Name: **Osinowo Kolawole Michael** **Reg No:** **170153326**

MACHINE LEARNING REPORT

Predicting Customers Airline Satisfaction

Table of Contents

Name: Osinowo Kolawole Michael **Reg No: 170153326**

Section 1

1.1 Brief Introduction and Domain Analysis

The aim of this report is to employ machine learning methods to predict customer satisfaction as accurately as possible. The following paragraphs contain a summary of domain analysis done for this project.

The factors that determine customers satisfaction vary depending on conditions such as; ethnic background, expectation and travelling purpose. To narrow down these factors, a study was carried out in 2019 by Hyun-Jeong et al. in [1] to understand customers experience and satisfaction through online reviews from airline passengers. This study, through statistical and NLP techniques, found out the following factors are what determine customer satisfaction:

- Value for Money
- Staff experience
- Food & Beverages
- Seat Comfort
- Ground Service
- Entertainment

The factors stated above are listed in order of their importance, with ‘Value for Money’ being the most significant factor. Putting these findings into consideration, this report will carry out analysis on the provided dataset for this project and see if the same relationship can be established.

Section 2

2.1 Data Pre-processing and Feature Engineering

Table 1 below provides a comprehensive description of the cleaning, pre-processing and feature engineering steps taken before the implementation of a machine learning model.

Table 1. A concise summary of the problem with the raw dataset and steps taken to eliminate these problems with the reason why these steps were taken.

Problem	Actions	Reason
Missing values in the dataset	Rows with missing values are removed	• Filling in the null values (970 instances) with either the mean or mode will not reasonable as it would encourage misclassification. • There was enough data (2537 instances) without missing values to train on.
String variables from the dataset, i.e.: • Departure Time • Quarter	Encoded the categorical variables, introducing two new features in the dataset; "**Encoded departure time**" and "**Encoded quarter**". • For the "Departure Time", this feature column is encoded according to the different times of the day, as shown in Table 2. • The "Quarter" feature, each quarter of each year is assigned an ordinal label, as shown in Table 3.	The features may contain useful information/ predictive properties for determining overall customers satisfaction. • The "Departure time" feature may be able to describe; at what time of the day a customer is likely to feel satisfied or what time of the day influences their satisfactory rate. • The "Quarter" feature may be able to describe; at what quarter of the year a customer is likely to feel satisfied with the airline service.

		This sort of information can help the airline/airport better understand their services and make it easy to find solutions to the problem.
Unwanted or redundant features or information, i.e.: • "Date Recorded"	This feature is dropped from the dataset.	The "Date Recorded" provides similar information as the "Quarter" feature but on a granular scale which is difficult to categorize. It is a redundant feature, hence, useless.
Skewed Distributions	The dataset is normalised by setting every feature/column to have a mean of 0 and a standard deviation of 1.	This is to prevent features with different distributions to negatively affect the performance of the machine learning model. This step brings the features to a common ground.

Table 2. The categorisation table of "Departure Time" feature.

Part of Day	Category	Time Range	Ordinal Label
Morning	Early morning	5 – 8 am	1
	Late morning	9 am – 12 pm	2
Afternoon	Early Afternoon	12 – 3 pm	3
	Late Afternoon	3 – 5 pm	4
Evening	Evening	5 – 9 pm	5
Night	Night	9 pm – 4 am	0

Table 3. The categorization table of "Quarter" feature.

Quarter	Ordinal Label
'1Q15'	1
'2Q15'	2
'3Q15'	3
'4Q15'	4
'1Q16'	5
'2Q16'	6
'3Q16'	7
'4Q16'	8
'1Q17'	9
'2Q17'	10

2.2. Data Visualisation and Exploration

Considering the dataset has ordinal features, i.e. (it is on a scale of 1-5), scatter plots do not provide much information on the relationship between features and the target variable. Figure 2.1 shows an example of the type of plot observed for all the features even after it is normalised. Another visualisation method that can be useful in this case is the distribution plot of the feature with respect to the target variables.

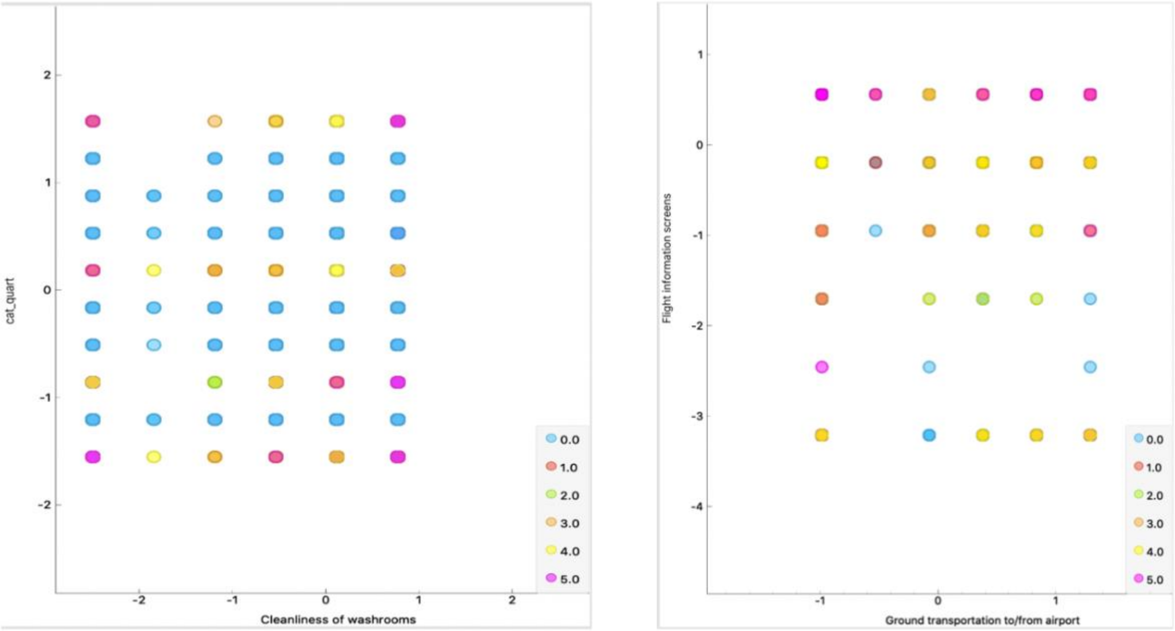


Figure 2.1 Sample scatter plots of features from the datasets (normalised).

