



# **GROUP ASSIGNMENT**

**TECHNOLOGY PARK MALAYSIA**

**CT099-3-3-BDA**

**Knowledge Discovery and Big Data Analytics**

**APU3F2209CS(IS)/APD3F2209CS(IS)**

---

Name : Yan Mun Kye (TP056066)

: Sia De Long (TP060810)

: Hor Shen Hau (TP061524)

: Tan Sheng Jeh (TP056267)

Lecturer : DR. MURUGANANTHAN VELAYUTHAM

Hand-in date : 1 May 2023

## Table of Contents

1.0 Introduction.....	4
1.1 Problem statement.....	4
1.2 Aim (Group).....	4
1.3 Objectives .....	4
2.0 Methodology .....	5
3.0 Literature review .....	7
3.1 Decision tree .....	7
3.2 Random forest.....	11
3.2.1 Ensemble Methods.....	11
3.2.2 Random Forest Algorithm .....	12
3.2.3 Prediction of Student Course Performance.....	12
3.3 Deep learning .....	14
3.4 K nearest neighbour .....	19
4.0 Data preparation.....	23
4.1 Exploratory data analysis.....	23
4.1.1 Visualisation on numerical data.....	24
4.1.2 Visualisation on categorical data .....	27
4.1.3 Multivariate analysis.....	29
4.2 Data pre-processing .....	37
5.0 Modelling.....	40
5.1 Model 1: Decision tree.....	41
5.2 Model 2: Random forest .....	47
5.3 Model 3: Deep learning .....	52
5.4 Model 4: K nearest neighbour.....	57
6.0 Evaluation .....	62
6.1 Comparison .....	62

6.2 Critical analysis and recommendations .....	63
References.....	65
Workload matrix .....	69

## **1.0 Introduction**

### **1.1 Problem statement**

Being a highly competitive industry with many competitors from all around the world, it is important for airline companies to maintain the satisfaction of passengers for them to be successful in the industry. Naturally, a satisfying flight experience will lead to an increase in passenger satisfaction, which in turn brings in loyal customers, positive reviews and ultimately increased revenue. There may be several factors, including inflight experience and airport service experiences. It is important to study the satisfaction of the airline company's passengers to ensure the company can identify their strengths and weaknesses while taking steps to maintain or improve the satisfaction level.

However, it may be a challenge to study passenger satisfaction as there may be too many factors imposing an influence, and it is highly subjective to classify "satisfactory experience". Therefore, this project aims to utilise data mining techniques to objectively analyse the factor affecting passenger satisfaction and to predict satisfaction levels. This way, actionable insights can be generated for the airline company so they can develop strategies for improving their service.

### **1.2 Aim (Group)**

To increase passenger satisfaction of an airline company

### **1.3 Objectives**

- To determine factors affecting passenger satisfaction via descriptive analysis
- To build machine learning models to predict passenger satisfaction
- To determine the suitability of each machine learning model to be used along with justification
- To determine actions to be taken with the results of descriptive and predictive analysis.

## 2.0 Methodology

The data mining methodology that will be used by the researchers will be the KDD methodology short for Knowledge Discovery in Databases. It can be officially defined as the process to extract useful information and data from large databases (Javatpoint, n.d.). However, in this context, the researchers will focus on the specific dataset chosen, which is the Airline Passenger Satisfaction dataset. The KDD methodology can be summed up with the figure below:

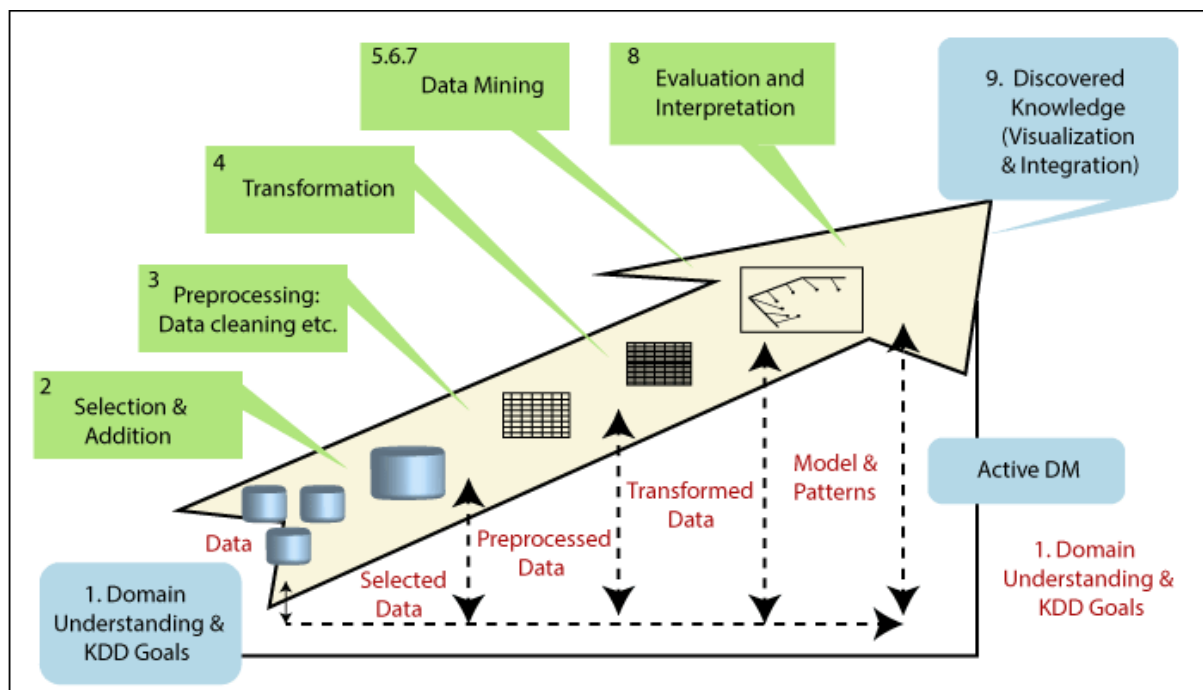


Figure 1 - KDD methodology

Firstly, the researchers have selected a subset of data to perform data mining. This is to define the scope of the project and to understand the business needs and objectives. The researchers need to fully understand the context behind the problem, only they can proceed with the data mining process to solve the particular problem.

Next, the researcher need to perform data preprocessing, which include steps like data cleaning, normalisation, missing value imputation and data augmentation. This aim to remove the anomalies within the dataset which may negatively affect the performance of the data mining model.

Next, data transformation will be carried out where the cleaned data is transformed into suitable format for the data mining algorithm, like matrix or graph. This is because different data mining algorithms have different requirements on the data format. This is especially true

for models which only can work on numeric input, in which categorical data need to be transformed into numeric format.

Next, the data mining itself where techniques and algorithms are used to extract useful information from the data. This may include relatively simple techniques like statistics, or more complicated techniques which include machine learning techniques. In general, the tasks can be divided into clustering, classification, association rule mining and anomaly detection.

After that, evaluation and interpretation of the results of the model will be taken place. The researchers will need to visualise the results, evaluate the quality of the discovered patterns and identify possible relationship among the data. The researcher also need to evaluate the model using evaluation metrics like accuracy, mean square error and precision to determine viability of the data mining method.

### 3.0 Literature review

#### 3.1 Decision tree - Yan Mun Kye (TP056066)

The field of machine learning is gaining more attention in various industries, including medical, customer relations, and finance. Machine learning has been applied in various real-world scenarios to help users make more informed decisions. Decision tree is one of the popular algorithms used to perform machine learning tasks. It has gained merits in its simplicity and straightforward implementation and interpretation (Chen et al., 2020).

A decision tree is a data structure consisting of leaf nodes and internal nodes. In the case of performing classification tasks, the internal node of a decision tree represents a feature or attribute that is used to split the data into smaller subsets. Each internal node represents a decision that needs to be made based on the feature or attribute. Leaf nodes store the label data information. During the training stage, the decision tree tries to find the best feature to do the splitting, in which results in the highest information gain compared to other features. There are several algorithms that can be used to calculate the best splitting feature, which include CART, C4.5 and ID3.

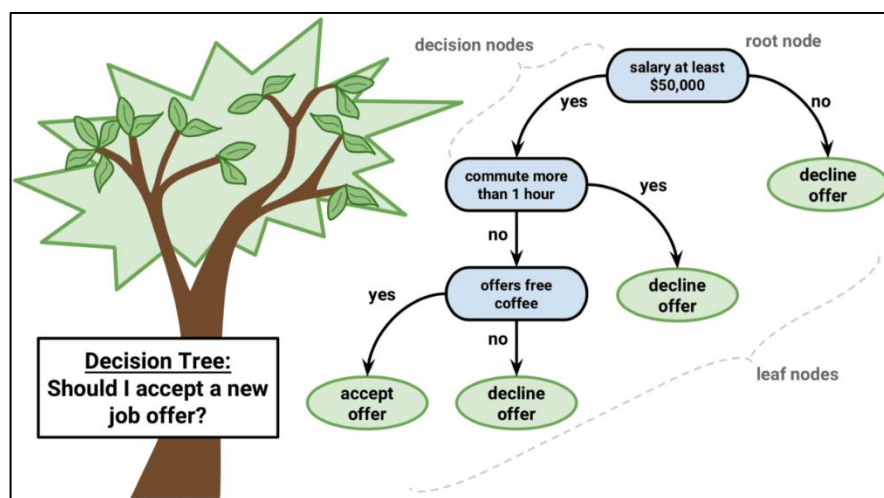


Figure 2 - Simple example for decision tree

Iterative Dichotomiser 3 (ID3) is a simple algorithm which employs a top-down, greedy approach to search for the property with the highest information gain. However, ID3 can only be used on categorical data. C4.5 is an extension towards the ID3 algorithm that is able to be used on both numerical and categorical data (Sharma & Kumar, 2016), however it uses information gain ratio as splitting criteria. C4.5 also enables handling missing values as missing attributes values are not utilised in the gain ratio calculations. Depth-first strategy is also used to grow by recursively splitting the data, and nodes are chosen based on depth-first

strategy. Classification and Regression Trees (CART) utilises Gini index to split data, which measures the divergence between the probability distribution of the attribute. One important note ability of CART is the ability to generate regression trees (Hssina et al., 2014). Below shows a comparison between the decision tree algorithms:

Features	ID3	C4.5	CART
Type of data	Categorical	Continuous and Categorical	Continuous and nominal attributes data
Speed	Low	Faster than ID3	Average
Boosting	Not supported	Not supported	Supported
Pruning	No	Pre-prunning	Post-prunning
Missing value	Cannot handle	Cannot handle	Can handle
Formula	Information entropy and information gain	Information split and gain ratio	Gini diversity index

Table 1 - Comparison between ID3, C4.5 and CART (Sharma & Kumar, 2016)

When the decision tree had finished training, it will produce a fully-grown tree with various splitting and several layers deep. However, a fully-grown tree is prone to overfitting. Therefore, pruning techniques are introduced to avoid overfitting. In general, there are pre-pruning and post-pruning methods. Pre-pruning methods uses early-stopping conditions like the predefined tree height or limited number of objects. Post-pruning methods is triggered after the fully-grown tree is produced. A validation method will be used to check each node to estimate the accuracy of the model. If removing a node satisfies both reducing complexity and improve accuracy, the node will be removed (Chen et al., 2020).

There are many instances of decision tree being used in various scenarios. (Zhu et al., 2018) had implemented decision tree on customer churn prediction, however their research focusses on the class imbalance within the dataset. They have performed their data mining on various datasets of the telecommunication industry. Using the each dataset, pre-processing steps are carried out and the decision tree model is applied. The authors had used CART and C5.0 along with pruning, unpruning setting, SMOTE, random over-sampling and random under-sampling techniques. They had concluded SMOTE sampling technique is best for both CART and C5.0 if only one pruning or oversampling technique can be selected.

Another implementation of decision tree on customer churn prediction is the research by (Li et al., 2023). The authors had combined one-dimensional convolutional neural network (1DCNN) and gradient boosting decision tree (GBDT) to predict customer churn. The



prediction of the 1DCNN will be fed into the GBDT for second forecast in case the result was not-churn. It was shown by the authors that the F1 score and recall rate of their proposed solution is significantly higher than other models when a comparative experiment was carried out. The model was able to produce high performance prediction while omitting the need for laborious human feature engineering.

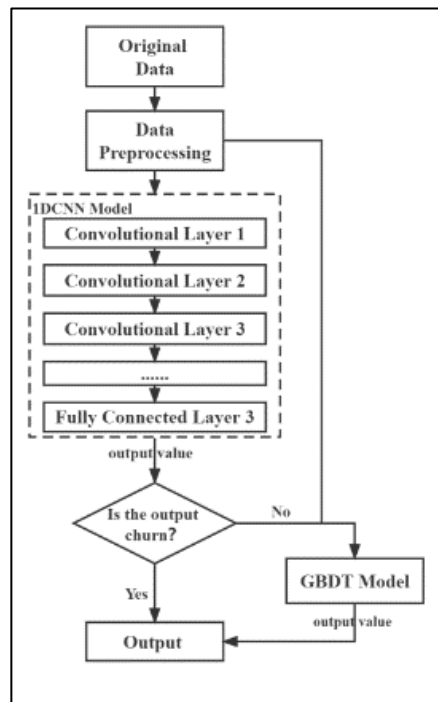


Figure 3 - structure of the combined model

Another use case of decision trees is being used in predicting diabetes implemented by (Wu et al., 2022). The authors leveraged decision trees and rough set to perform classification predictions. With their dataset containing 17 attributes, the implementation using rough sets, redundant attributes were removed. Using the nuclear attributes, the decision tree is built and ten-fold cross-validation is used to deal with over-fitting problem. This results in an improved classification accuracy and reduced time consumption in the classification task.

An effort in predicting customer credit risk using decision tree is performed by (Perera, 2019) using a dataset from finance/leasing company in Sri Lanka. Out of the 24 variables of the dataset, only 9 were chosen to develop the prediction model due to multicollinearity problems found using VIF factor values. The author was able to achieve up to 92%, making it a suitable model to predict payment of leasing customers in Sri Lanka.

Lastly, it is worth noting that there are also advancements in the decision tree algorithm, in which (Manapragada et al., 2018) had proposed a novel Extremely Fast Decision Tree

(EFDT). The proposed algorithm was able to achieve higher prequential accuracy than the previous model as shown by the authors. It was also illustrated that the EFDT algorithm learning is more rapid and less affected by the complexity of the learning task compared to the previous model.

### **3.2 Random forest - Tan Sheng Jeh (TP056267)**

Remote sensing has emerged to prove its value in various sectors; however, the success of image classification relies on different factors, involving the selection of a suitable classification procedure. (Belgiu & Drăguț, 2016) ascertained that supervised classifiers are majorly utilized because they look more robust compared to the model-based techniques. Various scholars including (Speiser et al., 2019) defined random forest as a famous machine learning procedure that can be utilized to establish prediction models. The classifiers can easily learn the features of classes aimed from the training samples as well as to determine the learned features within the unclassified data. The scholars argued that over the past two decades the application of this machine has had an increasing attention due to its excellent classification outcome as well as its processing speed. The classifier of RF results in reliable classifications utilizing predictions drawn from an ensemble of decision trees (Speiser et al., 2019). In addition, this classifier can be utilized successfully in selecting and ranking of variables having the great ability to discriminate between the targeted classes. The researchers argued that it is an essential asset given that the high dimensionality of remotely sensed data can make the selection of the most relevant variables a time-consuming, subjective task, and error prone.

