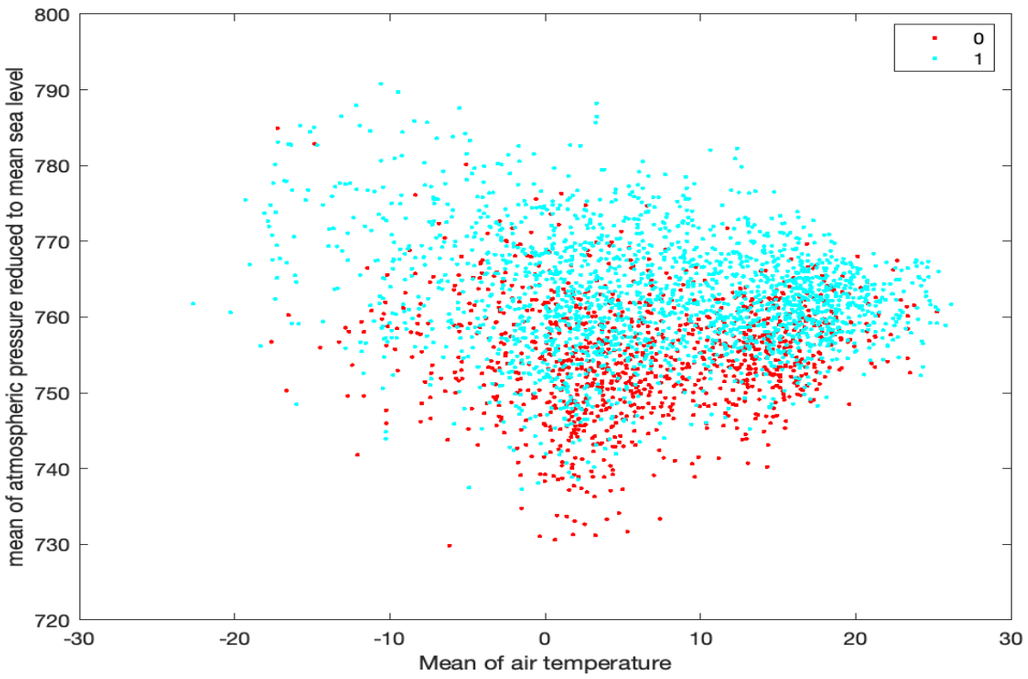
Vikas S. Kushwaha
Department of Engineering Science, LUT University
A220A0010: Analytics of Business