Figure 2.2 and 2.3 show the distribution of the new features introduced in the feature engineering process. Based on the Quarterly ratings in Figure 2.2 it can be observed that at following quarters; “1Q15” “4Q16”, “1Q17” and “2Q17”, customers had a higher satisfaction rate hence the increase in 4-5 rating coloured yellow and pink. Customers are also noticed to be less satisfied with the other yearly quarters.

Regarding the time of the day where customers have a significant increase in satisfaction, in Figure 2.3 it can be observed that customers are more likely to be significantly more satisfied with the airline/airport services at departure times between 9:00 pm – 11:59 am and specifically between 9am – 12pm.

These introduced features clearly contain significant predictive/informative features. A decision of using these features for the classification task will be determined after the dimensionality reduction and feature selection step in the next section.

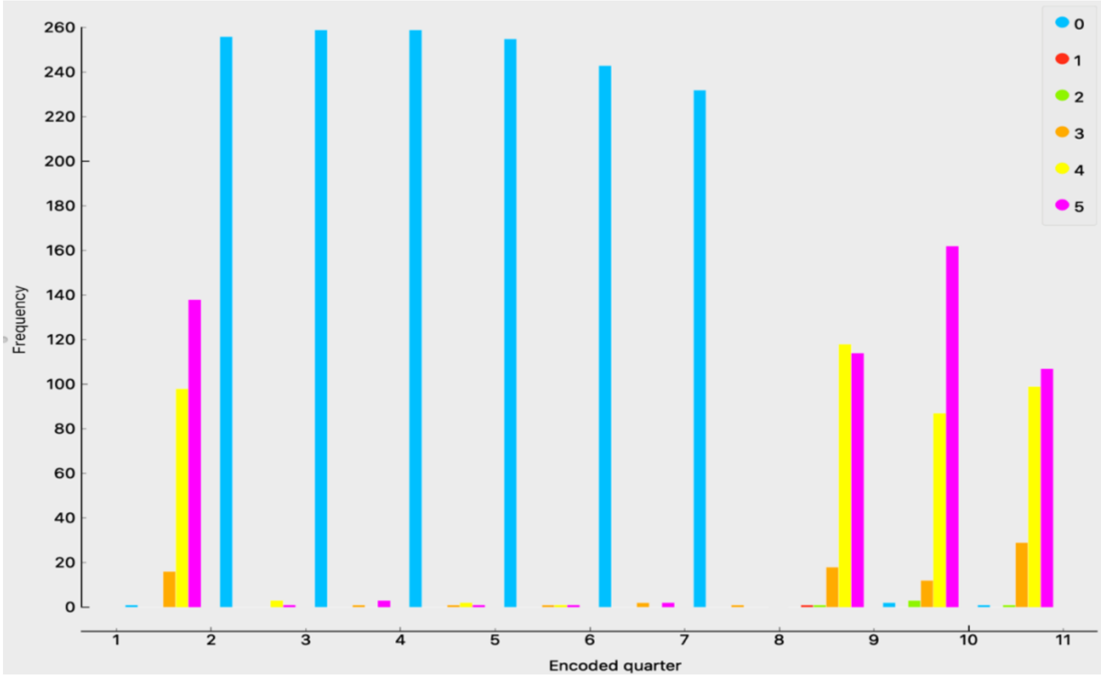


Figure 2.2. Distribution plot of Encoded "quarter" feature.

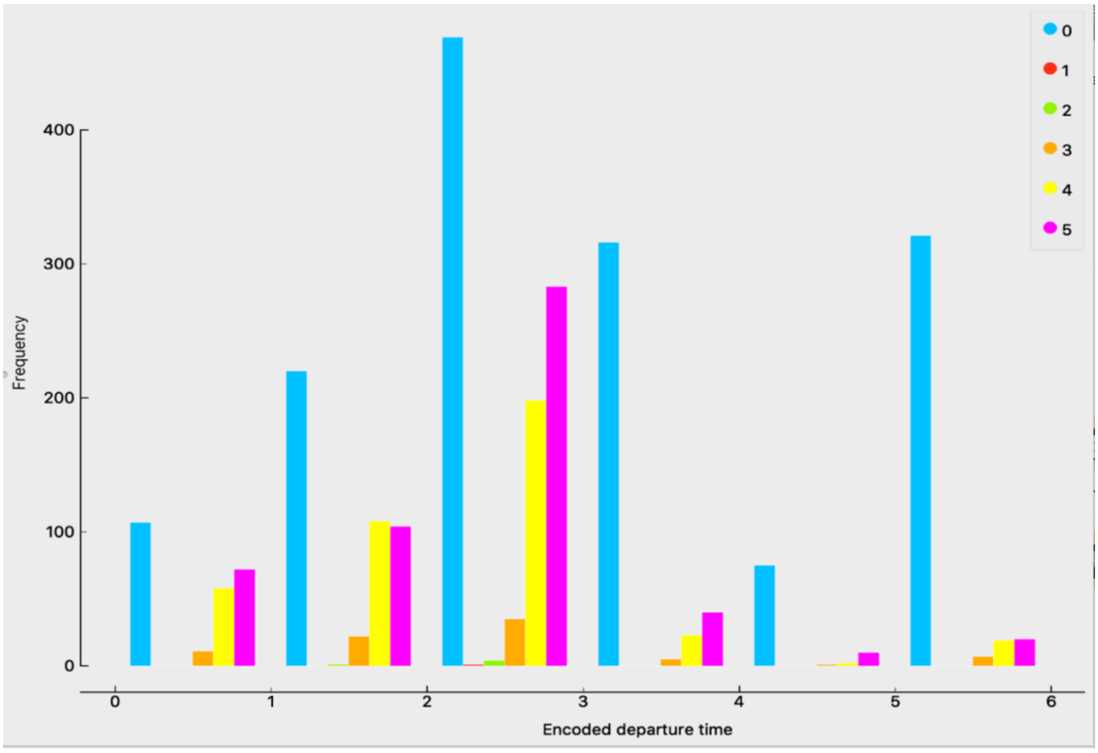
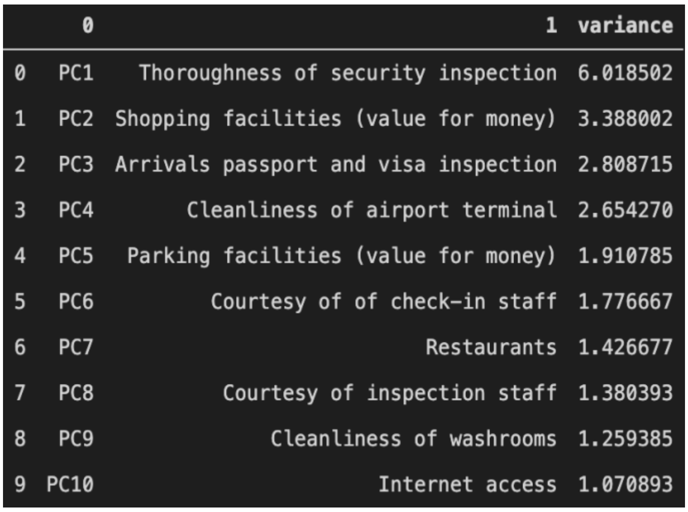


