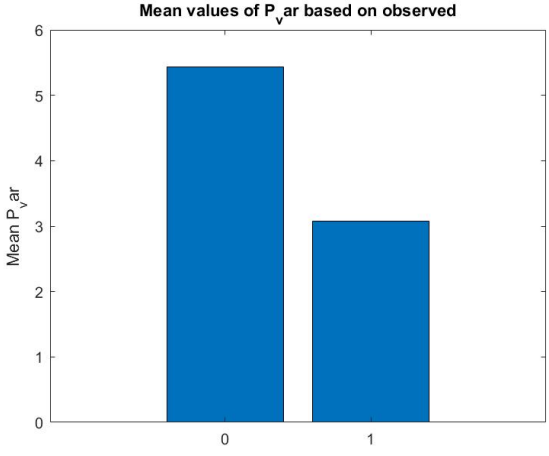

Figure 4: Mean of P_var

Task 1.5

Initially as there were 19 attributes, after completed this task, the date attribute was removed and YY, MM and DD attributes was added to the table and the code is mentioned in the live script.

Task 1.6

After the sub task 6, the Y1 which is observed attribute was separated as a column vector and U_mu which relative humidity was also record in a column vector and all the other variables was formed into a matrix X where there are 19 columns which represents attributes, and the code is mentioned in the live script.

Task 2

Task 2.1

The principal component analysis was adopted and based on the output explained which mentioned below, just the first two principle components explain the 94% of the total variance. Hence the optimal number of PC is 2.

```
explained = 19×1
   90.8774
    3.2430
    1.6603
    1.2473
    0.8146
    0.7873
    0.6092
    0.3443
    0.1388
    0.1090
      ⋮
```

Figure 5: Variance explained by each PC

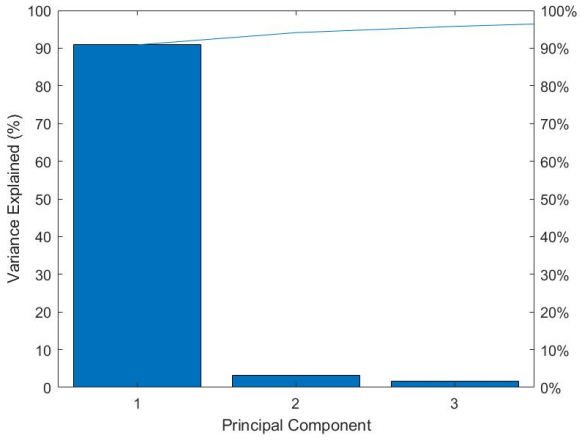

Figure 6: Pareto graph

Task 2.2

K means clustering was adapted for the PCA data, where the number of clusters are 2, 3, 4 and 5.