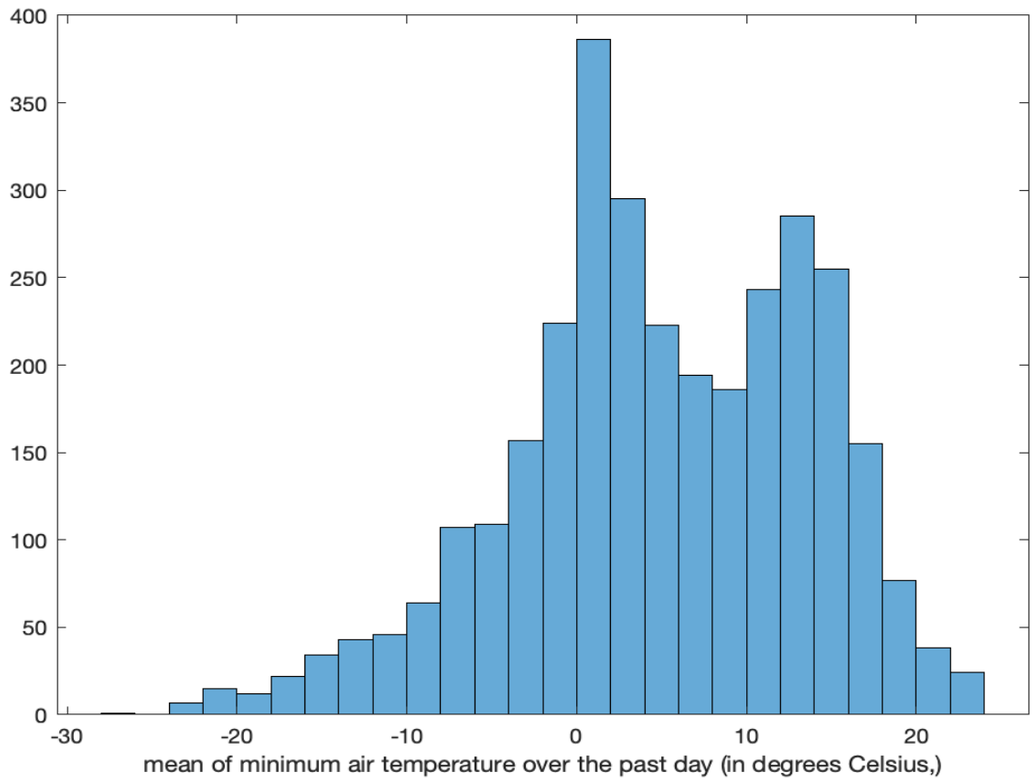
Analytics of Business

TASK 1

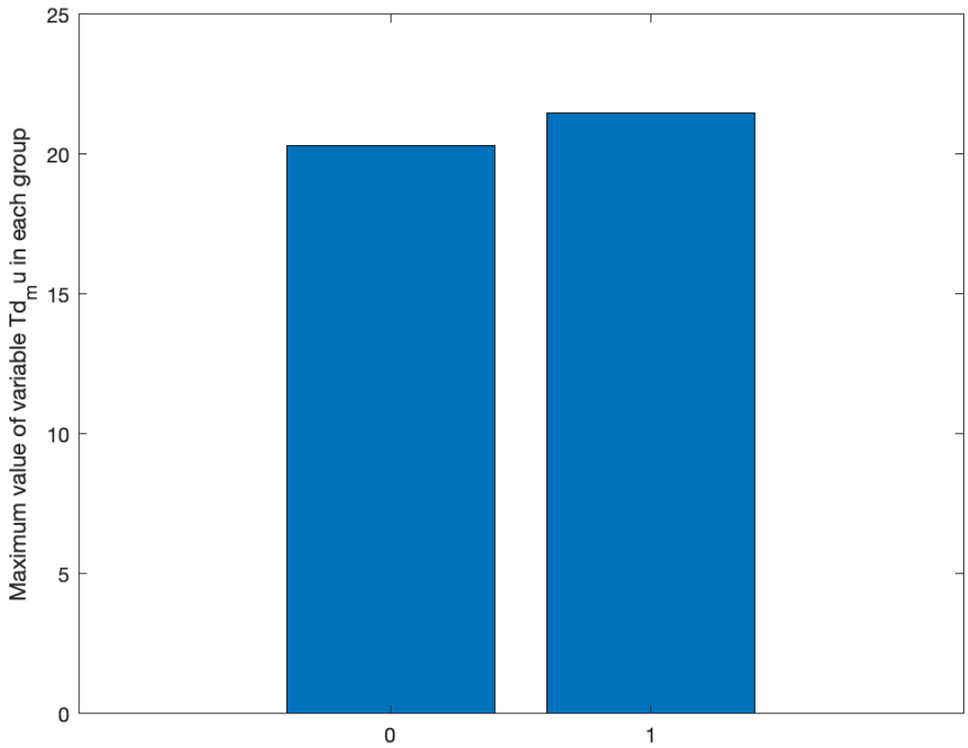
1. 2-D scatter plot



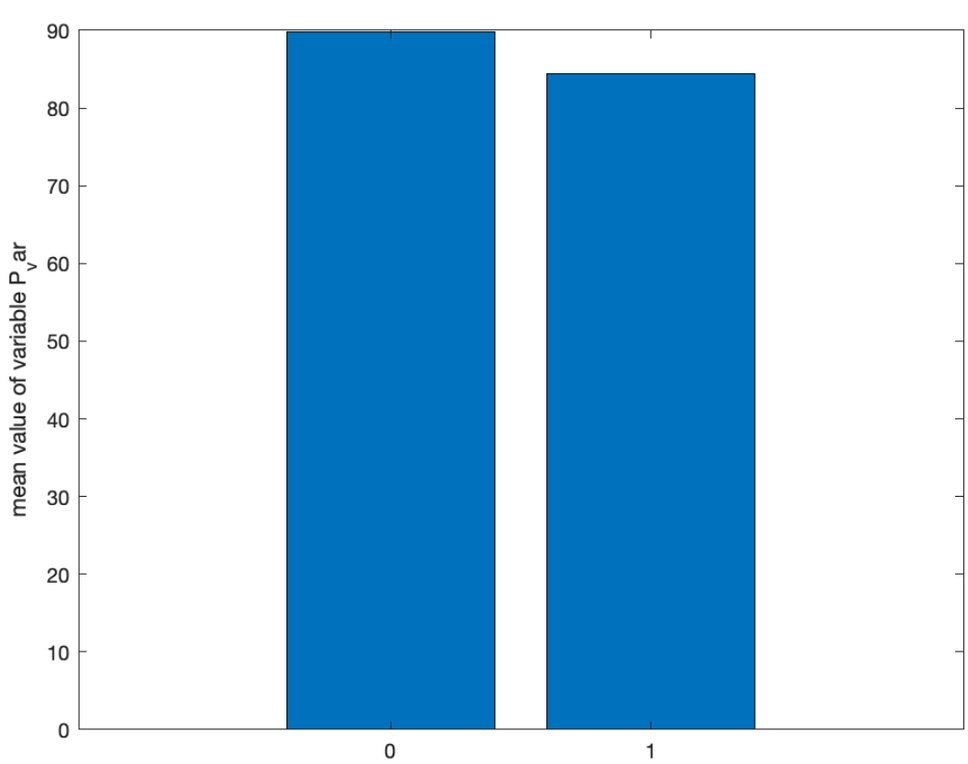
2. Histogram of a mean of minimum air temperature over the past day