Since the random forest model is made up of various decision trees as described by (Belgiu & Drăguț, 2016), it is important to understand the decision tree algorithm first. The decision trees seek in finding the best split in subsetting the data. These are typically trained via the classification and regression tree algorithm. According to the researchers, the metrics, including mean square error, information gain, or Gini impurity, can be utilized to evaluate the split quality. While decision trees are the most common supervised learning algorithms utilized, they are sometimes prone to problems, involving overfitting and bias. However, when different decision trees form an ensemble in the RF algorithm, more accurate results are predicted, specifically when trees are uncorrelated with each other.

#### **3.2.1 Ensemble Methods**

The Ensemble learning techniques are made up of a set of classifiers, for example, the decision trees. According to Speiser et al (2019), the predictions of these decision trees are aggregated to determine the most popular outcome. The most likely and well-known ensemble models are bagging, usually understood as bootstrap aggregation, and boosting.

Leo Breiman (link resides outside ibm.com) (PDF, 810 KB) launched the bagging model in 1996; where in this technique, a random sample of data within the training set is chosen with replacement. This means that the single data points can be selected even more than one time. Immediately the multiple data samples are generated, the researchers argued that these models are then trained one by one, however, trained depending on the kind of the task. This can be in the form of classification or regression, where the average or most of such predictions result in a more accurate estimate. The researchers further noted that this approach is commonly utilized to minimize variance within a noisy dataset. (Schonlau & Zou, 2020) argued that the ensemble classifiers within the remote sensing have supervised parametric classifiers, which include the Maximum Likelihood Classification (MLC), which can deliver excellent work when dealing with unimodal data. On the other hand, this type has challenges when dealing with multi-modal input datasets as these classifiers usually assume a normal distribution of data.

### **3.2.2 Random Forest Algorithm**

The random forest algorithm on the other hand represents an extension of the bagging technique because it uses both bagging and feature randomness to help develop an uncorrelated forest of decision trees. The feature randomness, which is also understood as the feature bagging or “the random subspace method, is geared to generate a random subset of features that ensures low correlation between the decision trees. Importantly, the random forests select a subset of the features. (Schonlau & Zou, 2020) argued that the RF classifier stability is an essential criterion for its integration into an operation setting. In the past, studies reported that the general accuracy of classification of the random forest classifier reduces when the algorithm is conducted or maybe trained on different study settings.

### **3.2.3 Prediction of Student Course Performance**

Education is a fascinating and essential sector which has experienced growth over the past years and has significantly impacted on the lives of many individuals. Various methods and techniques have been proposed to help establish high quality experiences to benefit the entire field. According to (Xu & Yin, 2021), in the past approaches of teachers and learners evaluation performance systems, in most areas, individuals have utilized average score or the total score to evaluate learners. Since then, random forest algorithm was proposed by Leo Breiman. According to the researchers, this system is composed of various sub models, where

the output of every sub model is combined together to provide the final result. By utilizing different classifiers for voting classification, random forest algorithms can minimize the error efficiently to enhance the accuracy of classification (Sheykhmousa et al., 2020). There are many available factors that can influence physical education. These may include some controllable and uncontrollable factors, which can either affect the student's performance directly or indirectly.

### **3.2.4 Random Forest Applications**

The random forest algorithm is one of the machine tools, which has been utilized across various industries including the IT industry, which has allowed them to come up with better decisive businesses.

#### **3.2.4.1 Finance**

According to Schonlau & Zou (2020), random forest is the most preferred algorithm ranked ahead of others. Researchers highlight that it minimizes the time spent during data management and pre-processing of tasks. The scholars also stressed that the tool can be utilized to detect fraud, to evaluate consumers having high credit risk, and option pricing challenges.

#### **3.2.4.2 Healthcare**

According to Schonlau & Zou (2020), the random forest algorithm as a tool has uses within computational biology, which allows the stakeholders within this sector, primarily the doctors to handle problems, for instance, the sequence annotation, biomarker discovery, and gene expression classification. In the end, through the data obtained by the random forest, doctors are able to produce estimates around the drug response towards certain specific modifications.

### 3.3 Deep learning - Hor Shen Hau (TP061524)

Deep learning models are basically machine learning models that are trained using a technique that teaches computers the things and ways that come as natural to humans, essentially the learning by example method. This enables computers to be able to perform feats such as distinguishing a car from a dog, recognize road signs and allow for modern technologies that we take for granted now such as autonomous driving and voice control in devices. Deep learning models essentially performs classification tasks on varying forms of media input such as images, text and sound. Deep learning of recent years have made big strides in advancement such as achieving all time high accuracy levels and outperforming humans in some tasks. This has made deep learning a popular choice for applications in various industries from autonomous vehicles to medical devices and even aerospace and defense. The science behind deep learning models is that the term “deep” refers to the hidden layer count in the neural network architecture that the deep learning model is based off hence its other name, deep neural networks. Unlike neural networks which traditionally only had 2-3 hidden layers, deep learning networks can have upwards of 150 or more. Deep learning models’ exceptional performance does come with a few caveats such as its requirement of large, labeled datasets to be paired with neural network architectures that are able to extract features and learn directly from the data without the need for any manual extraction. (*What Is Deep Learning?*, n.d.)

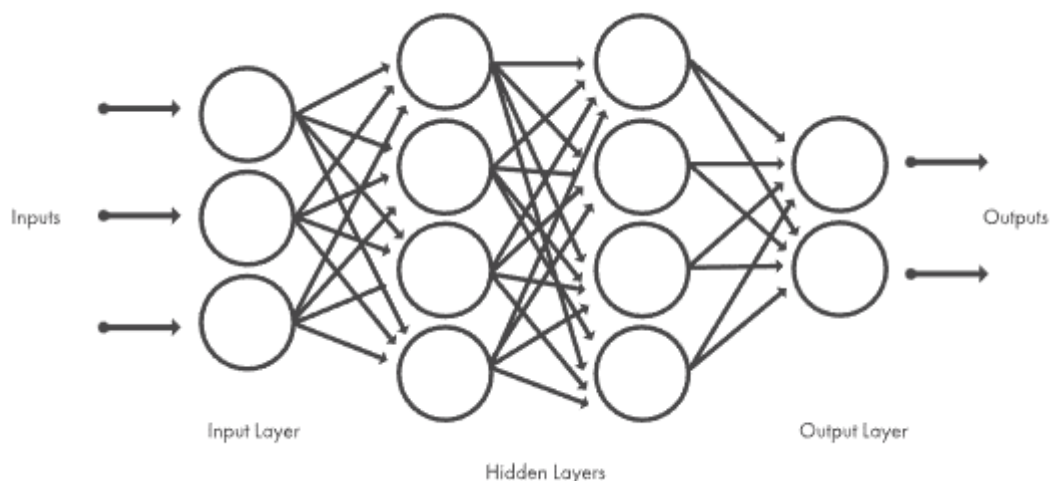


Figure 4 - Simple Neural Network showing the basic architecture of a deep learning model.

In a paper by (Rasool Fakoor & Azade Nazi, n.d.), they have proposed to using deep learning in the medical field with a noble intention of its application for cancer diagnosis. The proposed method is able to apply data on various kinds of cancer and automatically learn its features which will help in the detection and diagnosis the specific cancer. The researchers used the autoencoder neural network which is an unsupervised feature learning algorithm that only has one hidden layer to be paired with a softmax regression deep learning classifier to obtain their results. The classifier is then evaluated with a 10-fold cross validation while also compared it with two other classifiers, SVM and Softmax Regression as baselines. The researchers has obtained results that show their proposed method has achieved better accuracy over the baselines for cancer classification problems.

Meanwhile (Kowsari et al., 2017), proposes a text classification method using hierarchical deep learning to aid in the processing large document operations such as searching, retrieving and organizing. The researchers here have employed a unique deep learning approach where different kinds of deep learning algorithms were used together such as Deep Neural Networks (DNN), Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). The DNN architecture, which is the first layer of classification (left side of the Figure 3) only takes input from the previous layer and then outputs to the next layer. The second level in the hierarchy is also made up of a DNN which is connected to the output of the first DNN which provides a sample output scenario of where the first level of DNN is trained with the all the documents available while the second layer is trained only with the documents that are specified in the domain. The RNN and CNN are then used to perform classification on a lower level which produced significantly better accuracies when compared to more conventional approaches such as Naïve Bayes and SVM in document classification.

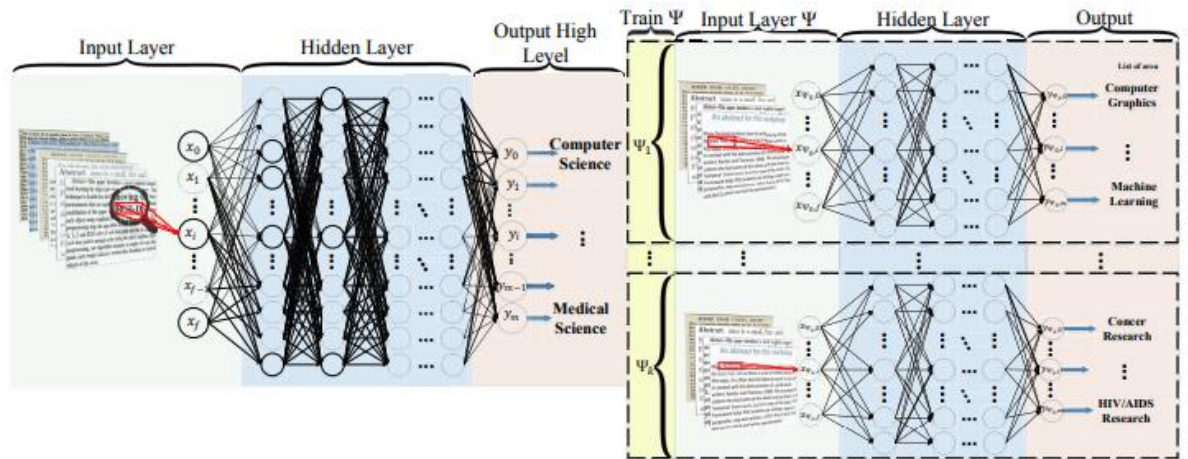


Figure 5 - Proposed Deep Learning Architecture for hierarchical text classification

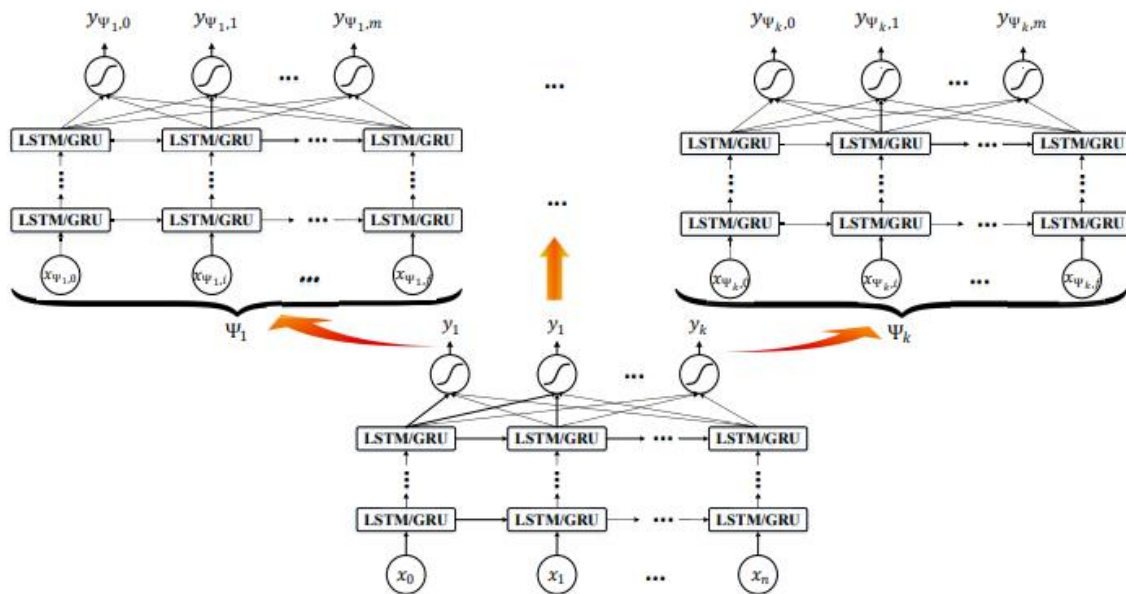


Figure 6 - Hierarchical Deep Learning Models DNN, RNN and CNN



In light of the pandemic that has struck humanity as a surprise four years ago, some researchers have proposed the use of deep learning models for social distancing detection to try and alleviate the pandemic problem whilst retaining most of their daily activities. The researchers here employed computer vision techniques with a deep convolutional neural network which was able to measure the distance between detected pedestrians from a top-down view. The model was able to measure and scale the distance between the pedestrians and then send pre-cautionary warnings to the system if any of the pedestrians' distances are below the preset threshold. The researcher used the YOLOv3 model as the Deep CNN model of choice to handle the computer vision component for pedestrian detection. The researchers were able to estimate distance between people with real testing. (Hou et al., 2020)

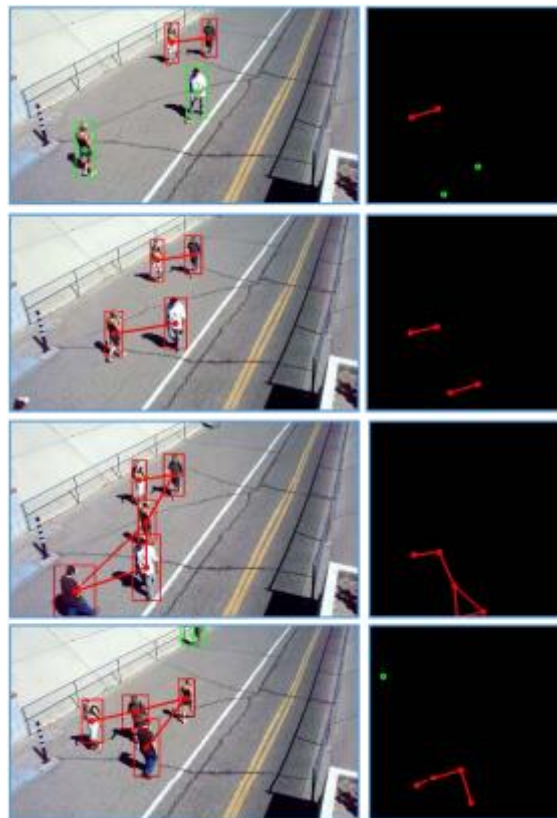


Figure 7 - Proposed Social Distancing Detection Demo

With transportation being the core of what makes a country prosper, modern day transportation systems have been becoming more efficient and faster with the implementation of deep learning techniques in problems that were previously addressed using traditional means such as statistics and analytical approaches. The researchers in (Veres & Moussa, 2020) proposes the use of deep learning for demand predictions which is able to tell where and when a trip will start to better allocated resources such as adding more busses to areas that are predicted to have higher demand and also be able to estimate the impact to transportation by new building constructions. Similar to the proposal of (Rasool Fakoor & Azade Nazi, n.d.), autoencoders are necessary here as the deep learning model used here requires a large amount of data. Ultimately the researchers were able to utilize deep learning in the application of transportation.

### 3.4 K nearest neighbour - Sia De Long (TP060810)

In the research paper (Zhang, 2016), the researcher introduces K-NN model in machine learning by demonstrate the usage of the model on a problem of categorising fruit, vegetable and grain. First and foremost, Unlabelled observations are categorised by the K-NN classifier by placing them in the category of the most comparable labelled samples. Both the training dataset and the test dataset contain information about the observations. Only two characteristics are used in order to display them on a two-dimensional plot. In practise, any number of predictors may be present, and the example may be expanded to include any number of traits. The knn() function by default uses the Euclidean distance, which can be determined using the equation below.

$$D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

Equation for K-NN algorithm (Zhang, 2016)

The variables p and q in the equation stand for the comparison subjects with n attributes. The Manhattan distance is one of many methods for calculating distance. Another idea that is mentioned in the research report is that the parameter k, which determines how many neighbours will be used in the K-NN algorithm, has a big impact on how well that algorithm performs in diagnostic tasks. According to the research, a big k minimises the effect of variance brought on by random error, but it also increases the chance of missing a subtle but significant pattern. In order to select a suitable k value, it is important to strike a balance between overfitting and underfitting. The researcher claims that some other data scientists advise setting k equal to the square root of the quantity of observations in the training dataset. Due to the importance of k value in K-NN model, this research paper showed one of the approaches to tune k value with the details of analysis.