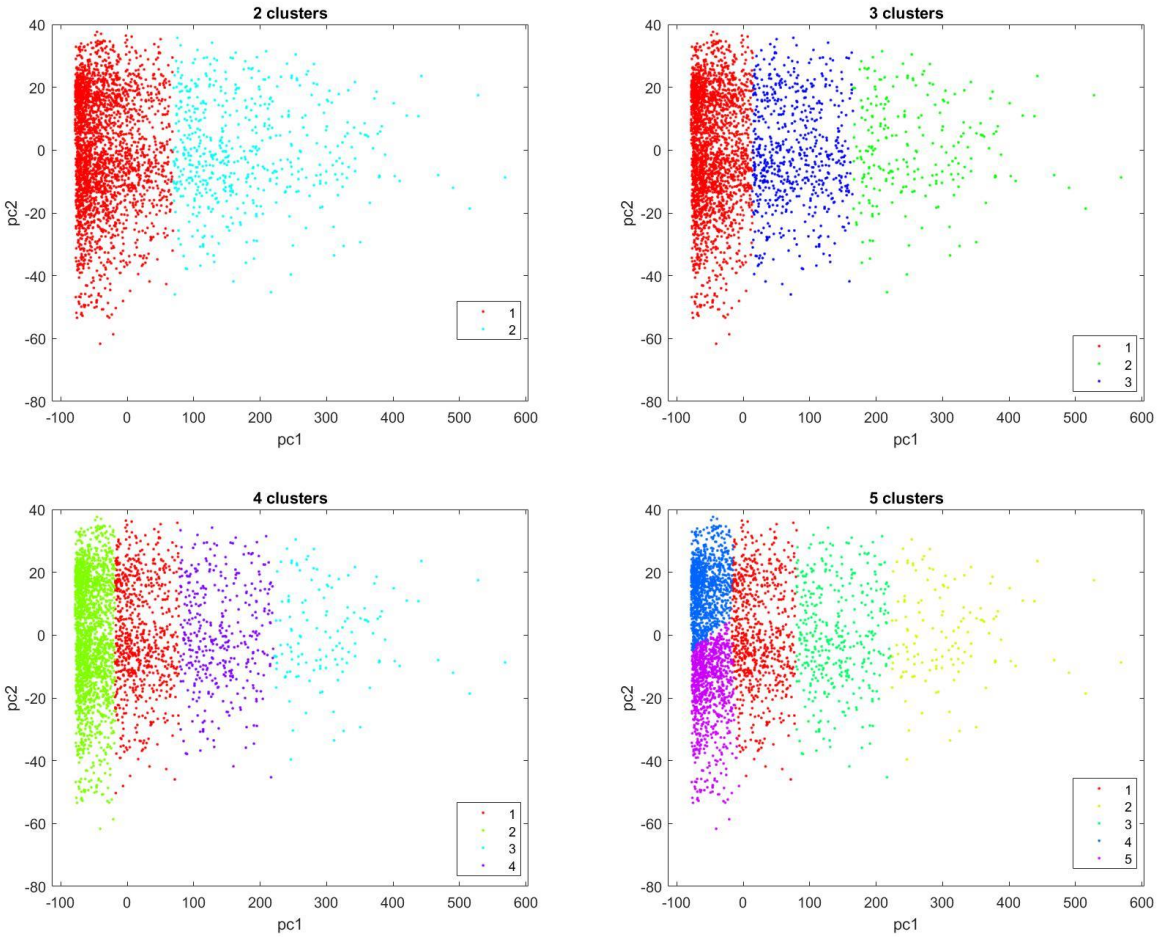

Figure 7: Scatter plots for clusters

In order to decide the optimal number of clusters, 3 different tests were performed, namely, based on silhouette criteria, CalinskiHarabasz criteria and DaviesBouldin criteria.

```
E =
  SilhouetteEvaluation with properties:

    NumObservations: 3202
         InspectedK: [2 3 4 5]
    CriterionValues: [0.8393 0.7436 0.6235 0.5221]
           OptimalK: 2

E2 =
  CalinskiHarabaszEvaluation with properties:

     NumObservations: 3202
          InspectedK: [2 3 4 5]
     CriterionValues: [6.0404e+03 5.6797e+03 5.4302e+03 4.9049e+03]
            OptimalK: 2
```

```
E3 =
  DaviesBouldinEvaluation with properties:

    NumObservations: 3202
         InspectedK: [2 3 4 5]
    CriterionValues: [0.5790 0.6846 0.8107 0.9148]
           OptimalK: 2
kbest3 = 2
```

Figure 8: Deciding the optimal number of clusters

And based on the all 3 criteria, it is decided that the optimal number of clusters is 2 clusters.

Task 2.3

For the sub task 3, a hierarchical clustering was adopted with the cosine distance and average linkage and when the cutoff is 0.6, it gave 19 clusters as shown in the below figure.