3. Two groups 0:not dry 1:dry
   The maximum value of the variable Td_mu in two groups group with not dry is 20.29
   The group with dry is 21.4625

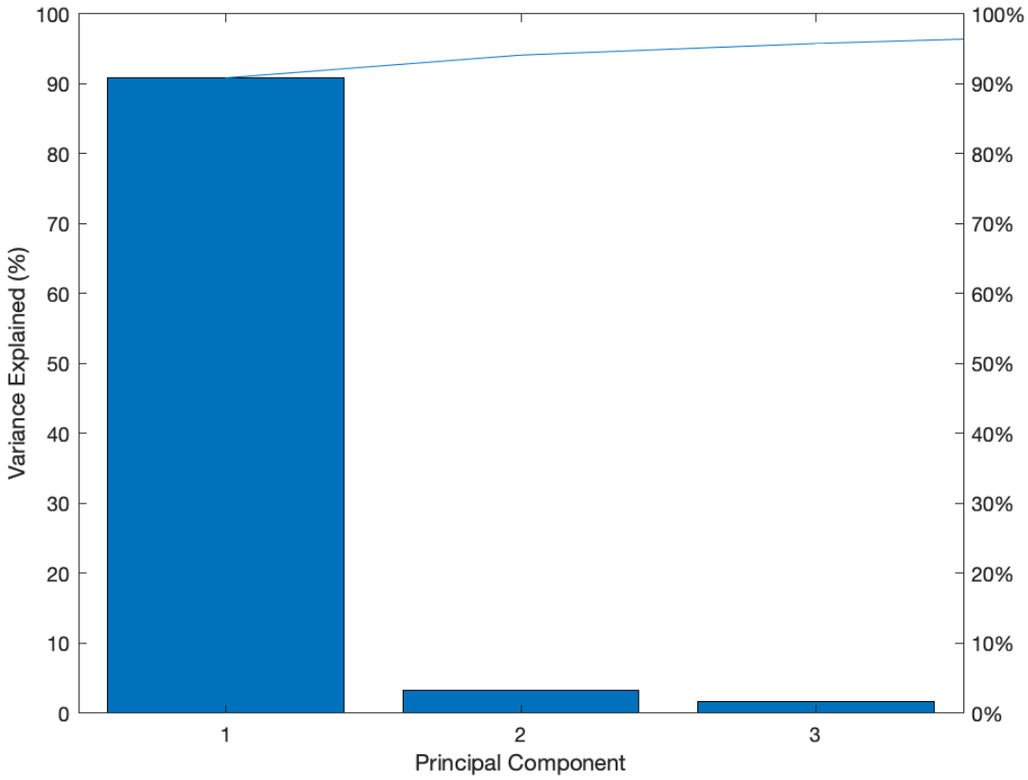


4. The mean value of variable P_var in each of the considered two groups, Mean of variable P_var in Group with not dry is 89.80. The mean of a group with the dry is 84.37.