In the research paper (Batista & Silva, 2009), the researcher studied on the influence of the parameters bring to the performance of K-NN model and the use of K-NN as a classification algorithm in this study is what the researchers are most interested in. The algorithm's main objective, for instance, is to select the class of a new case based on the classes of the k most comparable database components while each database element has a label assigned to it. In addition, this study discusses and delivers empirical data regarding how the key K-N factors affect performance. The number of nearest neighbours, distance

function, and weighting function are therefore the parameters involved and explored in this research, and the most common parameter selections for these three values were assessed. To achieve the aim of the research, experiment on different configurations on the parameters of K-NN model were conducted while 31 benchmark and "real-world" data sets were used over the course of the experiments. To conclude, the researchers came out with the conclusion of the best  $k$  value for each type of the K-NN model as shown as below.

Weighting	Distance	Null-hypothesis	Best $k$ value	Outperformed $k$ values
None	HEOM	Rejected	7	1, 15, 21, 27
None	HMOM	Rejected	5	1, 15, 21, 27
None	HVDM	Rejected	7	27
Similarity	HEOM	Rejected	7	1, 15, 21, 27
Similarity	HMOM	Rejected	5	1, 15, 21, 27
Similarity	HVDM	Rejected	7	27
Inverse	HEOM	Rejected	5	1, 27
Inverse	HMOM	Rejected	5	1, 21, 27
Inverse	HVDM	Not rejected	11	-

Figure 8 - Summary of statistical test results for the second null-hypothesis (Batista & Silva, 2009)

In the research paper (ENAS & CHO! , 1986), the researchers studied on the choice of the smoothing parameter and efficiency of K-NN classification. The research had discovered that the best value for  $k$  depends on the sample space's dimension, size, covariance structure, and sample proportions. The correct classification rates for the closest neighbour rules chosen using the  $k$  recommended by the simulations were at least as high as those rates for the logistic regression approach and the linear discriminant function. In short, the rule improved in efficacy as the covariance matrix difference grew and as the sample proportion difference grew. It is demonstrated that an adaptive rule that chooses  $k$  by repeatedly maximising the local Mahalanobis distance is effective. Based on the research, A family of nonparametric classification rules have been demonstrated empirically to perform as good as or better than the traditional parametric approaches for a combination of categorical and continuous variables for small to moderate sample sizes. The sample proportions and underlying covariance structure for small samples have been proven to affect the choice of  $k$  for the nonparametric closest neighbour rules' best performance. Though more research is necessary, the study offered the following general recommendations for choosing  $k$  for the best classification based on the size of the training set  $N$ :

## Difference Between Sample Proportions

		Small	Large
Difference Between Covariance Matrices	Small	$N^{3/8}$	$N^{2/8}$
	Large	$N^{2/8}$	$N^{3/8}$

*Figure 9 - Difference between sample proportions (ENAS & CHOI, 1986)*

According to the researchers, as the covariance matrices become more different or as the variations in sample sizes and distribution shapes grow, it is not surprising to see that the K-NN rules perform more efficiently than LDF and logistic regression methods. When the global dispersion for each population is radically different, the local optimality of the K-NN rules is significantly observed.

In the research paper (Agrawal, 2014), the researcher conducted research on the fundamentals of various existing data classification techniques for uncertain data using the K-NN approach. According to the research, finding a similar class of tuples is a difficult task because uncertain data incorporates tuples with various data. When compared to attributes with lower levels of uncertainty, attributes with higher levels need to be handled differently. Besides that, the research stated that there are many works has been done in this area but still there are certain performance issues in the K-NN classifier. The research defines the K-NN in the category of a lazy learner since the distance between training records is calculated in order to categorise an unknown record. The K nearest neighbours is selected based on the distance, and the class labels of these neighbours are utilised to determine the class label of the unclassified record. The illustration of K-NN is as shown as below.

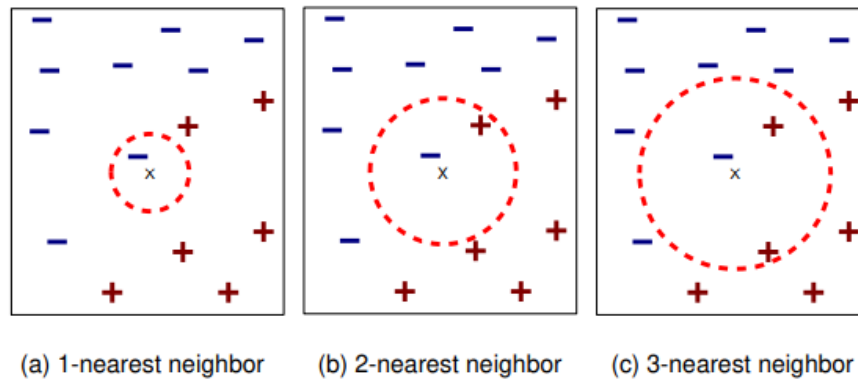


Figure 10 - K-NN illustration (Agrawal, 2014)

Moreover, the research concluded the K-NN issues in to three which first most crucial problem is choosing  $k$ . It might choose too many irrelevant data objects as nearest neighbours if it is too big. On the other hand, if  $k$  is too tiny, it might not yield the desired outcome. Secondly, the next issue is selection of distance units, the dimensionality of the data may have an impact on some distance measurements. Lastly, Because KNN classifiers are sluggish learners, it is relatively expensive to classify unknown objects using this method.

## 4.0 Data preparation

### 4.1 Exploratory data analysis

Name	Type	Missing	Statistics			Filter (25 / 25 attributes)	Search for Attributes
att1	Integer	0	Min 1	Max 103903	Average 51952		
id	Integer	0	Min 1	Max 129980	Average 64924.160		
Gender	Nominal	0	Least Male (51176)	Most Female (52727)	Values Female (52727), Male (51176)		
Customer Type	Nominal	0	Least disloyal [...] (18981)	Most Loyal Customer (84922)	Values Loyal Customer (84922), disloyal Customer (18981)		
Age	Integer	0	Min 7	Max 85	Average 39.380		
Type of Travel	Nominal	0	Least Personal Travel (32248)	Most Business travel (71655)	Values Business travel (71655), Personal Travel (32248)		
Class	Nominal	0	Least Eco Plus (7493)	Most Business (49665)	Values Business (49665), Eco (46745), ... [1 more]		
Flight Distance	Integer	0	Min 31	Max 4993	Average 1199.455		
Inflight wifi service	Integer	0	Min 0	Max 5	Average 2.730		
Departure/Arrival time conveni...	Integer	0	Min 0	Max 5	Average 3.060		
Ease of Online booking	Integer	0	Min 0	Max 5	Average 2.757		
Gate location	Integer	0	Min 0	Max 5	Average 2.977		
Food and drink	Integer	0	Min 0	Max 5	Average 3.202		

Showing attributes 1 - 25 Examples: 103,903 Special Attributes: 0 Regular Attributes: 25

Figure 11 - Attributes in the airline passenger dataset

The dataset is imported into RapidMiner. From the overall summary of the dataset, the researchers determined that the attr1 column is unused indexing column, while the id column is only used to uniquely identify each customer. Both the columns will be removed as it will not be used in the further steps.

Arrival Delay in Minutes	Real	310	Min 0	Max 1584	Average 15.179
--------------------------	------	-----	----------	-------------	-------------------

Figure 12 - missing values found

Next, it is noted that one of the column “Arrival Delay in Minutes” has 310 missing values. The researchers decided to remove all the rows with missing value as the number of records to remove is almost insignificant when compared to the whole dataset which has a total of 103,903 rows.

### 4.1.1 Visualisation on numerical data

Carrying on, the researchers will study the distribution of the columns with numerical data.

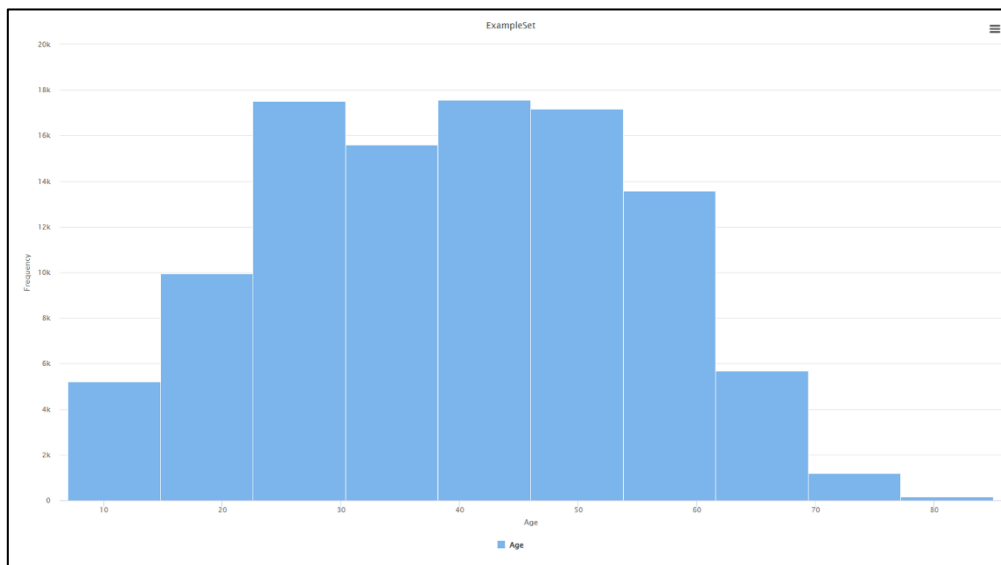


Figure 13 - distribution of age

The above histogram shows the distribution of the Age column. The data seem to approximately follow the normal distribution with very comparatively less people with the age 70 and above.

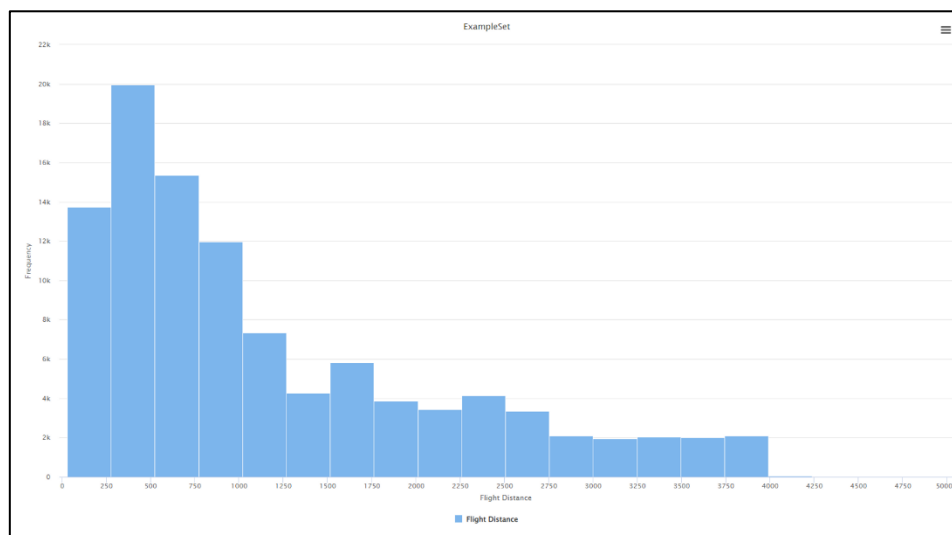


Figure 14 - distribution of flight distance

The distribution of Flight Distance shows that most of the flight are of shorter distances rather than long distances as it is heavily right skewed. It is evident that there are some observations with flight distance more than 4000, however, the number of observation is



small compared to the number of observation of the flight distance less than 4000. It is a good indication that these values are outliers. It will therefore be removed from the dataset.

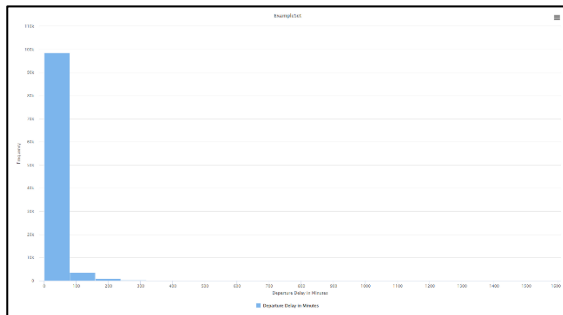


Figure 15 - distribution of departure delay in minutes

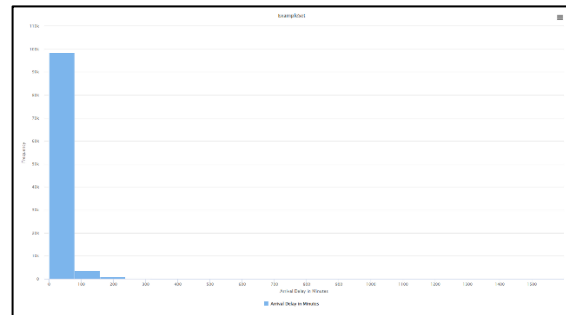


Figure 16 - distribution of arrival delay in minutes

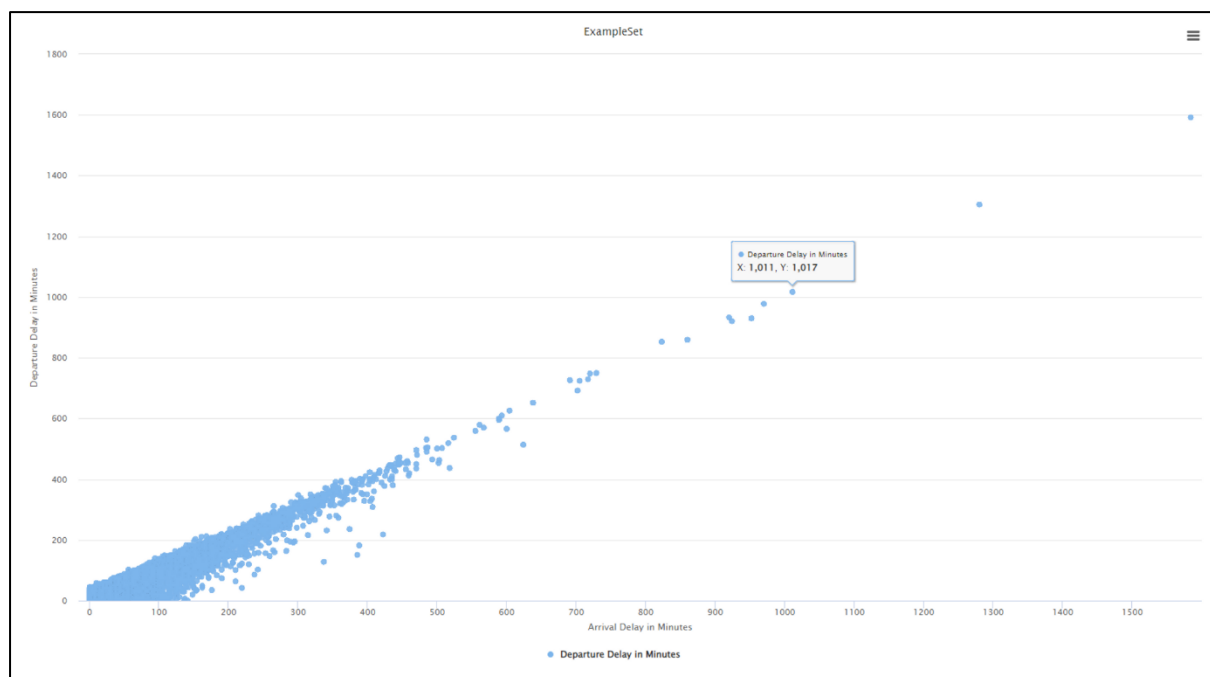


Figure 17 - scatter plot of departure against arrival delay in minutes

The 2 histogram above shows the distribution of both arrival delay and departure delay. Both suggest that most of the flights have less than 100 minutes of delay, with only minuscule number of flights that are delayed for more than that. In the scatter plot, we can see most of the values are concentrated from the range 0 to 550, and the values above 550 in both axis starts to be very scattered away. This suggests that these observations are outliers of the data and will therefore be removed. However, after removing the values more than 550, the distribution is still extremely right skewed. This may be due to the nature of the problem where most flights are minimally delayed, with a few exceptions where the flight will be

delayed for extreme periods. The scatter plot also suggests there is a linear relationship between departure and arrival delay, and it makes sense because when the departure time is delayed, the arrival time will naturally be delayed as well considering the flight time is the same.