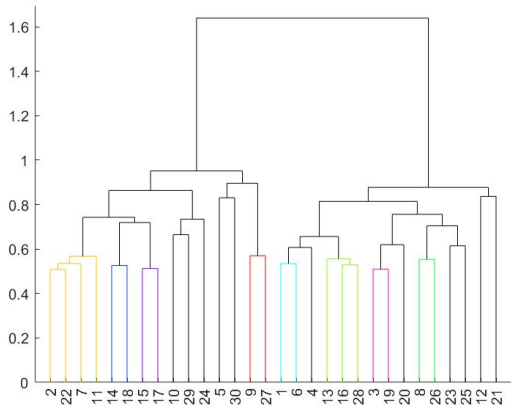

Figure 9: Dendrogram for hierarchical clustering

In order to determine the optimal number of clusters between 2 and 19, silhouette Evaluation was used, and it is shows that the optimal number of cluster is 2 even according to that.

```
Eva =
  SilhouetteEvaluation with properties:

    NumObservations: 3202
         InspectedK: [2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]
    CriterionValues: [0.6573 0.3433 0.1133 0.1085 0.0861 0.0811 0.0810 0.0993 0.1038 0.1003 0.0825 0.0733
           OptimalK: 2
```

Figure 10: Evaluation of Hierarchical clustering

Task 2.4

As instructed after adopting the Gaussian model the probability of falling the 8^{th} observation into cluster number 1, 2 and 3 is as follows.

- Cluster 1 : 0.012
- Cluster 2 : 0.5943
- Cluster 3 : 0.3937

Task 3

Task 3.1

Both predictors matrix X and the variable of interest is divided into training and testing data sets for the ratio of 0.7:0.3 and the code is mentioned in the attached live script.

Task 3.2

By using all the predictors attributes the linear regression model was fitted. And the fitted model explains 98.7% of the variance of the relative humidity.

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	166.81	24.17	6.9014	6.6941e-12
x1	-4.2234	0.066469	-63.539	0
x2	3.2381	1.4387	2.2508	0.024499
x3	-3.2752	1.438	-2.2776	0.022842
x4	-0.25393	0.031883	-7.9643	2.628e-15
x5	0.24538	0.041928	5.8525	5.5615e-09
x6	-0.4069	0.040033	-10.164	9.401e-24
x7	-0.06757	0.0035474	-19.048	4.7317e-75
x8	4.4901	0.019606	229.01	0
x9	0.043513	0.013499	3.2235	0.0012846
x10	0.32729	0.23923	1.3681	0.17141
x11	-0.32692	0.23913	-1.3671	0.17172
x12	-0.024046	0.029606	-0.81222	0.41675
x13	0.05018	0.0056836	8.8289	2.0981e-18
x14	-0.015577	0.0075001	-2.0769	0.037928
x15	0.0022651	0.00035436	6.3921	1.9883e-10
x16	0.036119	0.0086445	4.1782	3.0516e-05
x17	-0.019276	0.011863	-1.625	0.10431
x18	0.0089492	0.0097834	0.91473	0.36043
x19	-0.004206	0.0033983	-1.2377	0.21597

Number of observations: 2242, Error degrees of freedom: 2222
Root Mean Squared Error: 1.4
R-squared: 0.987, Adjusted R-Squared: 0.986
F-statistic vs. constant model: 8.56e+03, p-value = 0

Figure 11: Result of first linear regression model

Task 3.3

After removing the first and the second variable from the X matrix the result of linear regression is as follows.

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	274.46	40.443	6.7865	1.4686e-11
x1	-0.073966	0.0067534	-10.952	3.1648e-27
x2	-0.50481	0.053077	-9.5108	4.7279e-21
x3	-1.5549	0.051858	-29.984	3.5105e-166
x4	-2.2643	0.045768	-49.473	0
x5	-0.11084	0.0058362	-18.992	1.1653e-74
x6	3.9457	0.029572	133.43	0
x7	-0.018418	0.022551	-0.81674	0.41417
x8	-0.15331	0.28733	-0.53358	0.59369
x9	0.17877	0.28743	0.62196	0.53403
x10	-0.1299	0.049575	-2.6203	0.0088452
x11	-0.0015439	0.0094315	-0.16369	0.86999
x12	-0.1018	0.012366	-8.2321	3.098e-16
x13	0.0039241	0.00059247	6.6234	4.3885e-11
x14	-1.1075e-05	0.014465	-0.00076559	0.99939
x15	-0.059164	0.01987	-2.9776	0.0029368
x16	0.052265	0.016366	3.1935	0.0014254
x17	0.0012183	0.0056946	0.21394	0.83061