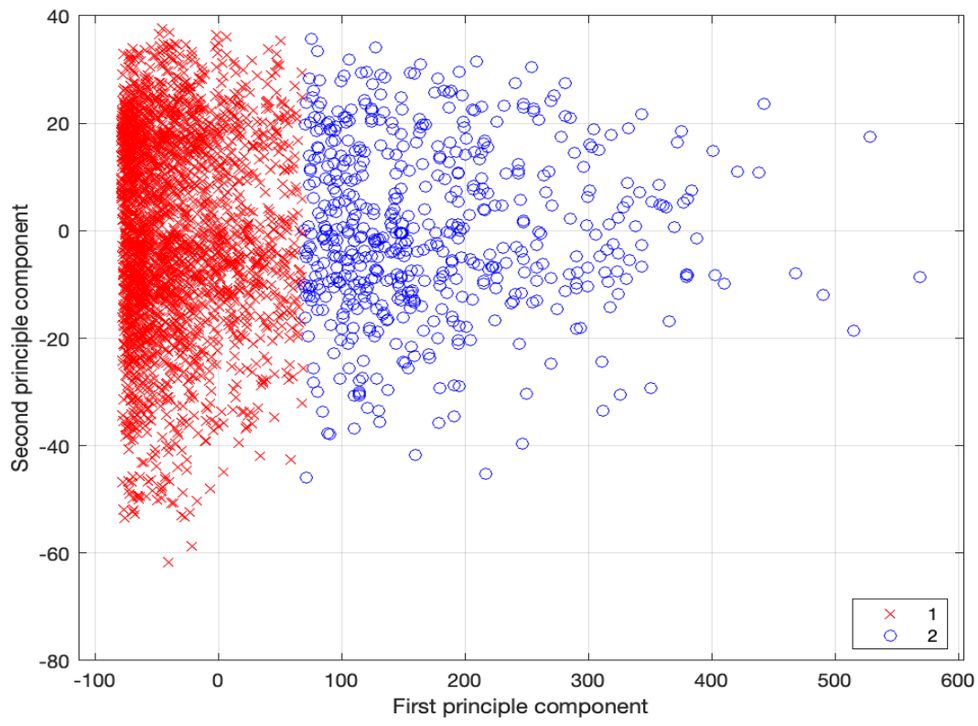


TASK 2

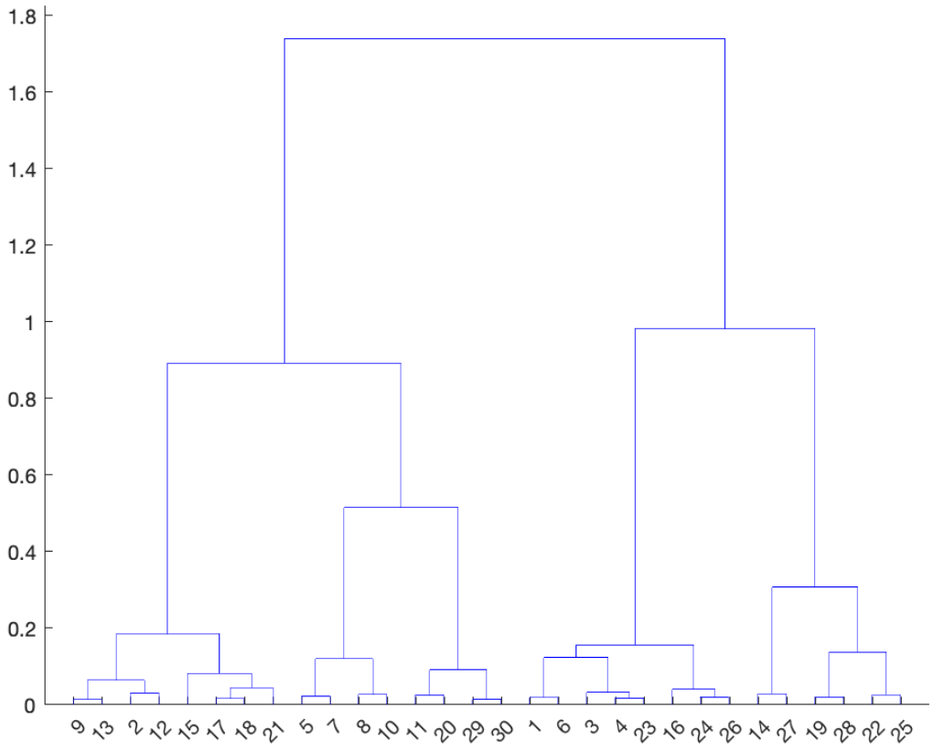
1. Cumulative explained variance plot
   We have selected 2 principle components which explained the 94% of the variance



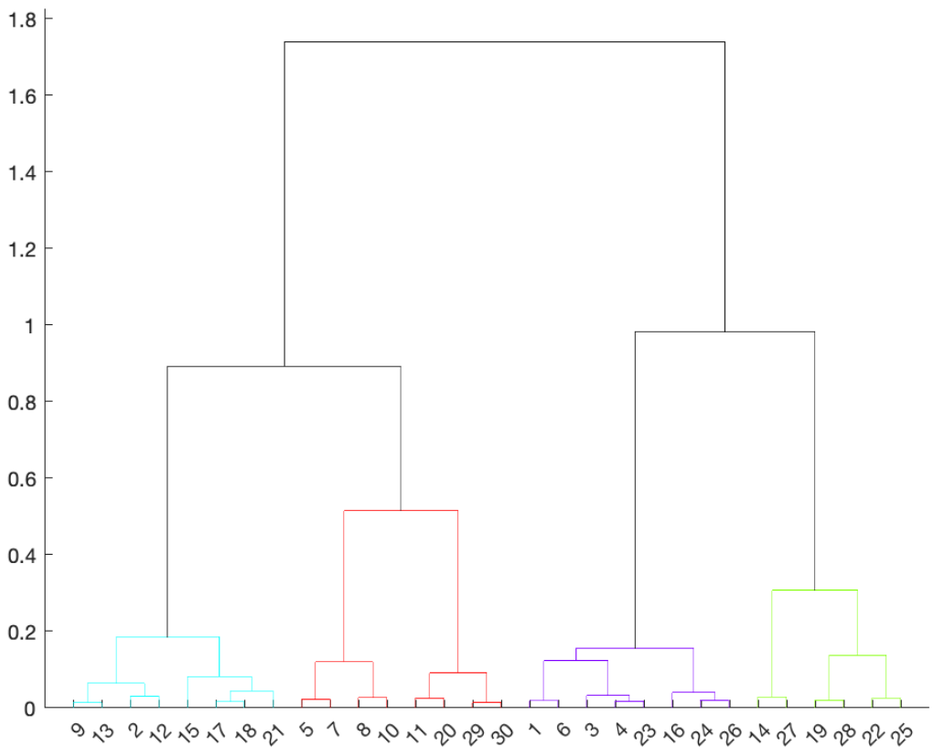
2. With the selected principle component, I generated 2, 3, 4, and 5 clusters. We used the silhouette method to evaluate and found that the 2 clusters had the highest silhouette value, thus I concluded that it was the best option. To visually analyse the indicated groupings, we created a scatter plot (2D).