The other columns include Inflight wifi service, departure/arrival time convenience, ease of booking online, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on-board service, leg room service, baggage handling, checkin service, inflight service and cleanliness is on a scale from 0 to 5. It will be difficult to determine the outliers in this case as the values are discrete.

### 4.1.2 Visualisation on categorical data

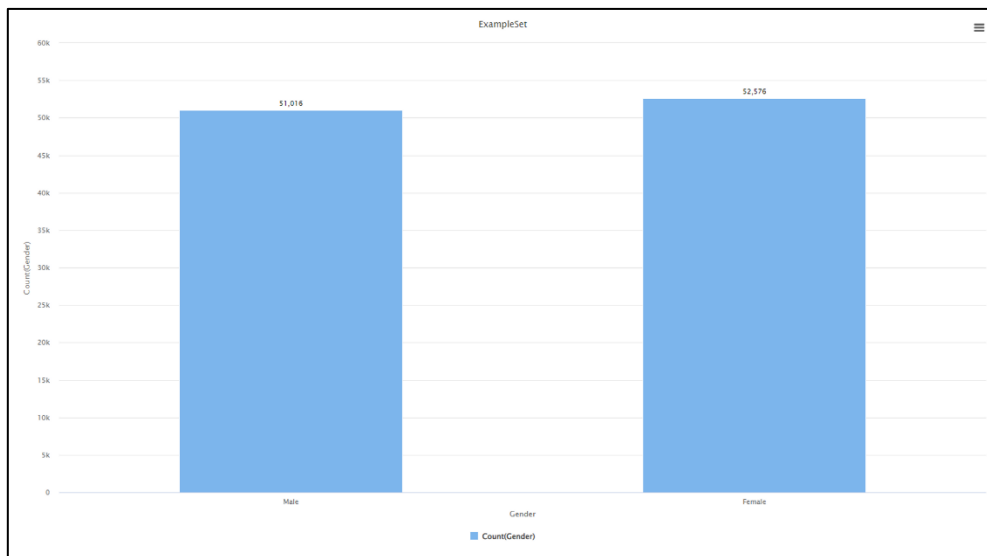


Figure 18 - bar chart of gender

The bar chart shows that this dataset is evenly distributed between male and female customers.

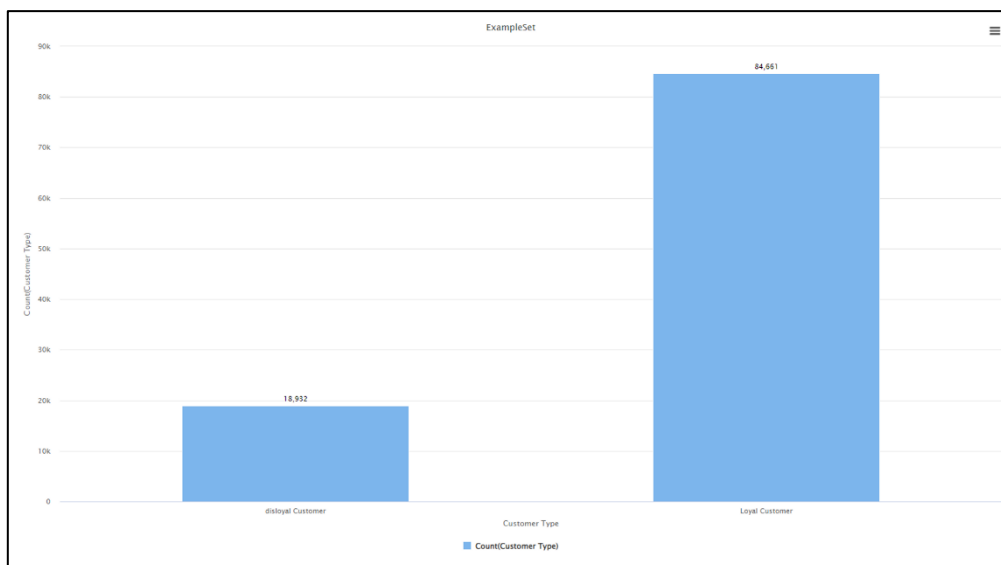


Figure 19 - bar chart of customer types

From the bar chart above, we can see there are proportionally more loyal customers compared to disloyal customers. This may be an indication that the airline company may need to focus on retaining the loyal customers.

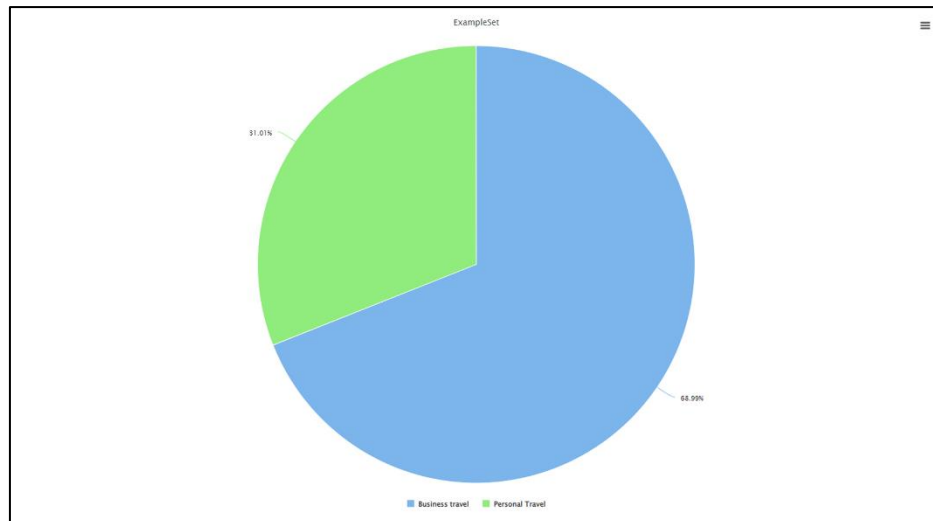


Figure 20 - Pie chart of travel type

The pie chart above shows 68.99% of customers take the flight for business travels. The airline company should pay more attention towards the needs of business travellers.

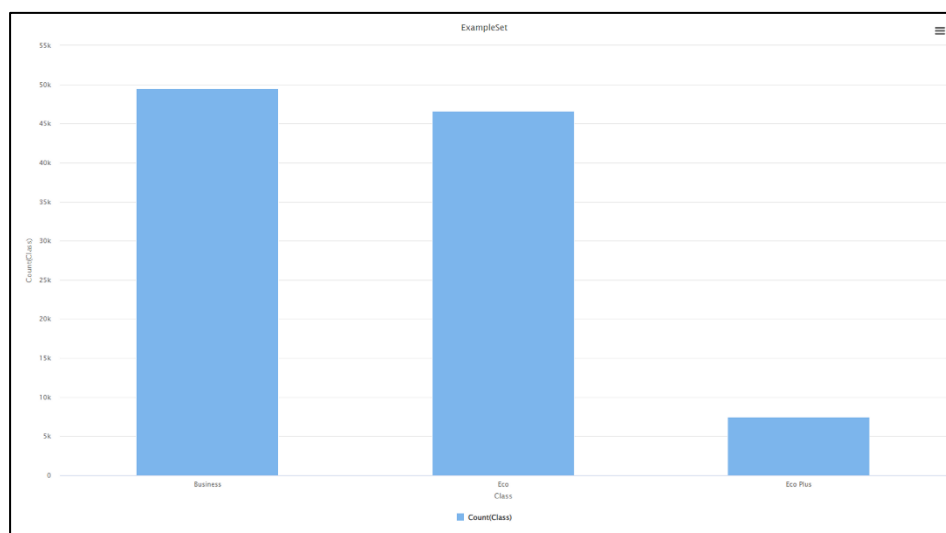


Figure 21 - distribution of each airline class

From the bar chart above, we can see similar distribution between economy class and business class, but the eco plus class is very less compared to the other 2. The airline company should look into reason as to why there is such a difference between the classes in various aspects like price and benefits provided.

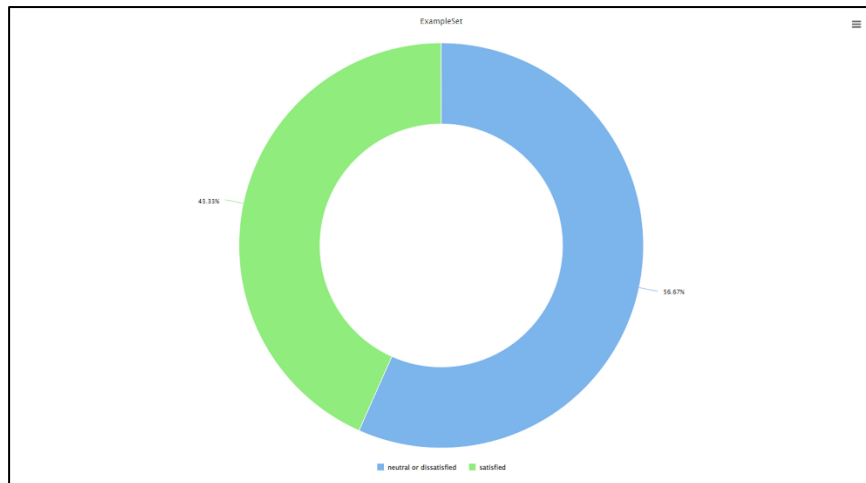


Figure 22 - doughnut chart of satisfaction

Lastly, the donut chart shows the distribution between satisfied and dissatisfied customers are roughly equal, with slightly more dissatisfied customers. This will ensure the model will be able to provide accurate prediction and does not have bias towards one class.

### 4.1.3 Multivariate analysis

#### 4.1.3.1 Who are the most satisfied?

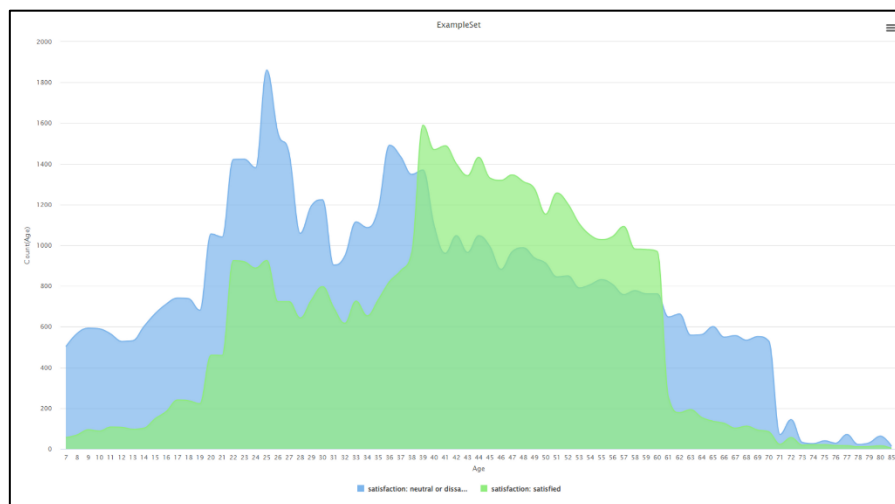


Figure 23 - distribution of age group by satisfaction

It is easy to read that there are 2 spikes for the neutral and dissatisfied group, which are the age range 22 to 27 and 36 to 39. While most satisfied customers are of the age range of 39 to 60.

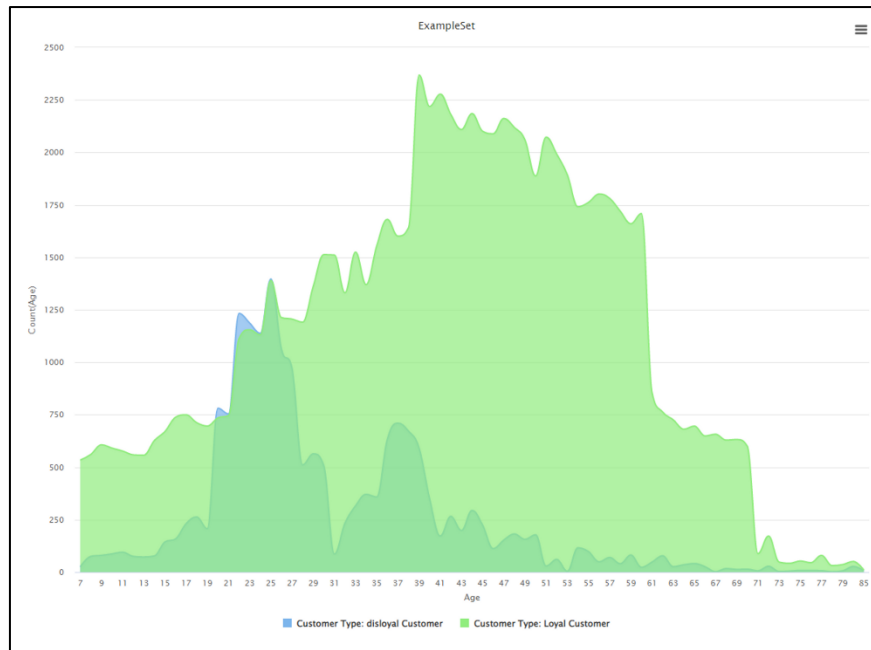


Figure 24 - distribution of age group by customer type

From another graph, we can see that disloyal customers tend to be the customers from the age range 20 to 27, while loyal customers peaks at the age range between 39 to 60. This observation is similar with the satisfaction of the customers. This suggests that dissatisfied customers tend to not be loyal customers, and it is logically so as customers would want to enjoy a satisfying flight experience rather than the opposite.