Number of observations: 2242, Error degrees of freedom: 2224
Root Mean Squared Error: 2.35
R-squared: 0.962, Adjusted R-Squared: 0.962
F-statistic vs. constant model: 3.31e+03, p-value = 0

Figure 12: The result of second linear regression model

After removing the first 2 variables the fitted linear regression's R square value has decreased to 0.962. Therefore, just based on the R squared values, the model one 1 is better.

Task 3.4

As per the two mean squared error values that is mentioned below, since the MSE value for the model 1 is lesser than the model 2, the model 1 that was fitted using all the explanatory attributes is better.

- Model 1 : 2.1855
- Model 2 : 5.8113

Task 3.5

Three different linear regression models was performed using the,

- First 2 PCs
- First 3 PCs
- First 4 PCs

And the results of those are as follows.

```
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate       SE          tStat       pValue
                   ________    _________    _______    __________

    (Intercept)      80.696      0.24025     335.88              0
    x1             0.025255    0.0025415     9.9373    8.4327e-23
    x2             -0.16863      0.01363    -12.373    4.6492e-34


Number of observations: 2242, Error degrees of freedom: 2239
Root Mean Squared Error: 11.4
R-squared: 0.101,  Adjusted R-Squared: 0.1
F-statistic vs. constant model: 126, p-value = 1.51e-52
```

Figure 13: PCR for the first 2 PCs

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	80.696	0.21324	378.42	0
x1	0.025255	0.0022558	11.196	2.3869e-28
x2	-0.16863	0.012098	-13.939	2.0528e-42
x3	-0.40879	0.016633	-24.577	2.799e-118

```
Number of observations: 2242, Error degrees of freedom: 2238
Root Mean Squared Error: 10.1
R-squared: 0.292,  Adjusted R-Squared: 0.291
F-statistic vs. constant model: 308, p-value = 2.49e-167
```

Figure 14: PCR for the first 3 PCs

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	80.696	0.17279	467.02	0
x1	0.025255	0.0018279	13.817	9.9215e-42
x2	-0.16863	0.0098027	-17.203	2.1834e-62
x3	-0.40879	0.013478	-30.331	1.4615e-169
x4	-0.54704	0.015982	-34.228	7.5516e-207

```
Number of observations: 2242, Error degrees of freedom: 2237
Root Mean Squared Error: 8.18
R-squared: 0.535,  Adjusted R-Squared: 0.535
F-statistic vs. constant model: 645, p-value = 0
```

Figure 15: PCR for the first 4 PCs

Just based on the R squared values, the PCR for the first 4 principal components has the highest among all three model. Based on the MSE values, we need to determine the model that has the lowest MSE and as per the below MSE values for the 3 models, the Principal Component Regression with the first 4 PCs is the best model among all 3.

- MSE for model 1 : 122.7393
- MSE for model 2 : 94.5255
- MSE for model 3 : 64.8524

Task 4

Task 4.1

As instructed the data was splitted into training and testing data and the code is mentioned in the live script.