Figure 2.3. Distribution plot of Encoded "departure time" feature.

Section 3

3.1. Dimensionality Reduction and Feature Extraction

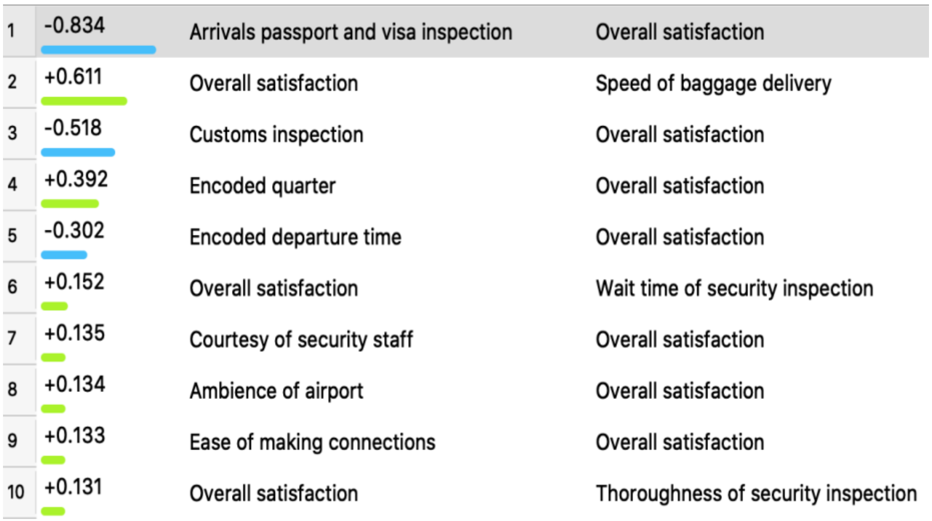
After the cleaning the original dataset, the new dataset has 35 features with 2532 instances. Some of the features in the dataset may not be useful and with more features the more the machine learning model gives in to the “curse of dimensionality”, where more features require more training data/instances. The Principal Component Analysis (PCA) technique is used to extract the ten most important features that can help achieve the classification task as accurately as possible without loss of information or reduced performance, as shown in Figure 3.1. This decomposition process provides features that contain the most information out of 35 other features.



	0	1	variance
0	PC1	Thoroughness of security inspection	6.018502
1	PC2	Shopping facilities (value for money)	3.388002
2	PC3	Arrivals passport and visa inspection	2.808715
3	PC4	Cleanliness of airport terminal	2.654270
4	PC5	Parking facilities (value for money)	1.910785
5	PC6	Courtesy of of check-in staff	1.776667
6	PC7	Restaurants	1.426677
7	PC8	Courtesy of inspection staff	1.380393
8	PC9	Cleanliness of washrooms	1.259385
9	PC10	Internet access	1.070893

Figure 3.1. Top 10 principal components and their corresponding feature names and variance.

Another way of selecting relevant features is based on their correlation with the target variable “Overall satisfaction”. Correlation measures the linear relationship between two continuous variables and the problem with the use of correlation is that it may not be as informative for ordinal /categorical features, regardless the Pearson’s correlation coefficient is used to rank the features, as shown in Figure 3.2.



1	-0.834	Arrivals passport and visa inspection	Overall satisfaction
2	+0.611	Overall satisfaction	Speed of baggage delivery
3	-0.518	Customs inspection	Overall satisfaction
4	+0.392	Encoded quarter	Overall satisfaction
5	-0.302	Encoded departure time	Overall satisfaction
6	+0.152	Overall satisfaction	Wait time of security inspection
7	+0.135	Courtesy of security staff	Overall satisfaction
8	+0.134	Ambience of airport	Overall satisfaction
9	+0.133	Ease of making connections	Overall satisfaction
10	+0.131	Overall satisfaction	Thoroughness of security inspection

Figure 3.2. Top 10 feature ranked based on their correlation coefficient.

From Figure 3.2, it can be observed that the new variables introduced through the feature engineering step i.e. ("**Encoded departure time**" and "**Encoded quarter**") at position 4 and 5 bare a relatively significant relationship with the target variable compared to the other variables from position 5-6. The PCA decomposition has also provided a meaning full projection of the input data set as shown in Figure 3.3, where an obvious separation occurs between the high ratings, i.e. (4-5) and lower ratings i.e. (1-3).