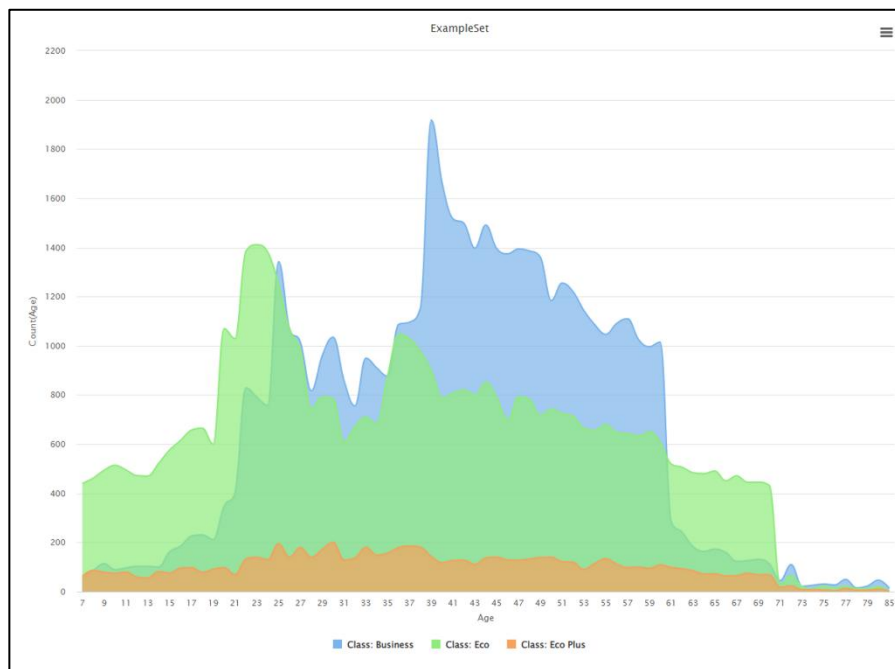


Figure 25 - distribution of age group by flight class

From this graph, we can see that most customers between 20 to 27 take the economy class flight, and a lot of business class customers are between the age 39 and 60. There is a similar pattern between the satisfaction, customer loyalty and flight class. The relationship suggest that customers between age 20 to 27 who take economy class tend to be unsatisfied with the flight experience, and thus will not vouch their loyalty towards the airline company. In contrast, customers between age range 39 to 60 who flies with business class tend to be loyal and satisfied customers.

The airline company should target the customers the age range 20 to 27 who takes the economy class flight, and provide them with extra promotions and amenities that would improve their flight experience and increase their satisfaction. This could include things like offering complimentary meals or drinks, providing more comfortable seating, or offering in-flight entertainment options.

In addition, the airline company should take care of the customers of age range 39 to 60 who takes the business class flight as to retain their loyalty. This can be done by continuing to provide excellent service towards this group of people, and offering special privellages for their loyalty towards the company.

#### **4.1.3.2 What are the inflight factors affecting flight satisfaction?**

There are several metric that is related to the inflight flying experience, which include flight distance, inflight wifi service, food and drink, seat comfort, inflight entertainment, on-board service, leg room service, inflight service and cleanliness.

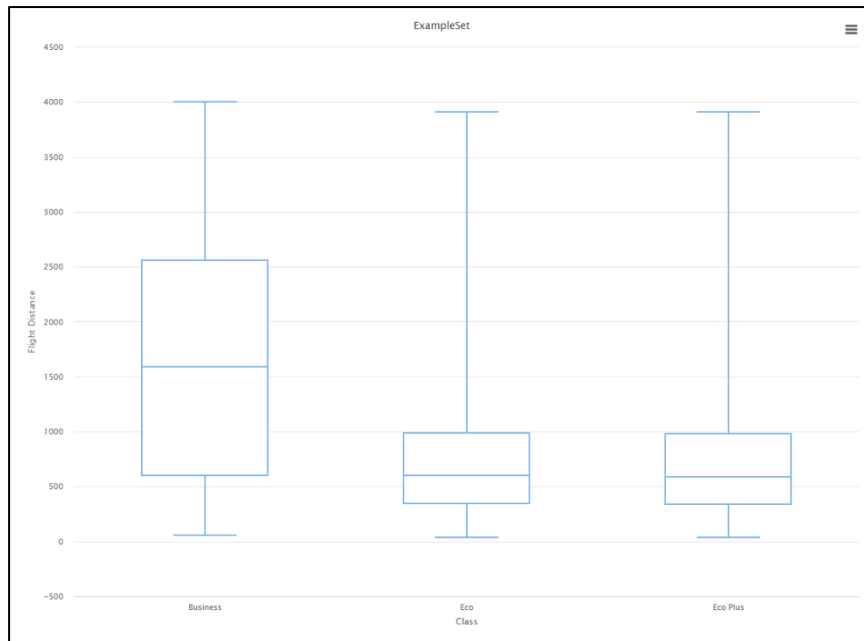


Figure 26 - boxplot distribution of flight distance group by flight class

From this boxplot, we can see that most economy class and economy plus class flyers are travelling shorter distances, while business class flyers have a higher average flight distance.

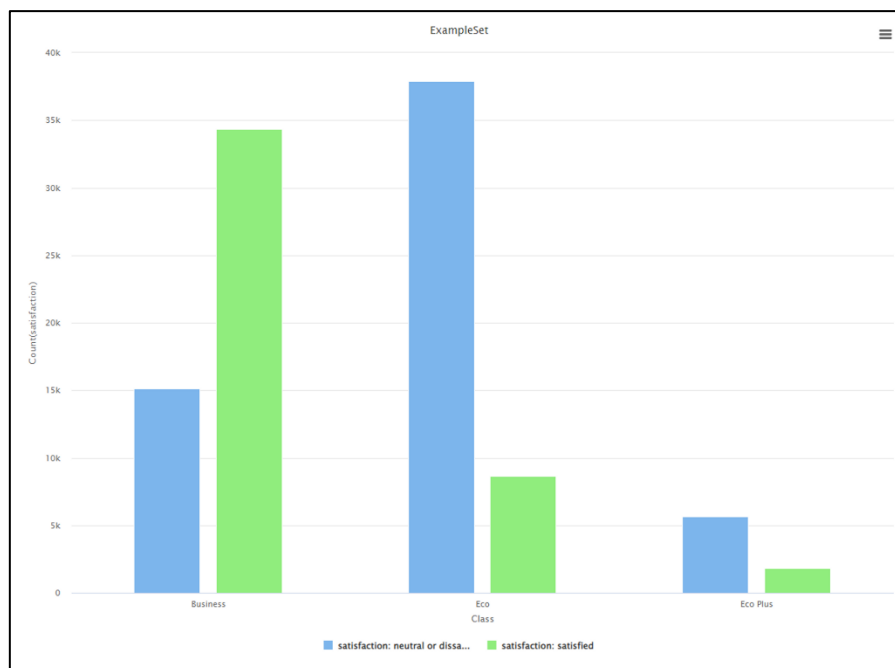
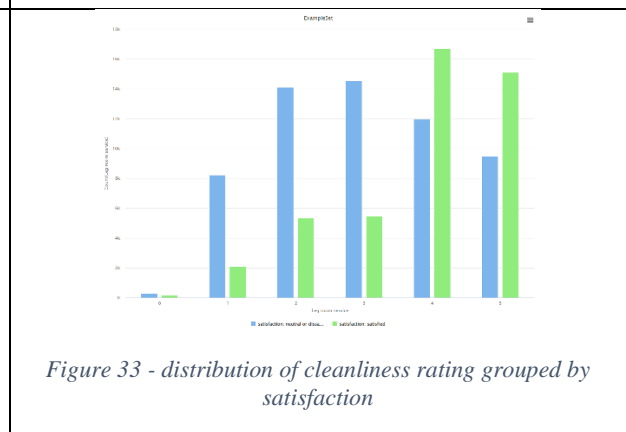
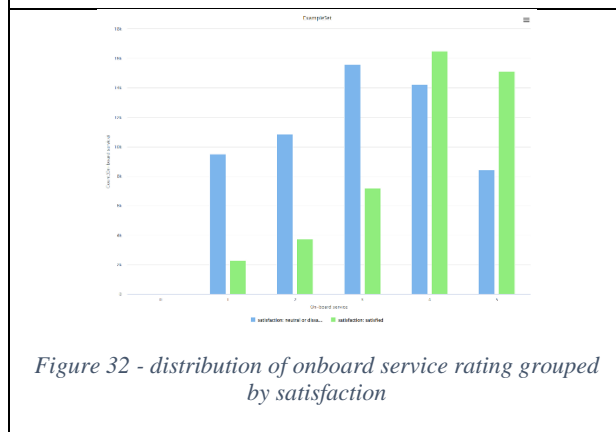
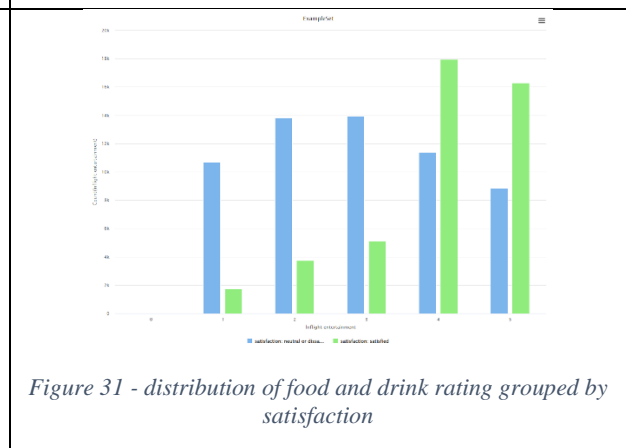
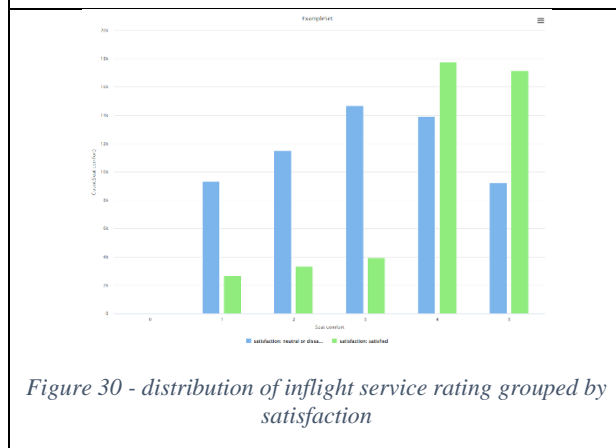
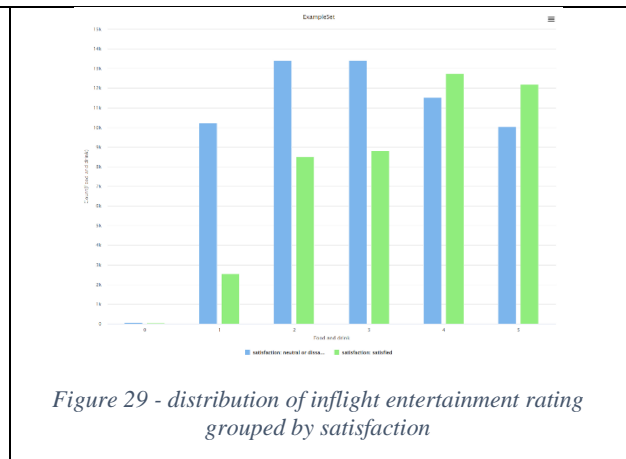
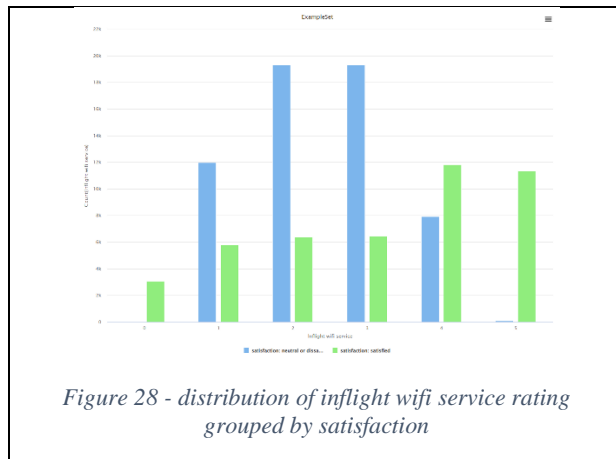


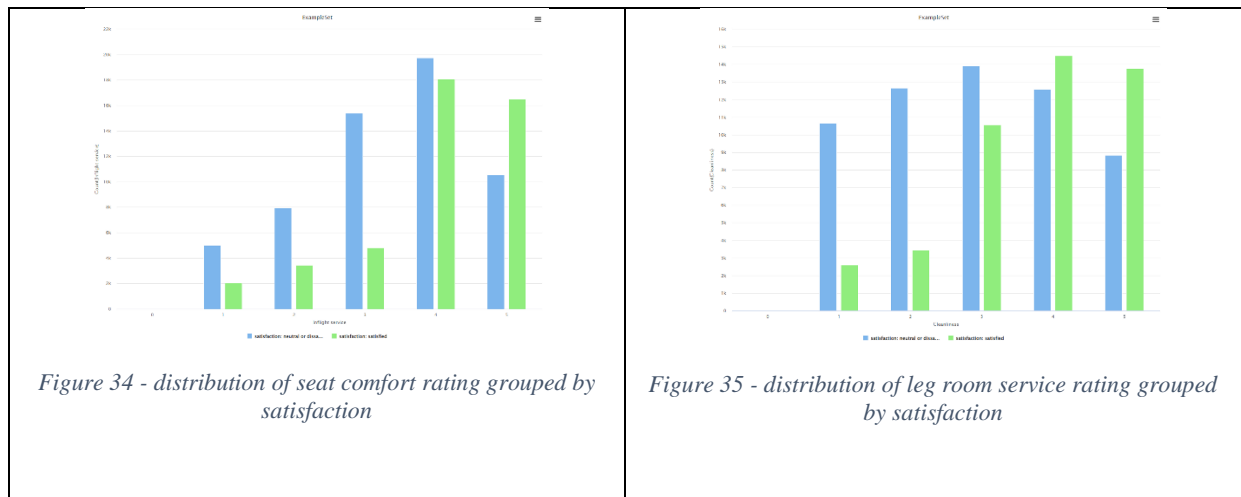
Figure 27 - count of customers in each flight class grouped by satisfaction

In this bar chart, it shows an overwhelming difference between the ratio of satisfied customers versus the dissatisfied customers. Business class has more than double satisfied compared to dissatisfied, while the economy class has more than four times dissatisfied customers compared to satisfied, similar with the economy plus class with more dissatisfied



customers than satisfied customers. Although business class customers may tend to fly longer distance, however they also make up a big portion of the satisfied customers. This show that the difference in service and benefits between the classes are a major factor in customer satisfaction.