3. Adopting Hierarchical Clustering, the Cosine distance, and linkage "average to the selected principle component in point 1 and visualizing the dendrogram.
   We have got 4 clusters using this method



Visualize the coloured clusters in the dendrogram



Evaluating the cluster

Eva =

Silhouette Evaluation with properties:

Num Observations: 3202
InspectedK: [2 3 4]
CriterionValues: [0.6771 0.2968 0.3292]
OptimalK: 2

From the silhouette value stored in variable Eva for cluster 2, it is visible that it is highest compared to cluster 3 and cluster 4 so we can evaluate cluster 2 as an optimal cluster. cluster 4 is good as compared to 3 but not cluster 2

4. GMM

Cluster	1	2	3
Probabilities	0.3936	0.01120	0.5944

From observation, we can see that observation 8 has the highest probability of lying in cluster 3.

TASK 3

1. We used a 70:30 ratio to divide data from the predictor matrix X and the target variable Y for training and testing, and saved training data in Xtrain and Ytrain and testing data in Xtest and Ytest.
2. Linear regression model using predictor matrix Xtrain and target variable Y2train

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 +
x13 + x14 + x15 + x16 + x17 + x18 + x19
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	138.85	24.51	5.665	1.6609e-08
x1	-4.3564	0.069671	-62.529	0
x2	1.6884	1.3335	1.2661	0.2056
x3	-1.7257	1.3327	-1.2948	0.19551
x4	-0.24274	0.032941	-7.3687	2.4163e-13
x5	0.32671	0.043755	7.4669	1.1727e-13
x6	-0.33848	0.041221	-8.2114	3.6639e-16
x7	-0.067287	0.0035606	-18.898	5.5316e-74
x8	4.478	0.020354	220.01	0
x9	0.059237	0.013838	4.2808	1.9416e-05
x10	0.25987	0.24179	1.0748	0.28258
x11	-0.25703	0.2417	-1.0634	0.2877
x12	-0.043101	0.02904	-1.4842	0.13789
x13	0.051311	0.0058756	8.7329	4.7865e-18
x14	-0.020019	0.0077471	-2.584	0.0098289
x15	0.0023799	0.00036281	6.5597	6.6851e-11
x16	0.046228	0.0090049	5.1336	3.0892e-07
x17	-0.0056878	0.01203	-0.47279	0.63641
x18	0.017068	0.0097602	1.7487	0.080476
x19	-0.0052721	0.0034398	-1.5326	0.1255

```
Number of observations: 2242, Error degrees of freedom: 2222
Root Mean Squared Error: 1.42
R-squared: 0.986,  Adjusted R-Squared: 0.986
F-statistic vs. constant model: 8.46e+03, p-value = 0
```

3. Removing the first and second variable from training data and making a linear regression model.

```
mdl2 =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
x12 + x13 + x14 + x15 + x16 + x17

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	291.34	40.495	7.1945	8.5324e-13
x1	-0.071947	0.006763	-10.638	8.1541e-26
x2	-0.5982	0.053859	-11.107	6.2175e-28
x3	-1.5719	0.052313	-30.049	8.7548e-167
x4	-2.1998	0.047325	-46.482	0
x5	-0.11138	0.0057931	-19.226	2.4883e-76
x6	3.8935	0.030021	129.69	0
x7	-0.049222	0.022751	-2.1635	0.030608
x8	-0.15678	0.23941	-0.65487	0.51262
x9	0.18612	0.2395	0.77714	0.43716
x10	-0.039384	0.048199	-0.81712	0.41395
x11	-0.0066548	0.0096229	-0.69157	0.48928
x12	-0.096309	0.0127	-7.5833	4.9154e-14
x13	0.0041249	0.00060043	6.8699	8.3071e-12
x14	-0.021023	0.014844	-1.4162	0.15684
x15	-0.068269	0.019907	-3.4295	0.00061579
x16	0.056215	0.016161	3.4784	0.00051416
x17	0.00047454	0.0057049	0.083182	0.93371