Task 4.2

```
Generalized linear regression model:
    Observed ~ [Linear formula with 20 terms in 19 predictors]
    Distribution = Binomial

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	-254.63	48.44	-5.2567	1.4667e-07
T.T_mu	0.31237	0.13765	2.2694	0.023247
T.Po_mu	-3.7258	2.7732	-1.3435	0.17911
T.P_mu	3.8318	2.7723	1.3822	0.16691
T.Ff_mu	-0.35175	0.063208	-5.5649	2.6225e-08
T.Tn_mu	0.45726	0.087644	5.2172	1.8165e-07
T.Tx_mu	-0.3747	0.080105	-4.6776	2.9024e-06
T.W_mu	0.060491	0.0078224	7.7331	1.05e-14
T.Td_mu	-0.45605	0.054568	-8.3574	6.4122e-17
T.T_var	0.21377	0.033486	6.384	1.7247e-10
T.Po_var	-0.802	0.51754	-1.5496	0.12123
T.P_var	0.79454	0.5171	1.5365	0.12441
T.Ff_var	-0.054751	0.05845	-0.93671	0.34891
T.Tn_var	0.037347	0.011849	3.1519	0.0016223
T.Tx_var	-0.021887	0.015358	-1.4252	0.15411
T.W_var	-0.0051224	0.00070787	-7.2364	4.6092e-13
T.Td_var	-0.1178	0.019891	-5.9221	3.1784e-09
T.YY	0.086394	0.023555	3.6677	0.00024477
T.MM	0.014147	0.01828	0.77389	0.439
T.DD	-0.0031394	0.0066219	-0.47409	0.63543

Figure 16: Logistic Regression model

Task 4.3

The code is mentioned in the live script.

Task 4.4

The AUC of the models was calculated, and the results is as follows.

Model	AUC
Logistic Regression	0.8603
Classification tree (first 3 PCs)	0.6685

Based on the AUC calculations, the logistic regression model is better than the classification model that was performed based on the first 3 principal components.

Task 4.5

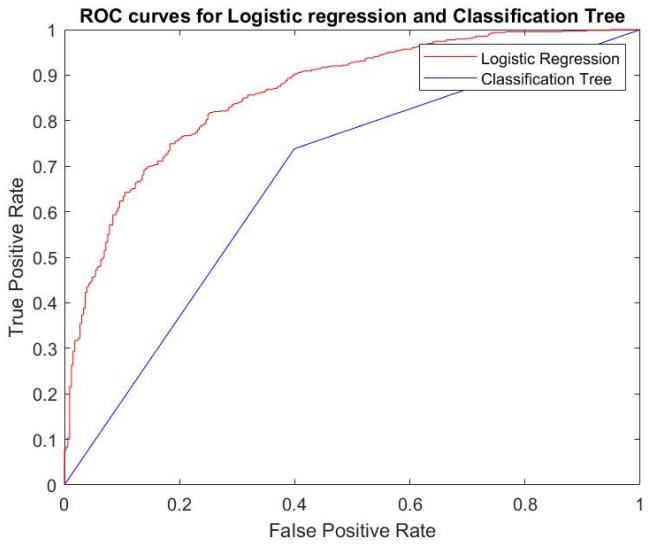

Figure 17: ROC for two models

Task 5

Task 5.1

The data X and the Y1 was splitted as Kfold partition where the number of k-folds is 4 and 3 models were fitted.