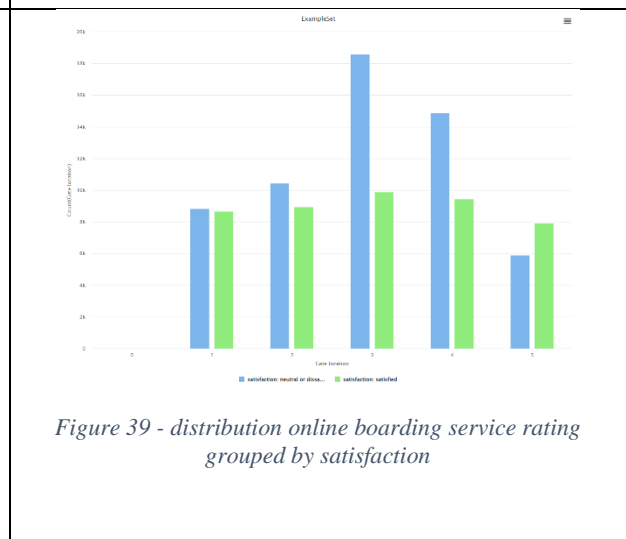
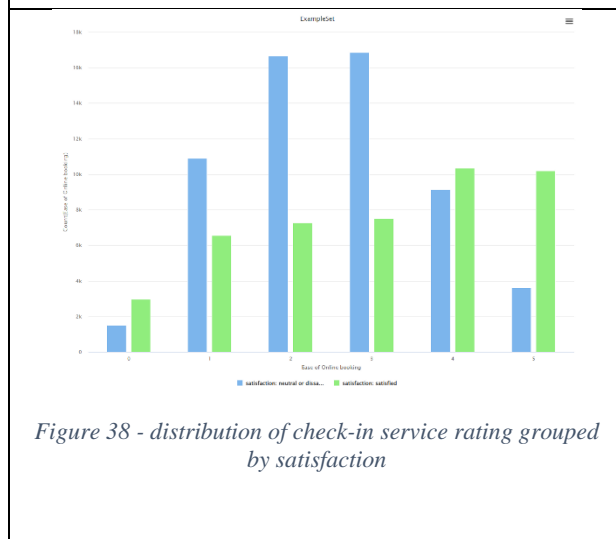
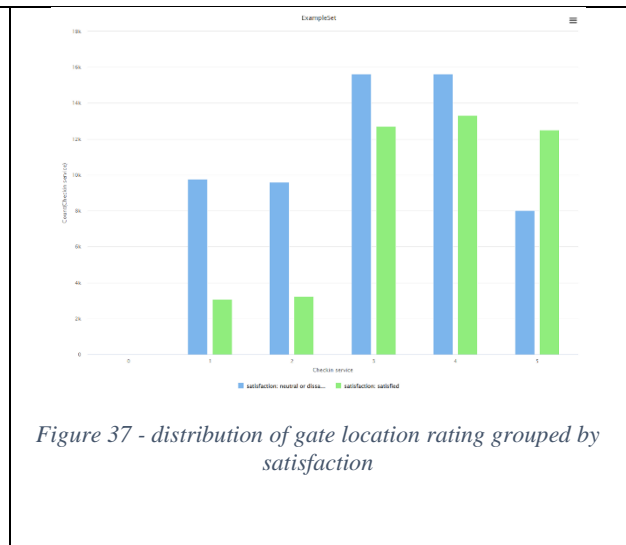
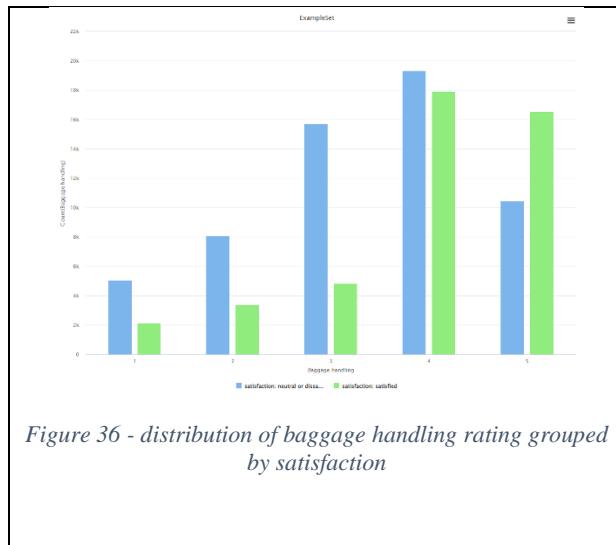




From these bar charts we can see there is a consistent pattern between customer satisfaction with each factors. Factors like wifi, food and beverage, seat comfort, inflight entertainment, onboard service, leg room service and cleanliness all has a similar pattern where customers that rate a score of 1 to 3 are typically dissatisfied, and in contrast, there are more satisfied customers that rate 4 and 5 for these services. This is not a coincidence as these are the factors that directly affect the experience of a customer within a flight. for the inflight services, the customers that rate a score of 1 to 3 are mostly dissatisfied, while customers that rated 5 are mostly satisfied. Customers that rated 4 has about the same proportion with dissatisfaction being slightly more. This may be an indication that the airline company need to put a higher bar on their service standards to satisfy more customers.

This information clearly shows that on-board service is impactful towards the customer satisfaction. The airline company should take this opportunity to focus on improving their on-board service in order to enhance the overall flight experience for their customers. This could involve providing additional training to flight attendants to ensure they are delivering high-quality service, improving the quality of food and beverage offerings, and providing more amenities and options to enhance the customer experience. Enhancing the overall customer service will enable the airline company to achieve a higher satisfaction rate for the customers, which ultimately lead to the increase in customer loyalty and profitability.

### 4.1.3.3 What are the processes that can be improved to increase customer satisfaction?



From all the bar charts above, we can see very clearly that the customers with a good online boarding experience tend to be satisfied, and the opposite is true where customers with a bad

online boarding experience tend to be dissatisfied as shown by the overwhelming amount of dissatisfied customer versus the satisfied customers on the rating score of 1 to 3.

In contrast, the gate location of the flight does not seem to be a contributing factor towards the customer satisfaction, as we can see from the graph. The number of satisfied customers in all 5 rating score for gate location is distributed fairly equally. However, there is a noticeable spike in the dissatisfied customers with score 3 and 4, which may have been caused by other factors.

For baggage handling, there seem to be an upward trend of higher score relating to higher number of satisfied customer. However, the same trend can be observed in the number of dissatisfied customers as well. There may be other details to be considered like baggage wait time and condition of baggage arrived. Therefore, the airline company should look into the details which is related to baggage handling as rating the general experience about the overall baggage handling process does not provide useful insights.

As for Check-in service, customers that rate 1 or 2 is more likely to be dissatisfied than the otherwise as shown in the graph. The same is also true for rating 3 and 4, but the ratio of dissatisfied to satisfied customers are significantly smaller. This may be an indication that the airline company need to improve their overall check-in service to minimize the number of dissatisfied customers. Only with a rating 5, the number of dissatisfied customers reduce, while the number of satisfied customers stays fairly constant.

Lastly for ease of online booking, there seem to be an increase in number of satisfied customers with the higher scores. However, the increase is gradual and not as prominent as to when compared with other factors. The number of dissatisfied customers are also on a decrease with the rating of 4 and 5. This shows that the online booking contributes towards the customer satisfaction. The high number of dissatisfied customers that rated low for this factor implies that the process does not need to be overly perfect, but a bad experience would definitely affect the customer satisfaction negatively.

## 4.2 Data pre-processing

As the data has some missing value and unused variables, the researchers had taken steps to remove these unwanted data.

Firstly, the dataset has columns with label id and att1. From the data dictionary given in the dataset and by observation of the raw data, the researcher had determined that these are just attributes that use to label each unique customer feedback. As it does not contain any useful information that can be mined, both the attributes will be removed.

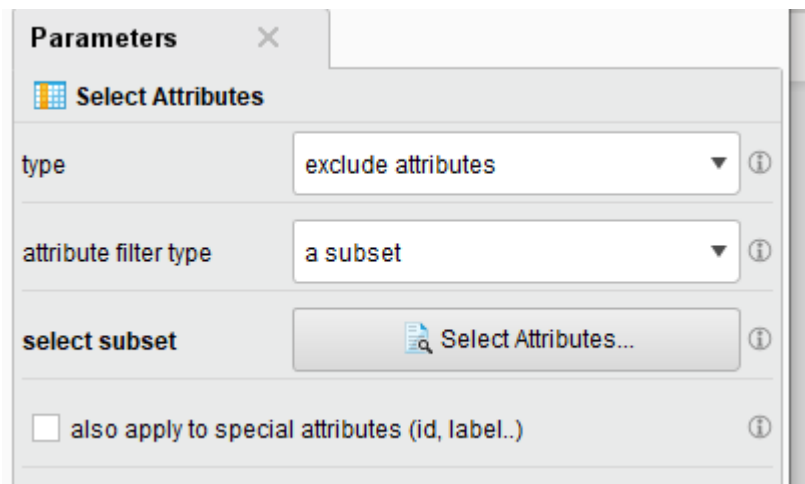


Figure 41 - properties of the Select Attribute block

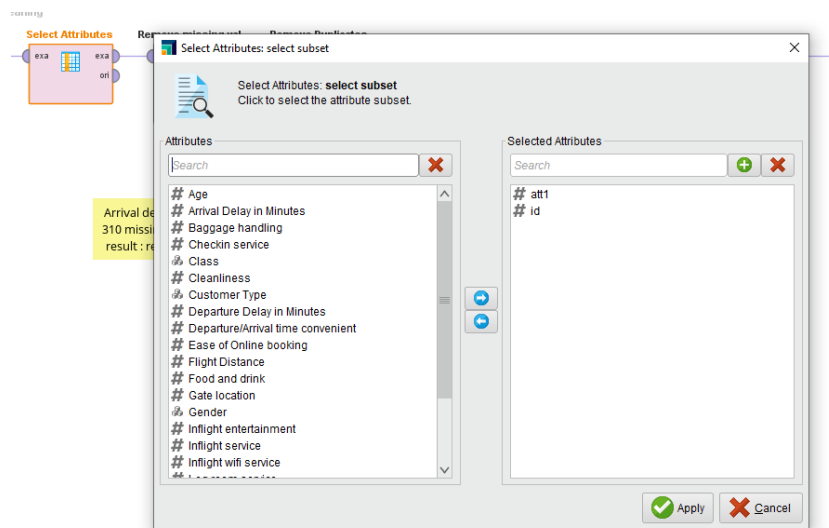


Figure 42 - id and att1 is selected to be excluded

Next, there are some missing values found in the Arrival delay in minutes attribute. There are 310 missing value as shown in Section 4.1. In this case, the missing value will be removed. As there are more than 100,000 observations in this dataset, removing 310 will not make a

significant impact towards the dataset. Filter Example block is used to achieve removal of missing value.

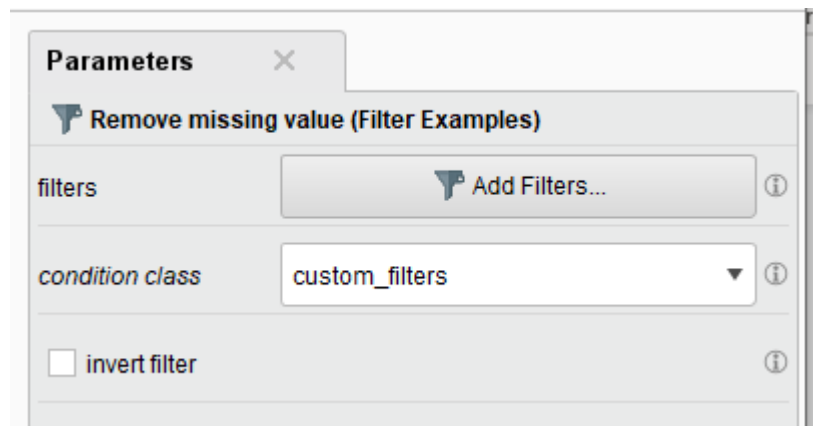


Figure 43 - filter properties

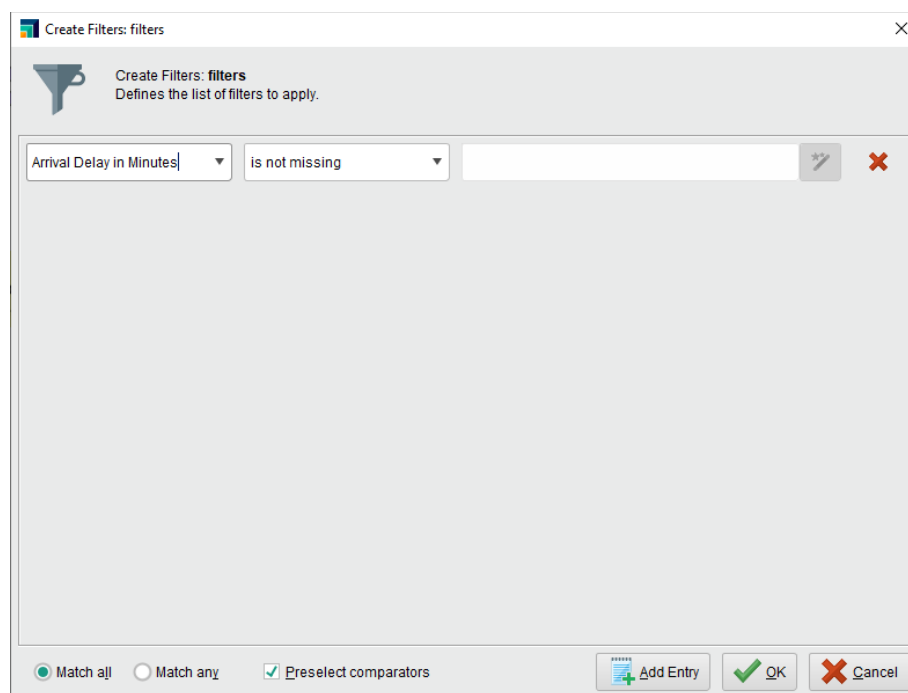


Figure 44 - filter attributes and criterion

Next, duplicated values will be removed. This is to prevent the models from learning any unintended patterns when there exist duplicates of data. The Remove Duplicates block is used.

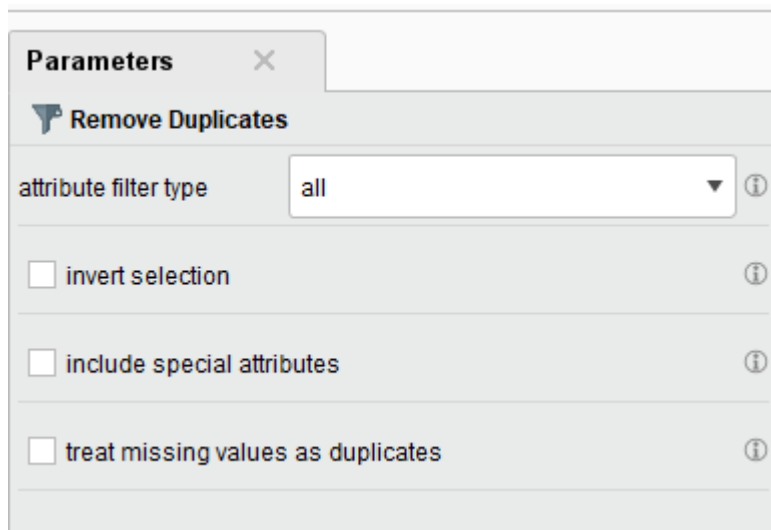


Figure 45 - remove duplicates properties

Overall, the process for data cleaning is as follow:

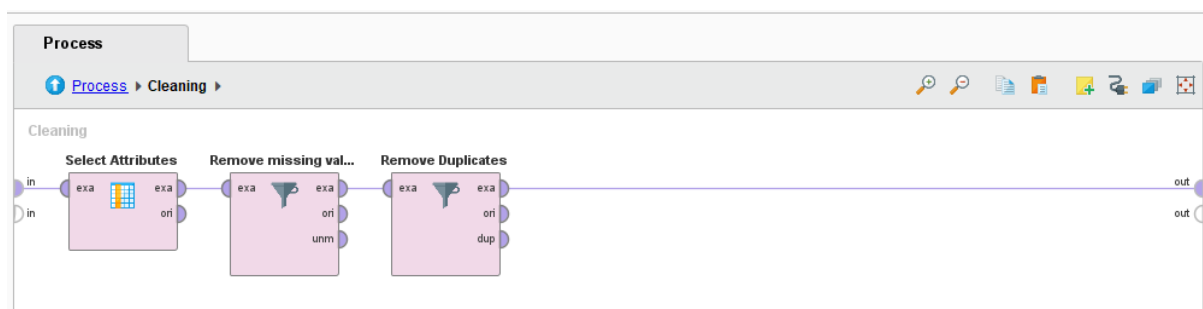


Figure 46 - overall process for data cleaning

## 5.0 Modelling

The airline passenger satisfaction dataset can be used to build various data models to predict passenger satisfaction levels based on their flight experience. The dataset can be utilized for both classification and prediction tasks, where the goal is to predict whether a passenger is satisfied or unsatisfied with their flight experience. Classification algorithms such as decision trees, random forest, and KNN models can be used to classify passenger satisfaction levels based on their demographic details and service ratings. These models can help airlines identify which factors contribute most to passenger satisfaction and prioritize their service improvements accordingly. In addition to classification models, prediction models such as deep learning models can be used to predict the overall satisfaction levels of passengers based on their flight experience. These models can be trained on a large dataset and can provide airlines with insights into how different factors affect overall passenger satisfaction levels. By building and training these models on the airline passenger satisfaction dataset, airlines can gain valuable insights into how to improve their services and enhance passenger satisfaction levels. This can lead to increased customer loyalty and repeat business, ultimately benefiting both the airline and the passengers.