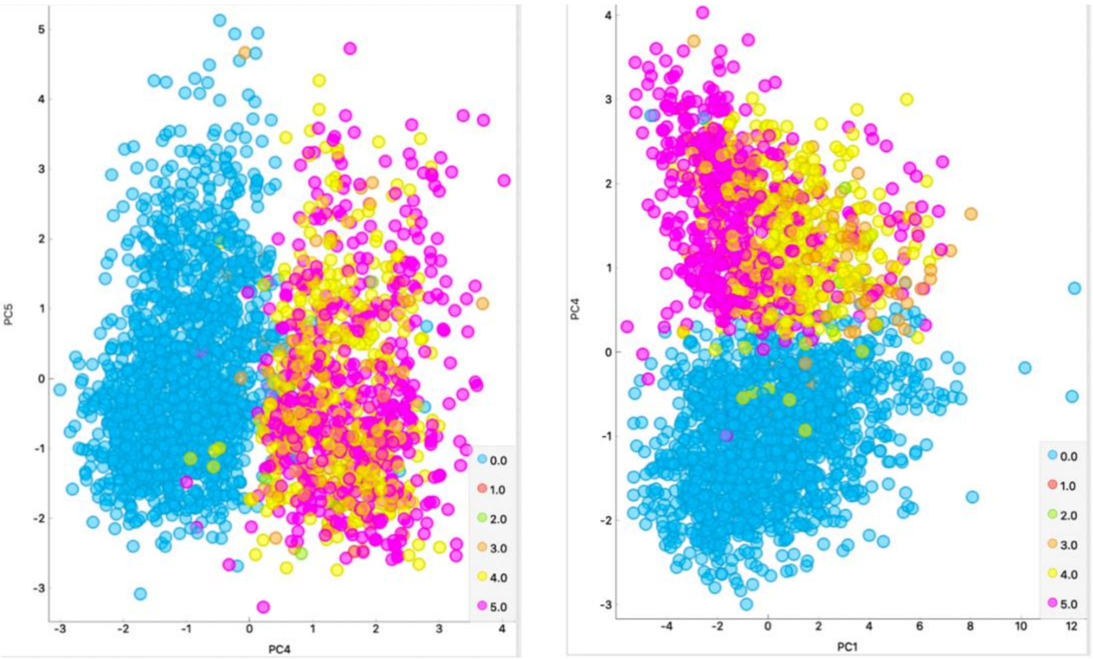


Figure 3.3. Visual representation/projection of some principal component.

3.2. Support Vector Machine Implementation

After the dataset is decomposed into its ten principal components, the resulting dataset is passed into the multiclass SVM model, the pipeline for the SVM model is shown in Figure 3.3. This pipeline is also implemented in the MATLAB code.

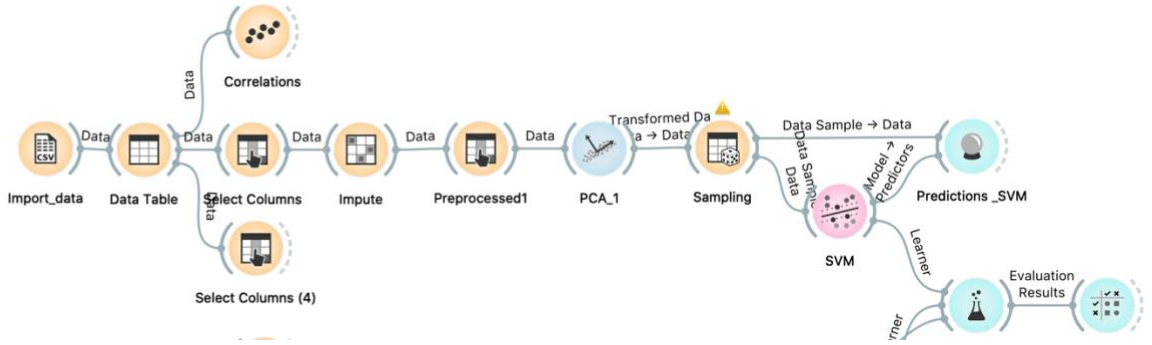


Figure 3.3. SVM pipeline.

3.4. Decision Tree Implementation

The Decision Tree is used as a second model for comparison. The decomposed dataset is used as the input for this model. The following pipeline is also implemented in the MATLAB code.

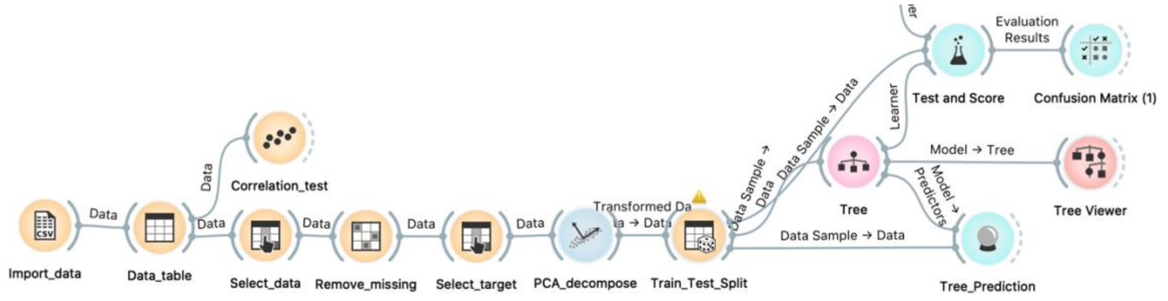


Figure 3.4 Decision Tree pipeline.

3.5 Cross-Validation and Hyperparameter optimisation

The K-fold cross-validation technique is implemented to avoid overfitting and underfitting in both models used. Using 22 folds, the training and test set is split with 80:20 ratio, respectively.

Figure 3.5 shows the learning curve for the Decision tree and SVM model; the SVM model has an ideal curve with relatively ideal bias, and low variance as the training and validation error gradually converges to a plateau. Although the error is not completely eliminated, this shows the SVM model is able to generalize as more data is added to the training set and the performance on both the training and validation set improves simultaneously. The Decision tree model has a low bias as low variance (slightly more than the SVM model), the performance on the training set improves alongside the validation set. The curve for the Decision tree shows it is also able to avoid overfitting, generalize and predict with more accuracy, with lower error on the validation set compared to the SVM model.