Task 5.2

The accuracy and the AUC for the fitted models are as follows.

Model	Accuracy	AUC
Support Vector Machine	0.7823	0.8514
Classification tree (default splits)	0.7337	0.8067
Classification tree (max splits 5)	0.7475	0.7674

As per the accuracy and the AUC values, as we are looking forward to having a model which has the highest value, the model 1 which is SVM would be the better model among the performed models. Hence the final model would be the SVM.

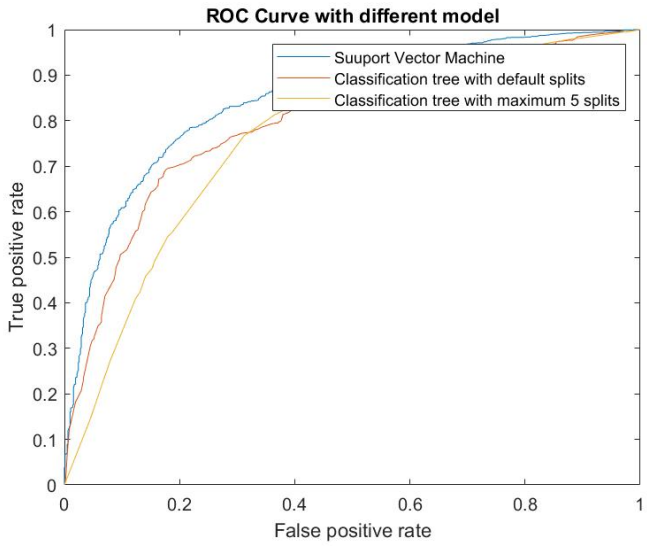

Figure 18: ROC for the performed models

Task 5.3

As the final model since the SVM was selected, the model was assessed based on first training data and then using the testing data.

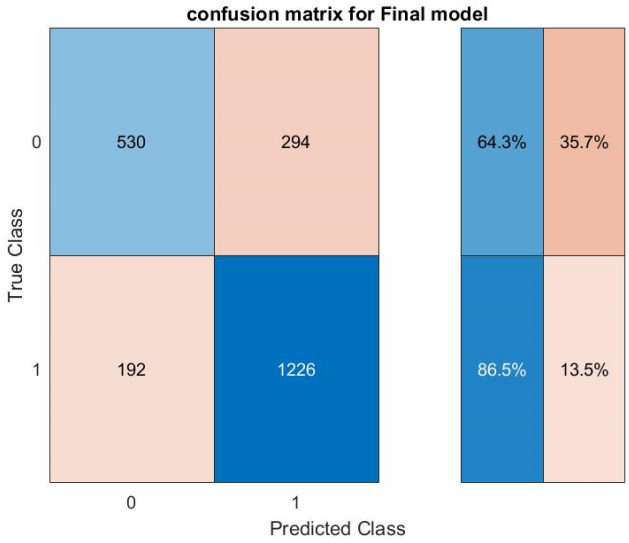

Figure 19: The confusion matrix for the training data

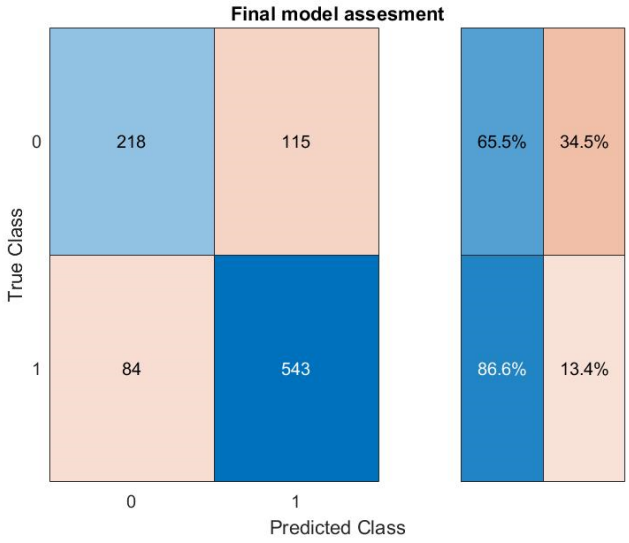

Figure 20: Confusion matrix for testing data

As per the confusion matrix for the training and testing data, the not dry attribute has been correctly classified 65% and the dry attribute has been correctly classified 86%.

Task 5.4

After splitting the data based on k folds, in order to choose the best model each AUC value of the models were considered and the model that has the highest AUC was considered plus accuracy was calculated for each model and the model that has the highest value is considered the best model. According to the given accuracy and AUC values, SVM model has the highest hence that was selected as the best model and for the sub task 3 predictions of SVM was chosen.

Two confusion matrices were performed for training and testing and the final model assessment was decided based on the testing data.