```
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: 3.35e+03, p-value = 0
```

4. The MSE for models 1 and 2 (after removing the variable) is 1.900 and 5.3171 respectively. In comparison to model 2, MSE for model 1 is low. On the basis of MSE, we may conclude that model 1 is better performing than model 2.

5. We created three models using the different principal component and perform testing with the testing data below is the MSE values for the three model

MODEL	Model with first 2 principal component	Model with first 3 principal component	The model with first 4 principal component
MSE	121.1737	93.6594	68.2676

From the above we can easily conclude that model with the first 4 principal components is better performing than the other 2.

TASK 4

1.We have considered the same holdout data with 70:30 ratio with Y1 variable.

2.Logistic regression model

```
    aglm =
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
x12 + x13 + x14 + x15 + x16 + x17 + x18 + x19
    Distribution = Binomial

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	-240.44	48.163	-4.9923	5.9676e-07
x1	0.24141	0.13173	1.8326	0.066858
x2	-4.2362	2.5263	-1.6769	0.093565
x3	4.3313	2.5253	1.7151	0.086322
x4	-0.25491	0.06256	-4.0747	4.607e-05
x5	0.36909	0.086367	4.2735	1.9243e-05
x6	-0.27765	0.078088	-3.5557	0.00037703
x7	0.062217	0.0075765	8.2119	2.1773e-16
x8	-0.38771	0.052531	-7.3806	1.5759e-13
x9	0.22031	0.032877	6.7011	2.0682e-11
x10	-0.88531	0.49684	-1.7819	0.074772
x11	0.87692	0.49639	1.7666	0.077296
x12	-0.11348	0.056792	-1.9981	0.045707
x13	0.041964	0.011585	3.6223	0.00029203
x14	-0.023812	0.015593	-1.5271	0.12674
x15	-0.0048175	0.00067731	-7.1127	1.138e-12
x16	-0.1288	0.019927	-6.4635	1.0228e-10
x17	0.083123	0.023429	3.5478	0.00038849
x18	0.0083137	0.017674	0.47039	0.63808
x19	-0.0049734	0.0065256	-0.76213	0.44598

```
2242 observations, 2222 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 1.05e+03, p-value = 4.54e-210
```

Based on AUC we can conclude about performance of model

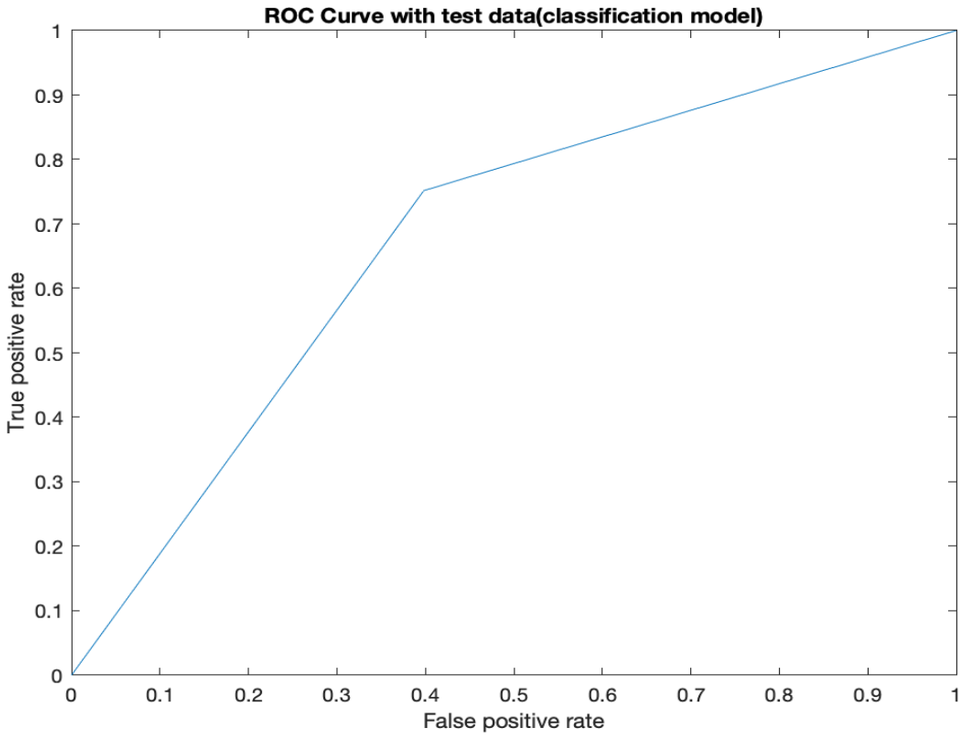
AUC of logistic regression model

```
auctest2 = 0.8799
```