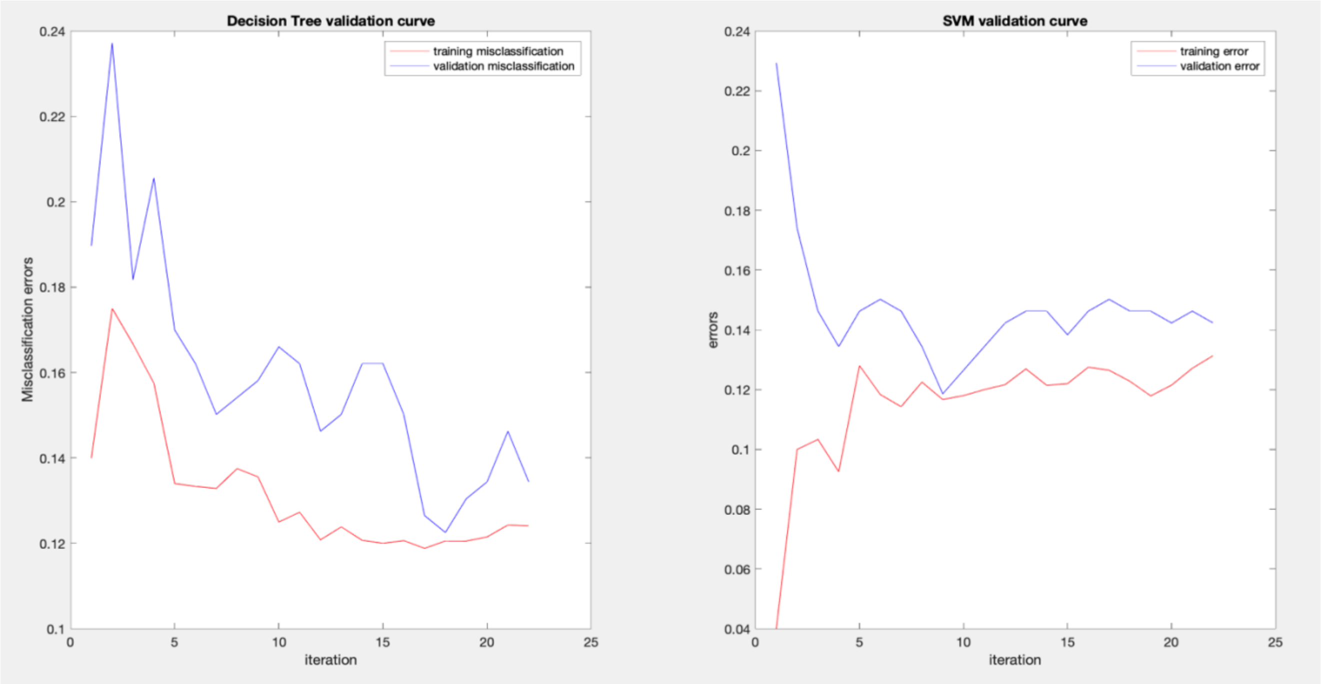


Figure 3.5. The learning curve for the Decision tree and SVM model.

Figure 3.6 on the left, shows the hyperparameter optimisation process for the Decision tree. The performance of the decision tree model is evaluated with respect to change in the "minimum leaf size" of the tree. The depth of the tree is automatically determined by the algorithm. The optimal minimum leaf size is 17, where the least classification error is observed. On the right of Figure 3.5 show the performance of the SVM model for each kernel with respect to change in "margin gap parameter" (C). The "linear" kernel performs better than the "polynomial" and "gaussian/RBF" kernels. The best Margin parameter to use for the linear-SVM model is 100. Although, a margin gap of 100 does not give significantly better performance that a margin of 10 or 50.

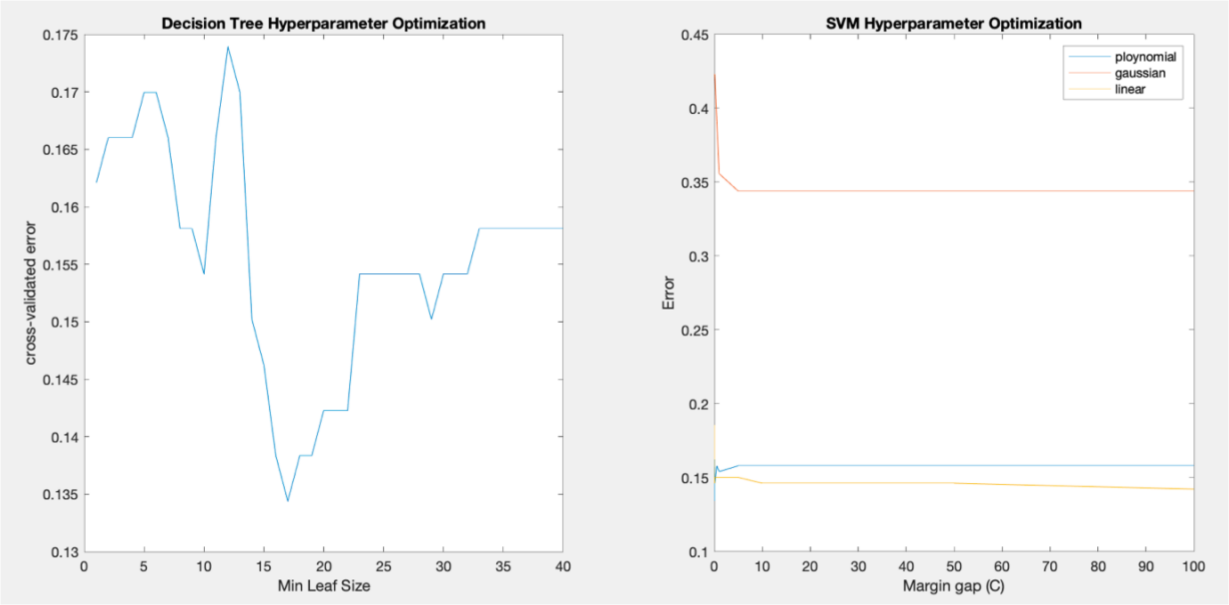
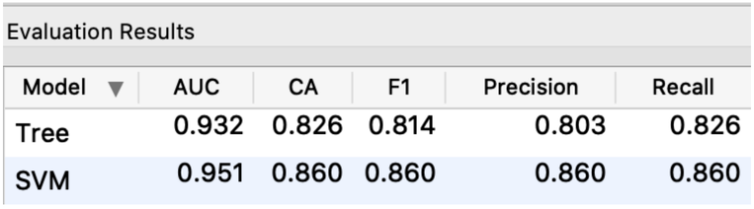


Figure 3.6. The error plot of hyperparameter tuning for Decision tree and three SVM kernels.

3.6 Model Evaluation and Results

Given the aim of the two models used in this project is a classification task, the confusion matrix and other metrics derived from the confusion matrix such as; Area Under the Curve (AUC), Recall, Precision, is used to evaluate the models. Using the Orange3 pipeline, the two models gave an almost similar performance on the validation set with the linear-SVM having the edge over the Decision tree after hyperparameter tuning, as shown in Figure 3.7. The SVM model has a significantly higher precision and recall rate meaning the SVM model is the optimal method for the aim of this project.