Classification is a machine learning technique used to sort data points into different categories or classes based on their features. The main objective of classification is to develop a model that can precisely predict the class unmarked data points (Kranz, n.d.). To achieve this, classification algorithms use a labeled training dataset to establish the relationship between the data points' features and their class labels. After the model has been trained, it becomes capable of assigning class labels to new data points by utilizing the learned relationships and assigning them to the most appropriate class label with the highest probability. This research study uses four different data mining algorithms, including Deep Learning, Random Forest, KNN, and Decision Tree, on the Airline Passenger Satisfaction dataset to determine which algorithm can provide the most accurate prediction model for Passenger Satisfaction. These data mining algorithms are implemented using RapidMiner, an open-source data mining tool. The purpose of this study is to find the algorithm that can build the best model to predict the satisfaction of airline passengers based on the available data, such as flight details, in-flight services, and demographic information. By comparing the accuracy of these algorithms, we can identify the most effective algorithm to use for this



particular dataset, which can help improve the overall passenger experience in the airline industry.

### **5.1 Model 1: Decision tree - Yan Mun Kye (TP056066)**

Decision tree is a data mining algorithm that employs a hierarchical model to represent decisions and potential outcomes because it is a model that resembles a tree, with nodes and branches, where each node represents a choice and each branch a potential result or effect of that choice. The formation of the tree starts from a root which is the initial decision of the choice, while the decision's potential directions are represented by the branches. It will divide the data into smaller subsets based on the values of specific features in a recursive manner until a stopping requirement is met. The splitting process will always select the best feature to split the data based on a measure of information gain, which measures the reduction in entropy or impurity of the data after the split. Eventually, the objective is to construct a tree that accurately depicts the connections between the features and the target variable. In machine learning, decision trees are frequently used for classification and regression tasks because they can handle both categorical and numerical data and the tree is simple to interpret. However, decision trees can be prone to overfitting, where the model is too complex and fits the training data too closely, leading to poor generalization on new data (IBM, n.d.).

From the result, decision tree shows that is easy to understand, interpret and visualize because the structure of the tree provides a visual representation of the decision-making process, making it easy for non-experts to understand how the model arrived at a particular decision. Decision trees can be used to generate if-then rules that can be easily understood and communicated. This makes decision trees a popular choice in domains where the interpretability of the model is important. Besides that, since it can handle both numerical and categorical data, it makes decision tree versatile and useful in a wide range of applications. Unlike other algorithms that require the data to be pre-processed or transformed before they can be used, decision trees can work with raw data in its original form. This cause decision tree needs less time and effort compared to other algorithm and can lead to more accurate results since the data is not distorted by pre-processing steps. Not to mention, decision trees can also handle missing values and outliers, which is useful when dealing with real-world data that is often incomplete or noisy (AnalytixLabs, 2022).

However, decision trees are prone to overfitting, which means that they can capture noise in the data and produce a model that fits the training data too closely, resulting in poor performance on new data. Overfitting can occur when the tree is too complete, too many branches or the splitting criterion is too strict which can be prevented by setting a minimum number of instances per leaf or using regularization methods but these techniques can also result in underfitting where the model will be too simple and fails to capture the underlying relationships in the data. Besides that, Decision trees is sensitive to minor changes in the data or the training process which can lead to instability and inconsistency in the model, it means that the tree structure can vary greatly depending on the specific subset of data used for training and make it difficult to compare and interpret the results. Additionally, decision tree is deterministic where it will always produce the same result given the same input data, it will be a disadvantage when dealing with noisy or uncertain data (AnalytixLabs, 2022). When used in the context of an airline passenger satisfaction dataset, a decision tree can provide numerous benefits.

First and foremost, it can help in determining the most critical factors that influence passenger satisfaction. By analyzing the structure of the decision tree and the features utilized to divide the data, valuable insights can be gained into which factors have the most significant impact on passenger satisfaction.

Besides, a decision tree can be used to forecast passenger satisfaction levels based on input features. This can be particularly advantageous for airlines that are interested in enhancing the passenger experience by concentrating on the factors that have the most substantial effect on satisfaction.

Furthermore, a decision tree can be utilized to identify different passenger segments with varying levels of satisfaction. This can be useful for airlines that aim to tailor their services to various passenger segments to boost overall satisfaction levels.

In conclusion, decision trees are potent tools that can be employed to analyze and predict passenger satisfaction in the airline industry. They offer valuable insights into the most critical factors that affect satisfaction, facilitate the customization of services, and improve the overall passenger experience.

The given decision tree shows the hierarchy of factors that contribute to passenger satisfaction levels based on the various service ratings provided in the airline passenger satisfaction dataset.

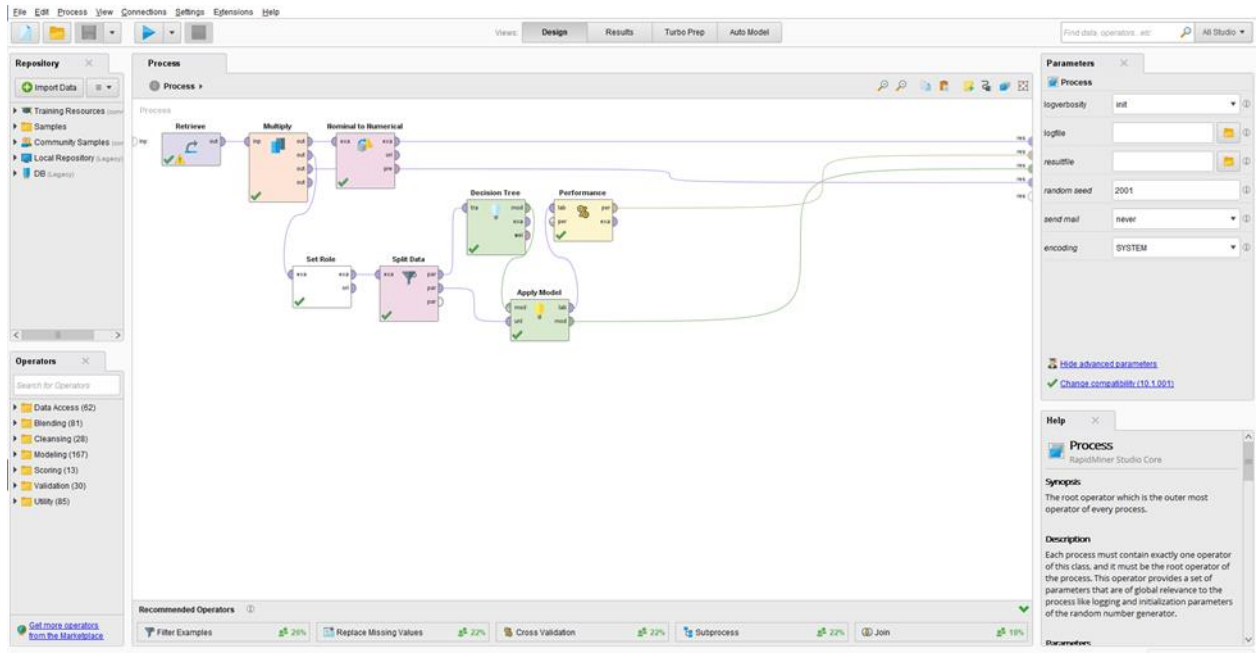


Figure 47 - Design of Decision Tree

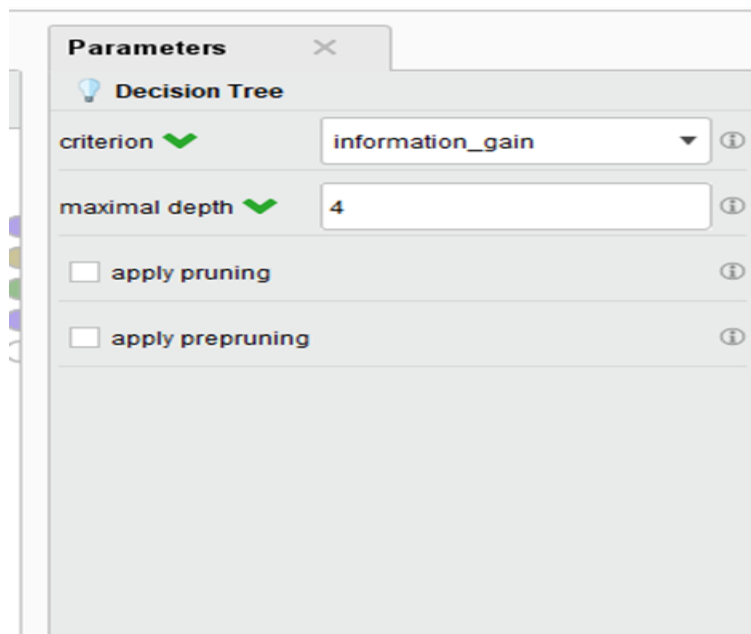
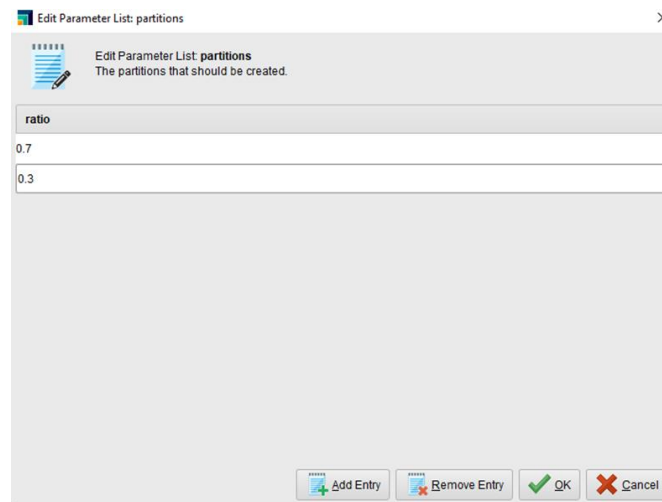


Figure 48 - Parameters Set for Decision Tree



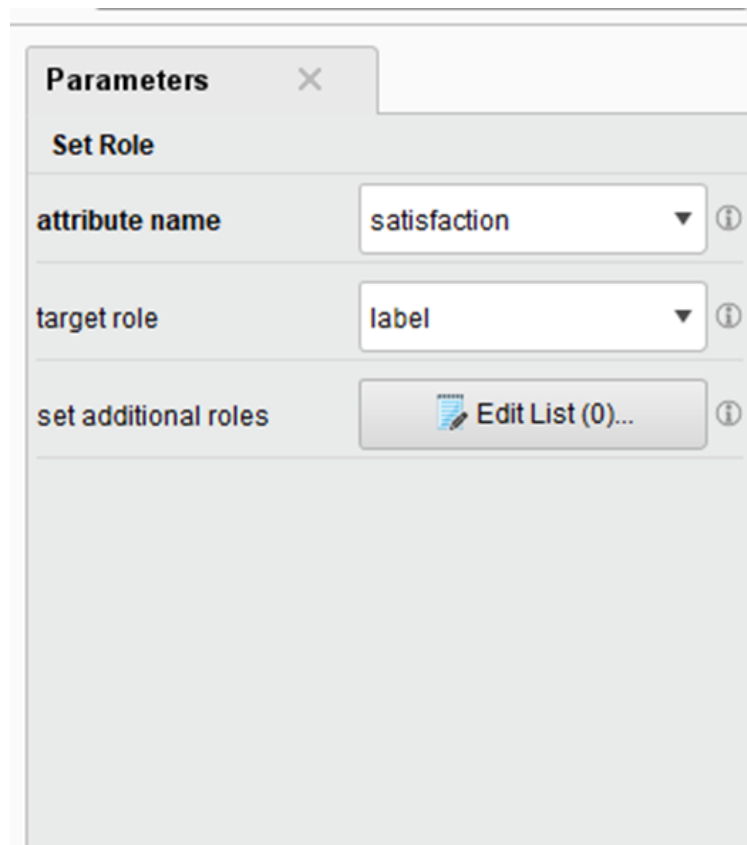
Edit Parameter List: partitions

Edit Parameter List partitions  
The partitions that should be created.

ratio
0.7
0.3

Add Entry Remove Entry OK Cancel

Figure 49 - Splitting Data of Decision Tree



Parameters

Set Role

attribute name satisfaction

target role label

set additional roles Edit List (0)...

Figure 50 - Target Variable

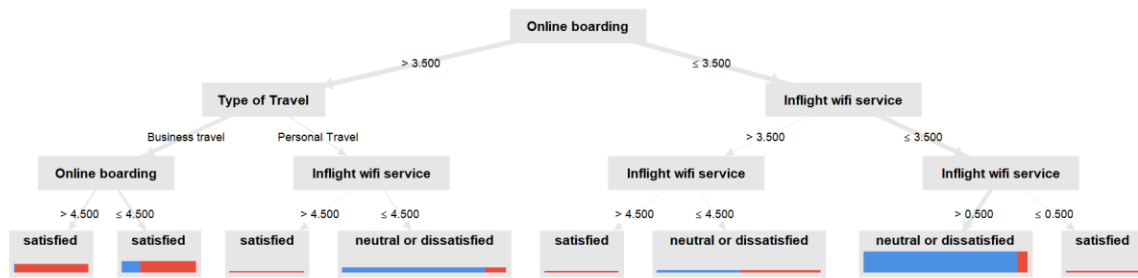



Figure 51 - Decision tree Graph

The top node in the decision tree shows that the online boarding rating is the most important factor in predicting passenger satisfaction levels. If the online boarding rating is greater than 3.5, the decision tree branches out into two paths based on the type of travel.

For business travelers with an online boarding rating greater than 4.5, the decision tree predicts high satisfaction levels. For business travellers with an online boarding rating less than or equal to 4.5, the decision tree predicts neutral or dissatisfied satisfaction levels. For personal travelers with an inflight wifi service rating greater than 4.5, the decision tree predicts high satisfaction levels. For personal travelers with an inflight wifi service rating less than or equal to 4.5, the decision tree predicts neutral or dissatisfied satisfaction levels (Baswardono et al., 2019).

If the online boarding rating is less than or equal to 3.5, the decision tree branches out into two paths based on the inflight wifi service rating. For passengers with an inflight wifi service rating greater than 3.5, the decision tree predicts High satisfaction levels if the inflight wifi service rating is greater than 4.5 else the decision tree will predict it as neutral or dissatisfied satisfaction levels if the inflight wifi service rating is less than or equal to 4.5. For passengers with an inflight wifi service rating less than or equal to 3.5, the decision tree predicts either High satisfaction levels if the inflight wifi service rating is less than or equal to 0.5 else the decision tree will predict it as neutral or dissatisfied satisfaction levels if the inflight wifi service rating is more than 0.5.

Overall, the decision tree provides insights into which service ratings are most important in predicting passenger satisfaction levels, and how these ratings can be used to improve overall passenger satisfaction.



	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	15754	1738	90.06%
pred. satisfied	1855	11731	86.35%
class recall	89.47%	87.10%	

accuracy: 88.44%

*Figure 52 - Decision tree performance metrics*

The decision tree model was applied to the airline passenger satisfaction dataset, and the model achieved an accuracy of 88.44%.

The confusion matrix shows that out of the 31078 instances, the model correctly predicted 15754 of the neutral or dissatisfied passengers and 11731 of the satisfied passengers. However, the model misclassified 1855 satisfied passengers as neutral or dissatisfied and 1738 neutral or dissatisfied passengers as satisfied.

Based on the confusion matrix, we can calculate other performance metrics such as precision, recall, and F1 score. Precision is the proportion of true positive predictions among all positive predictions, while recall is the proportion of true positive predictions among all actual positive instances. The F1 score is the harmonic mean of precision and recall.