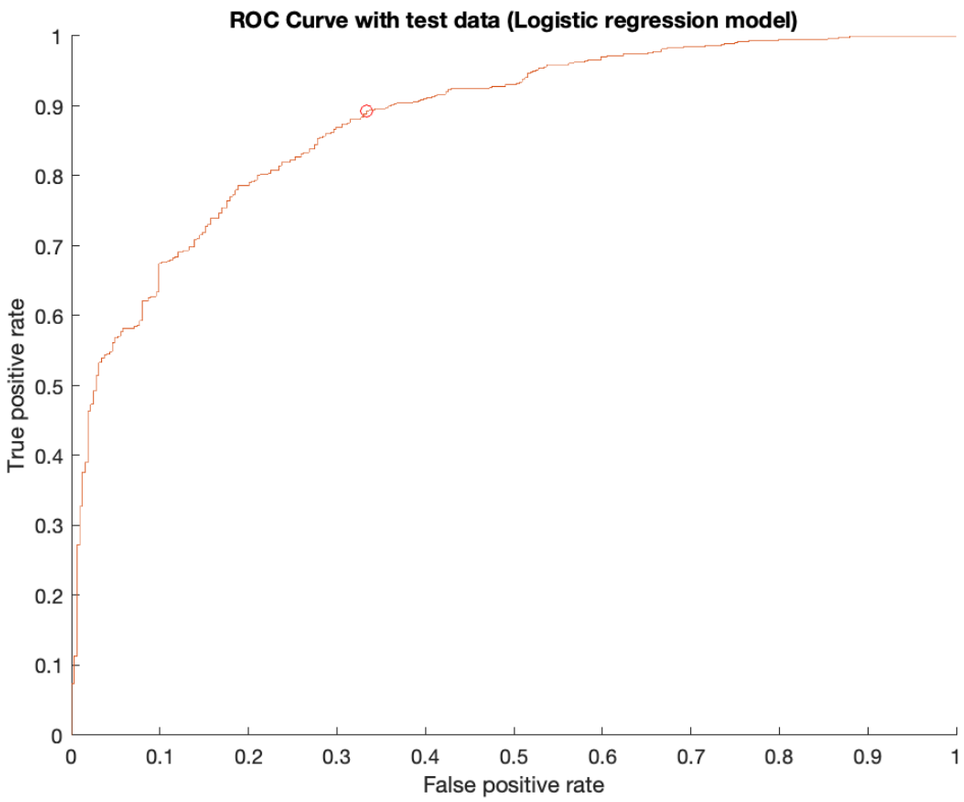
AUC of classification model

```
auctest = 0.6767
```

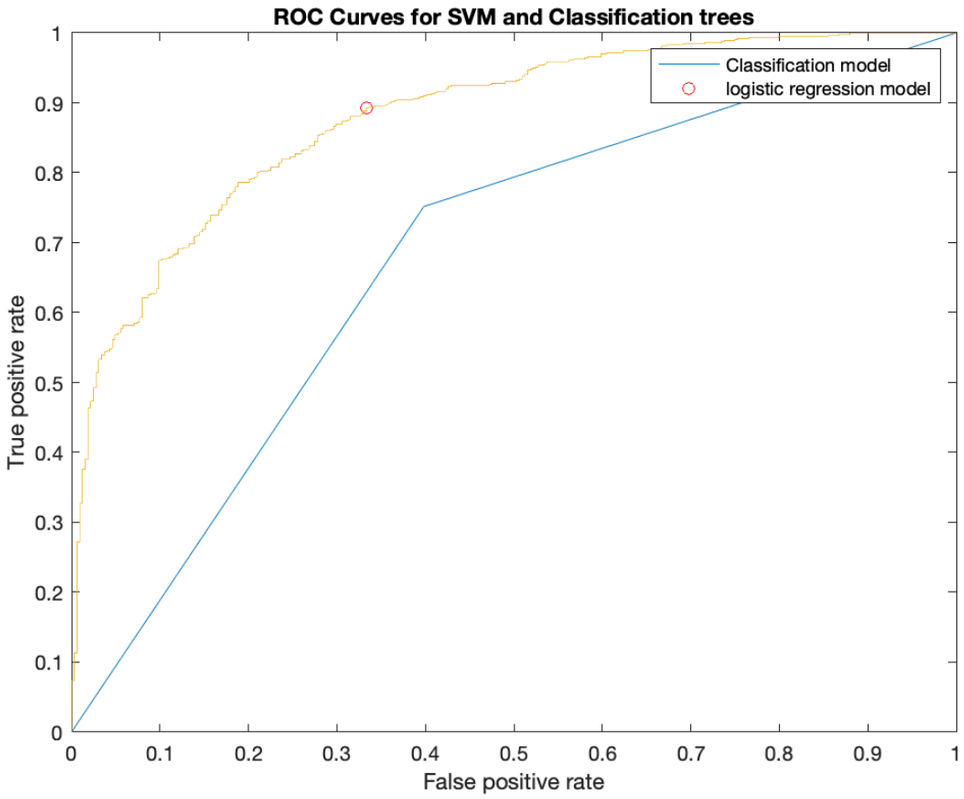
3.ROC curve with test data in classification model