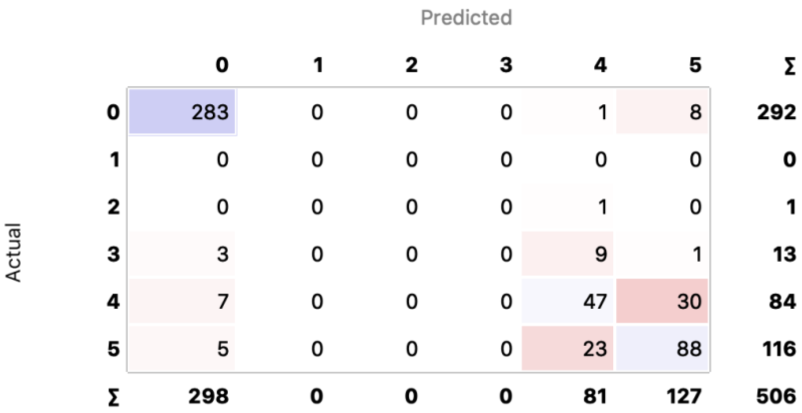
It can be observed in Figure 3.9 and Appendix 4, that the most important Principal Components (PC) common to the MATLAB and Orange3 implementation are PC4 and PC1 corresponding to “Cleanliness of airport terminal” and “Thoroughness of security inspection” respectively. This implies that the two PCs contribute a significant amount of information to the model, and the satisfaction of the customers can almost be entirely determined by these two features.



Evaluation Results

Model ▼	AUC	CA	F1	Precision	Recall
Tree	0.932	0.826	0.814	0.803	0.826
SVM	0.951	0.860	0.860	0.860	0.860

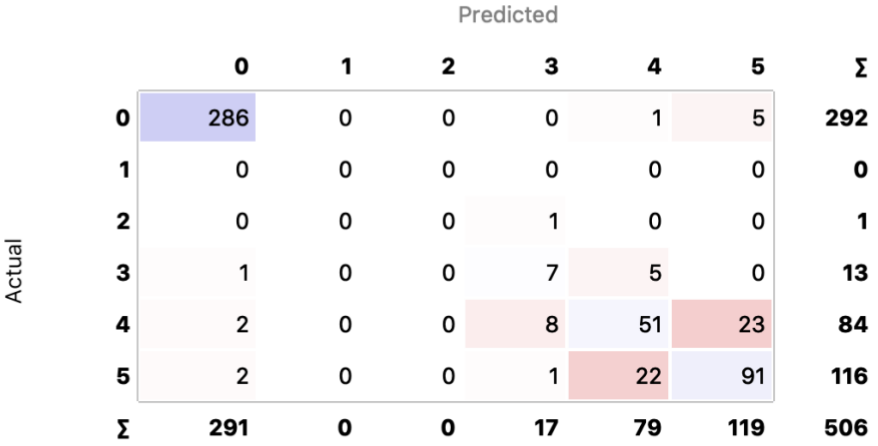
Figure 3.7. The evaluation metrics for the Linear-SVM and Decision tree models.



Predicted

Actual	0	1	2	3	4	5	Σ
0	283	0	0	0	1	8	**292**
1	0	0	0	0	0	0	**0**
2	0	0	0	0	1	0	**1**
3	3	0	0	0	9	1	**13**
4	7	0	0	0	47	30	**84**
5	5	0	0	0	23	88	**116**
Σ	**298**	**0**	**0**	**0**	**81**	**127**	**506**

Figure 3.8. Confusion matrix of the Decision tree model.



Predicted

Actual	0	1	2	3	4	5	Σ
0	286	0	0	0	1	5	**292**
1	0	0	0	0	0	0	**0**
2	0	0	0	1	0	0	**1**
3	1	0	0	7	5	0	**13**
4	2	0	0	8	51	23	**84**
5	2	0	0	1	22	91	**116**
Σ	**291**	**0**	**0**	**17**	**79**	**119**	**506**

Figure 3.8. Confusion matrix of the linear-SVM model.

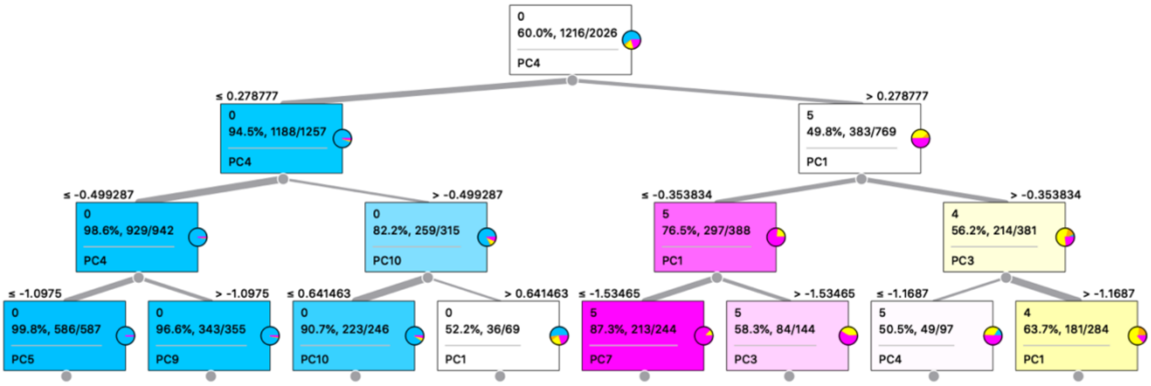


Figure 3.9. Decision Tree Diagram for Orange implementation.

Chapter 4

4.1. Discussion and Conclusions

After performing the data cleaning, pre-processing and feature engineering tasks, the ten components that correlate with the customers satisfaction are listed in Figure 3.2 with the top 3 correlating features being “Arrival passport and visa inspection”, “Speed of Baggage delivery” and “Customs inspection”. From experience, these features will have a significant contribution to the experience of a customer most especially “speed of baggage delivery”. Most people do not want to wait for their baggage for more than 2-5 hours as they have just landed and are tired and jet-lagged. It can also be observed that the feature engineering process that produced the 4 and 5 top correlating features, i.e. (Encoded departure time and Encode quarter) have a relatively significant correlation with the overall satisfaction.

The top 10 PCs shown in Figure 3.1 seem to make more sense, how they can be strong factors that determine customers satisfaction. For instance, the principal component (PC1) corresponding to “Thoroughness of security inspection”, It is highly understandable how this feature would contribute to satisfaction levels, security is one of the most fundamental concerns for anybody travelling, even before comfort and feeding. Also, given the history of terrorist attacks at airports from the late 1990s till date, it is easy to see how security has a major contribution to the customer’s satisfaction. Then other factors such as value for money at shopping facilities and cleanliness of airport terminals, would intuitively make sense as to how they contribute to customers satisfaction. People travelling would like to buy some things for their loved ones or colleagues and if the tax at the airport shopping centres is too, customers may feel disappointed buying the item or feel discouraged to buy anything in the first place. This will clearly have a significant impact on the satisfaction level of the customers.

From the information learned so far in the feature selection and data exploration process, recommending the following steps would potentially improve the customers satisfaction:

- Investigate why the customers seem to have a lower satisfaction rate between 3pm to 9pm and take action to address the problem within that time frame. This could be providing more entertainment to lighten the mood or providing more staff to help guide the customers through the airport and assisting with their baggage as customers are more likely to be stressed mid-day to night-time.

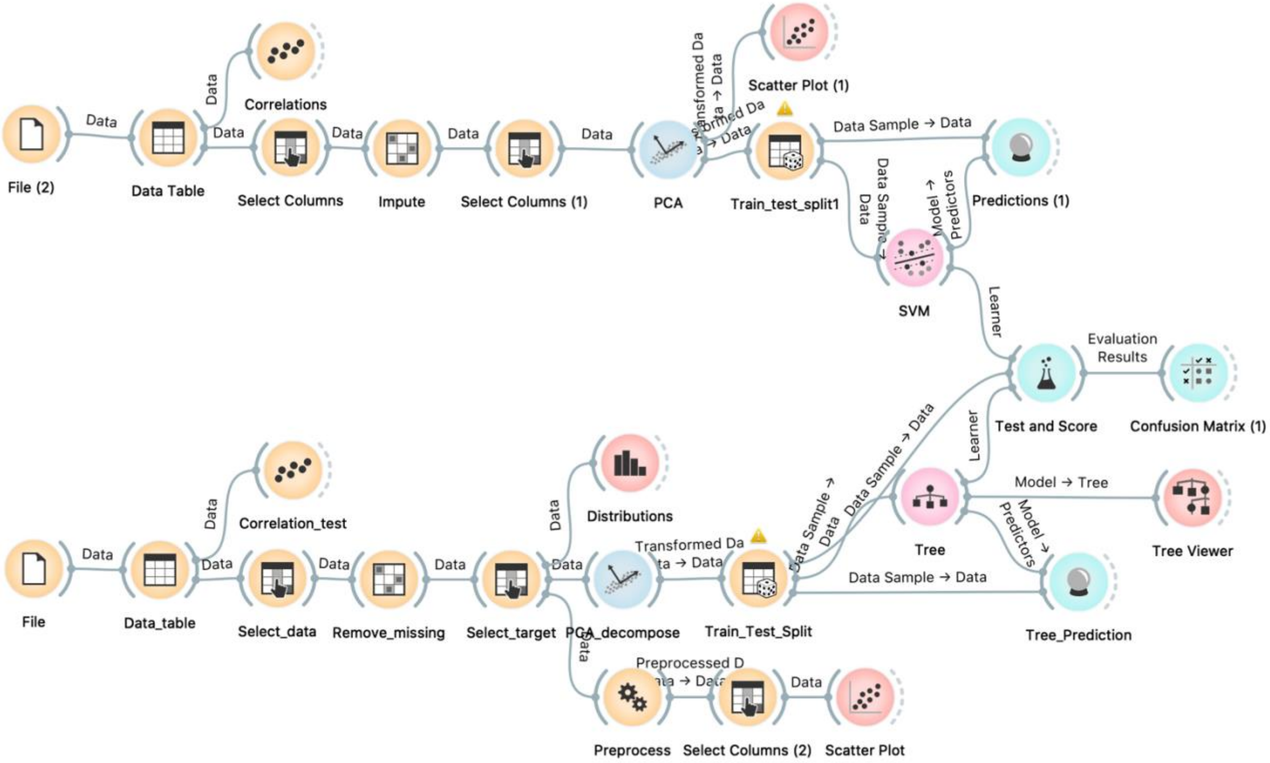
- It would be helpful to study why the airline/airport had bad satisfaction rating between the second quarter of 2015 and second quarter of 2016. It also needs to be determined if what-ever factor that caused a stall in customers satisfaction is likely to occur again and put measures in place to address that problem.
- It is also a good plan to ensure security and cleanliness/hygiene are of top priority for airlines and airports as they are the two most important factors that contribute to a customer's satisfaction.

Regarding the two models implemented in this project. Although both models gave a good performance and did not overfit or underfit, The SVM model seemed to have been more generalized to the data with less variance than the Decision tree model. Moving forward, it would be beneficial to use the SVM as a clear separation can be observe from the PCA plots in Figure 3.3. Using a linear hyperplane, it would be easy to classify the different satisfaction ratings.

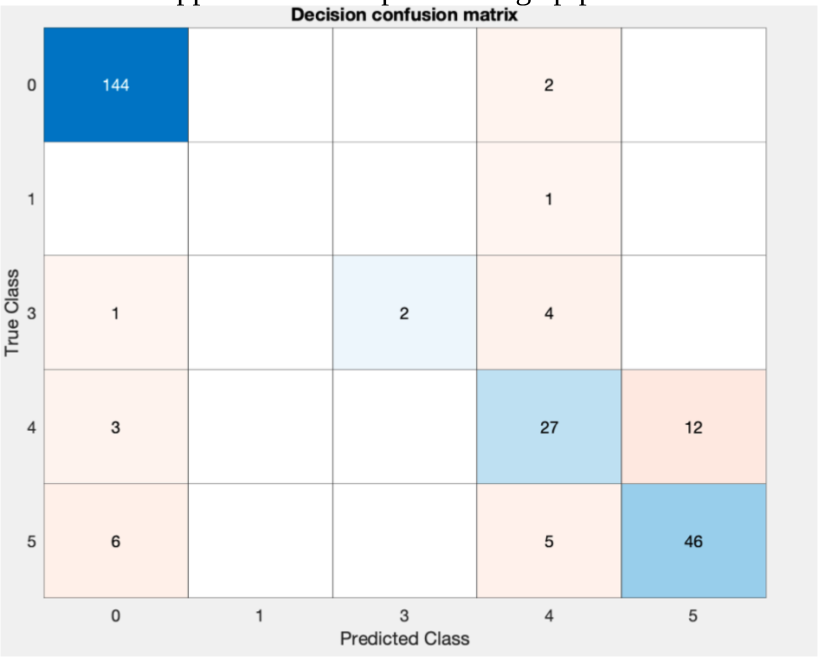
References

Ban, Hyun-Jeong, and Hak-Seon Kim. "Understanding customer experience and satisfaction through airline passengers' online review." Sustainability 11.15 (2019): 4066.