4. ROC curve with test data in logistic model



5. Visuallizing two ROC curve together



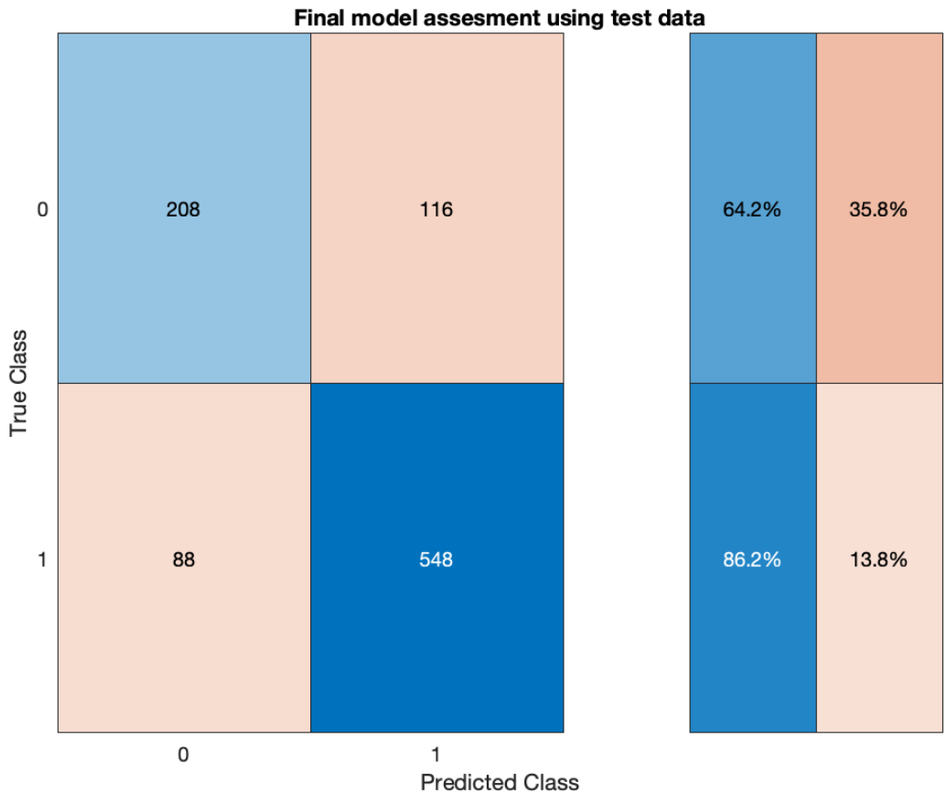
TASK 5

1. We used the split data in the ratio of 70:30 for training and testing
We made three model SVM model, classification tree with min 15 present at leaf node, classification tree(maximum number of split 5)
Below is the AUC and accuracy of all three model

models	SVM model	Classification tree (leaf node with min 15)	Classification tree (max split 5)
AUC	0.8407	0.8086	0.7451
Accuracy	0.7788	0.7511	0.7324

2.We found out the SVM is the best model with gaussian kernel function and auto as a kernel scale based on AUC and accuracy.

3.Confusion chart




```
accuracy_final_model = 0.7875
```

4.
Process

As we have already divided the data using holdout partition in previous task with 70:30 ratio, We get Xtrain,Xtest,Y1train,Y1test. Using Training dataset, train different models by mentioned criteria (SVM and classification tree) and considering k fold=4. We trained and cross validated the model with the 4 kfold, and then performed AUC and accuracy tests of all three models, and found that the SVM model with gaussian kernel function outperformed the other two models in both aspects .As a result, we chose the SVM model as our final model.

For making the model and cross validation in point 1, We used the data Xtrain Y1train we get after holdout partition in 70:30 ratio, even for cross validation the three models.(k=4)
For assess of performance of the final model we have chosen (SVM) , we used the Xtest data to predict and later we made confusion chart with the help of Xtest data prediction on target variable and Y1test data.