Appendix



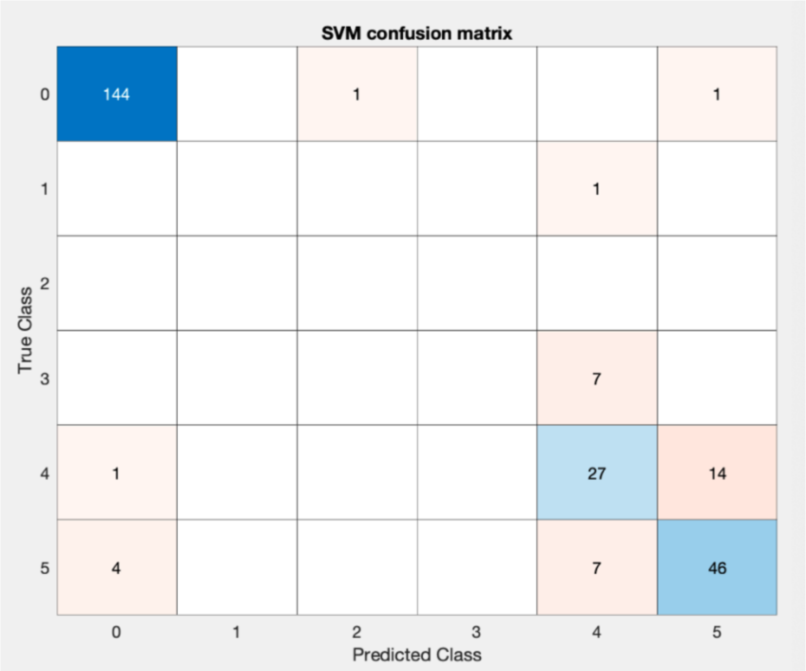
Appendix 1. Complete Orange pipeline.



Decision confusion matrix

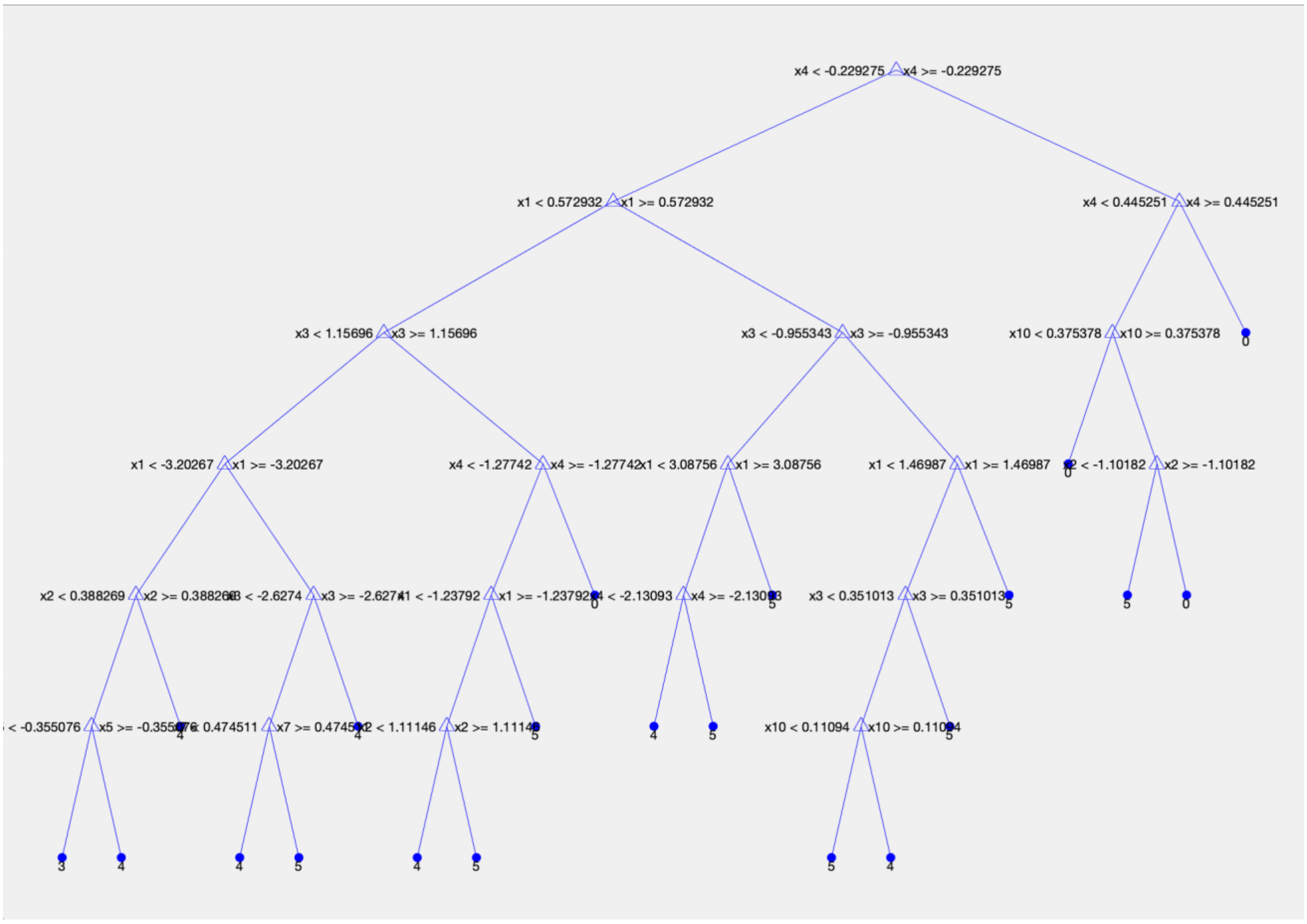
True Class \ Predicted Class	0	1	3	4	5
0	144			2	
1				1	
3	1		2	4	
4	3			27	12
5	6			5	46

Appendix 2. MATLAB confusion matrix results for Decision tree.



True Class \ Predicted Class	0	1	2	3	4	5
0	144		1			1
1					1	
2						
3					7	
4	1				27	14
5	4				7	46

Appendix 3. MATLAB confusion matrix results for SVM model.



Appendix 4. Tree diagram from MATLAB Implementation.