For the given confusion matrix, the precision for predicting satisfied instances is 86.35%, while the precision for predicting neutral or dissatisfied instances is 90.06%. The recall for predicting satisfied instances is 87.10%, while the recall for predicting neutral or dissatisfied instances is 89.47%. The F1 score is 86.72% for predicting satisfied instances and 89.76% for predicting neutral or dissatisfied instances.

Overall, the decision tree model performed well in predicting passenger satisfaction levels, with high accuracy, precision, and recall.

## 5.2 Model 2: Random forest - Tan Sheng Jeh (TP056267)

Random Forest are applied to classification and regression issues which uses an ensemble learning technique to merge various decision trees into a reliable and precise model. In a random forest, numerous decision trees are constructed and trained using various subsets of the training data that were randomly selected with replacement. Each tree is trained on a slightly different subset of data throughout this procedure, which is referred to as bagging, and it helps to lower the variance of the model. The random forest builds each decision tree by iteratively dividing the data into smaller subsets based on the most important traits. In regression problems, the divides are chosen to minimise the mean squared error or maximise the separation between the classes. After all the decision trees have been constructed, the model forecasts the class or value either using the results of the various trees individually or by using the vote of the majority. This lowers the possibility of overfitting and enhances the model's generalisation capabilities (Yiu, 2019).

Random Forest are known for its high accuracy and robustness in predicting outcomes, even in complex datasets with a large number of features. By combining multiple decision trees, random forest can capture complex interactions between features and provide accurate predictions while the random selection of subsets of data for each tree helps to reduce the risk of overfitting and improve the generalization performance of the model. Random Forest can handle large datasets with high dimensionality and heterogeneous feature types, such as numerical, categorical, or text data. It can also deal with missing values and outliers in the data, by using the majority vote or the average of the trees to make predictions. Besides that, Random Forest can also provide a measure of the importance of each feature in the dataset. By analysing the feature importance scores of the model, the team can identify the most significant factors that contribute to the outcome of interest. This feature selection technique can help to reduce the dimensionality of the dataset, improve the model's performance, and gain insights into the underlying processes that drive the outcome (Jagandeep, 2020).

However, Random Forest can be computationally intensive, especially when dealing with large datasets or a large number of features since building multiple decision trees and selecting the best features for each tree requires significant computational resources and can be time-consuming. The complexity of the model can make it challenging to interpret the results and gain insights into the underlying processes that drive the outcome. It is not easy to understand how the ensemble makes predictions, or how each feature contributes to the prediction. The decision trees in the random forest can be very complex and deep which

makes them hard to visualize and analyse and it does not provide a clear explanation for why a certain prediction is made, or how confident it is about the prediction. Besides that, although Random Forest models are useful at predict outcomes within the range of the training data accurately. Nonetheless, it not provides accurate predictions outside the range of the training data. This limitation can be a problem in some applications where the model needs to make predictions in uncharted territory, such as predicting the behaviour of a system under new conditions (Jagandeep, 2020).

When applied to an airline passenger satisfaction dataset, a random forest can provide numerous benefits. Firstly, it can help identify the most significant features that impact passenger satisfaction. By analyzing the feature importance evaluated by the random forest, valuable insights can be gained into which factors have the most substantial impact on passenger satisfaction.

Other than that, a random forest can be used to predict passenger satisfaction levels based on the input features. This can be especially beneficial for airlines seeking to enhance the passenger experience by focusing on the factors that have the greatest influence on satisfaction.

Apart from that, a random forest can be employed to categorize passenger segments with different satisfaction levels. This can be useful for airlines that aim to customize their services to various passenger segments to enhance overall satisfaction.

Moreover, random forests are recognized for their ability to handle incomplete data and noisy datasets, making them a robust option for real-world datasets such as airline passenger satisfaction data.

In conclusion, a random forest can be an effective tool for examining and forecasting passenger satisfaction in the airline industry. It provides insights into the most critical factors that impact satisfaction, facilitates the customization of services, and enhances the overall passenger experience.



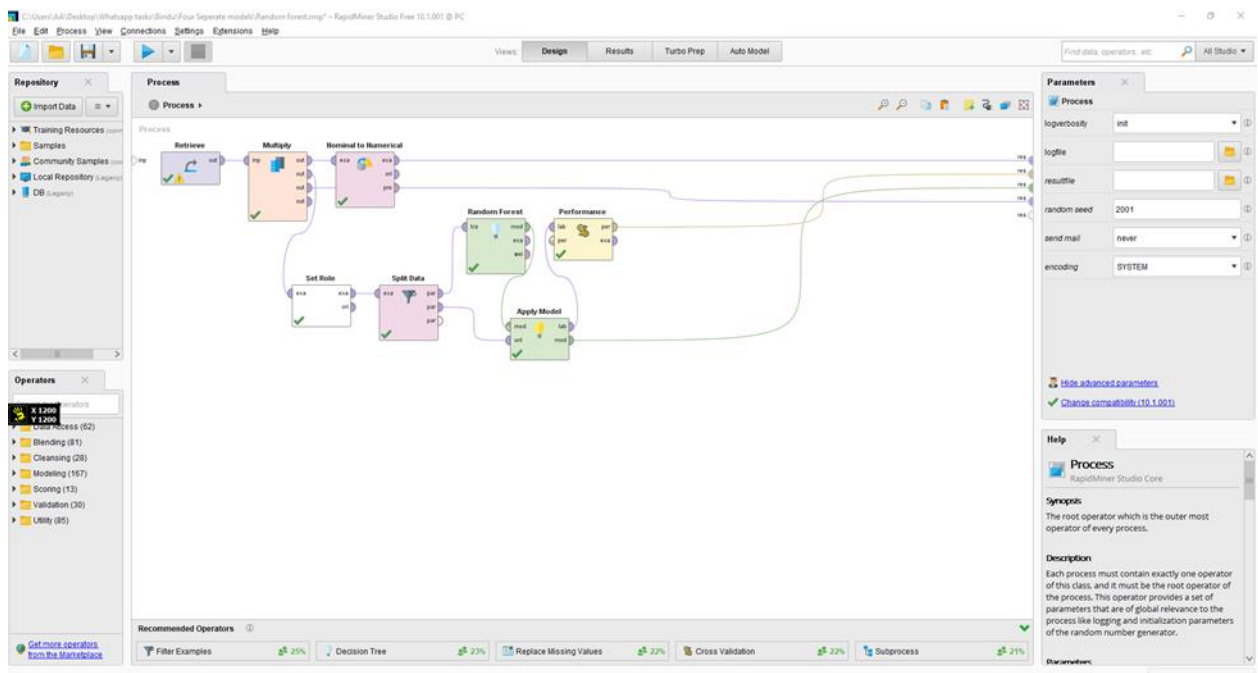


Figure 53 - Random Forest Model

The screenshot shows the 'Edit Parameter List: partitions' dialog box. It contains a text area with the title 'Edit Parameter List: partitions' and the instruction 'The partitions that should be created.' Below this, there is a list of partitions with the header 'ratio'. The list contains two entries: '0.7' and '0.3'. At the bottom of the dialog, there are four buttons: 'Add Entry', 'Remove Entry', 'OK', and 'Cancel'.

Figure 54 - Splitting Dataset of Random Forest

**Parameters**

**Random Forest**

number of trees ✓ 3

criterion ✓ gain\_ratio

maximal depth ✓ 4

☐ apply pruning ✓

☐ apply prepruning ✓

☐ random splits

☒ guess subset ratio

voting strategy confidence vote

☐ use local random seed

[Hide advanced parameters](#)

[Change compatibility \(10.1.001\)](#)

Figure 55 - Random Forest Parameters Setting

**Parameters**

**Performance (Performance (Classification))**

main criterion first

☒ accuracy

☐ classification error ✓

☐ kappa

☐ weighted mean recall

☐ weighted mean precision

☐ spearman rho

☐ kendall tau

☐ absolute error

☐ relative error

[Hide advanced parameters](#)

Figure 56 - Performance Classification

Based on the given information, a random forest algorithm has been used to build a model for predicting passenger satisfaction levels using the airline passenger satisfaction dataset. The model has achieved an accuracy of 82.45%.

To evaluate the performance of the model further, the confusion matrix can be analyzed (Baswardono et al., 2019). The confusion matrix provides information on the true positive, true negative, false positive, and false negative predictions made by the model. In this case, the confusion matrix indicates that:

- There were 14702 true negative predictions, meaning that the model correctly predicted that 14702 passengers were neutral or dissatisfied.
- There were 10923 true positive predictions, meaning that the model correctly predicted that 10923 passengers were satisfied.
- There were 2546 false positive predictions, meaning that the model incorrectly predicted that 2546 passengers were satisfied when they were actually neutral or dissatisfied.
- There were 2907 false negative predictions, meaning that the model incorrectly predicted that 2907 passengers were neutral or dissatisfied when they were actually satisfied.

Overall, the random forest algorithm has shown promise for predicting passenger satisfaction levels in the airline industry. However, further optimization and tuning may be necessary to achieve higher accuracy and better performance.

	true neutral or dissatisfied	true satisfied	class precision
accuracy: 82.45%			
pred. neutral or dissatisfied	14702	2546	85.24%
pred. satisfied	2907	10923	78.98%
class recall	83.43%	81.10%	

*Figure 57 - Performance Metrics for Random Forest*

### 5.3 Model 3: Deep learning - Hor Shen Hau (TP061524)

Deep learning for data mining is the process of analysing and deriving insights from massive, complicated datasets using artificial neural networks with numerous layers. In data mining, deep learning aims to automatically build hierarchical representations of the input data that may be used to spot patterns, forecast the future, and uncover new information. Several layers of artificial neurons are commonly used in deep learning models for data mining where each layer learning increasingly complicated features from the data. With the use of optimisation techniques that modify the weights of the neurons to reduce the discrepancy between the expected and actual outputs, these models are trained on enormous volumes of labelled data. Due to its capacity to handle enormous, unstructured information, such as photos and text, as well as its capacity to discover intricate patterns and relationships within the data, deep learning has grown in popularity in recent years. As a result, it is now an effective tool for many different applications, such as fraud detection, consumer segmentation, and predictive maintenance (Brownlee, 2019).

From the result deep learning shows one of its significant advantages which is its ability to learn complex representations of data automatically. This means that it can extract meaningful features from raw input data, such as images or text, without requiring manual feature engineering by humans. This ability to learn hierarchical representations allows deep learning models to achieve state-of-the-art performance on a wide range of tasks, including image recognition, speech recognition, and natural language processing. Besides that, deep learning models can be trained on massive amounts of data using distributed computing, allowing them to learn from millions of examples in a relatively short amount of time. Additionally, Deep learning algorithms can learn from large amounts of data and improve their performance over time. They can also adapt to changing environments and data distributions, and generalize well to new situations. Deep learning can capture complex and nonlinear relationships in the data, and solve problems that are difficult for traditional machine learning methods (Geeks For Geeks, n.d.).

However, deep learning requires large amounts of data to be trained effectively. This can be a challenge for smaller organizations or those working with specialized or niche datasets, it is not always be available or easy to obtain for high-quality and relevant data. Without sufficient data, deep learning models will memorize the training data rather than generalizing to new examples, it can lead to poor performance on new data and limit the model's usefulness. Besides that, it can be challenging to understand how the model arrives at

its predictions or decisions. This is especially true for deep neural networks with many layers, where the internal workings of the model can be complex and difficult to interpret. As a result, deep learning models may not be suitable for applications where interpretability and transparency are essential, such as in the legal or healthcare industries (Geeks For Geeks, n.d.).

With deep learning algorithms, it becomes possible to learn and extract meaningful patterns and features from large and complex datasets that would be challenging to analyze using traditional machine learning algorithms (Stancin & Jovic, 2019).

When it comes to an airline passenger satisfaction dataset, deep learning can be advantageous in multiple ways. Firstly, it can help to uncover intricate patterns and relationships between input features and passenger satisfaction. This can be valuable for airlines looking to gain a deeper understanding of how various factors influence satisfaction and devise strategies to improve overall satisfaction levels.

Besides, deep learning can be utilized to make more accurate predictions about passenger satisfaction based on input features. By training a deep learning model on a vast dataset of passenger satisfaction data, the model can learn to make more precise and accurate predictions about satisfaction levels based on diverse input factors.

Also, deep learning can be employed to develop customized recommendations and services for different passenger segments based on their unique preferences and behaviors. By training a deep learning model on large and diverse passenger data, airlines can identify patterns and create tailored services for various passenger segments to boost their overall satisfaction levels.

In summary, deep learning can be a powerful tool for examining and predicting passenger satisfaction in the airline industry. It offers insights into complex patterns and relationships between input features and satisfaction, enables more precise predictions of satisfaction levels, and facilitates the development of tailored services to enhance the overall passenger experience.

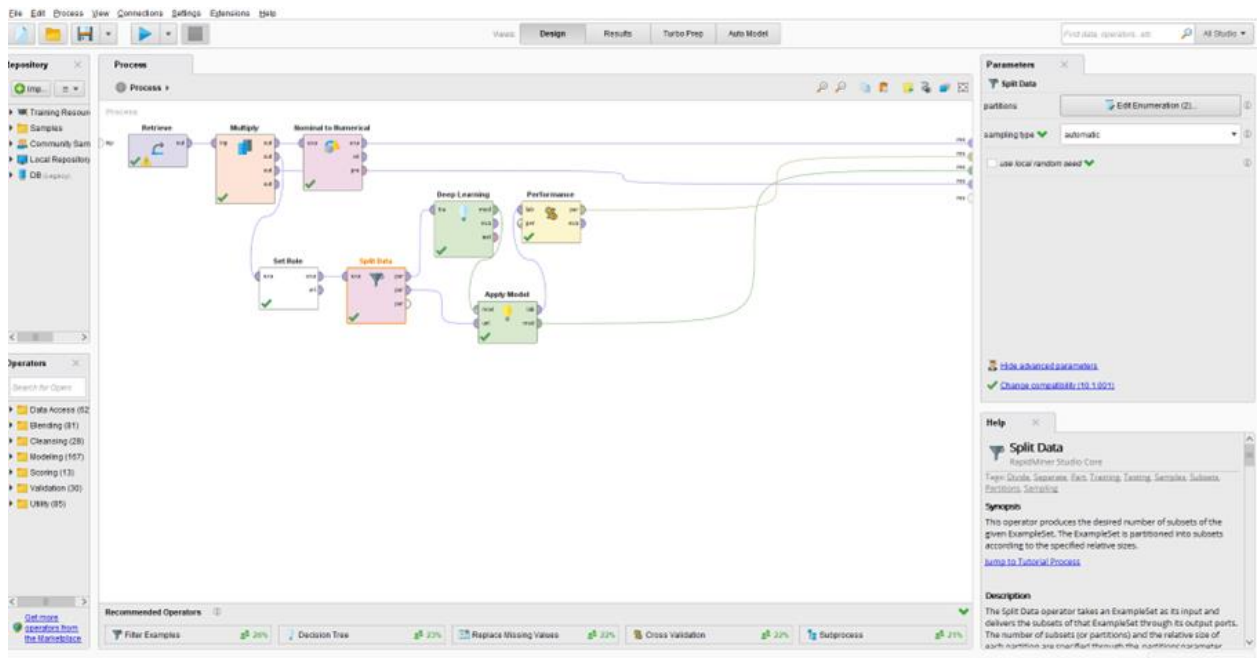


Figure 58 - Deep learning model

**Parameters**

**Deep Learning**

activation ☒ **Rectifier**

hidden layer sizes [Edit Enumeration \(2\)...](#)

☐ reproducible (uses 1 thread) ☒

epochs

☐ compute variable importances

train samples per iteration

☒ adaptive rate

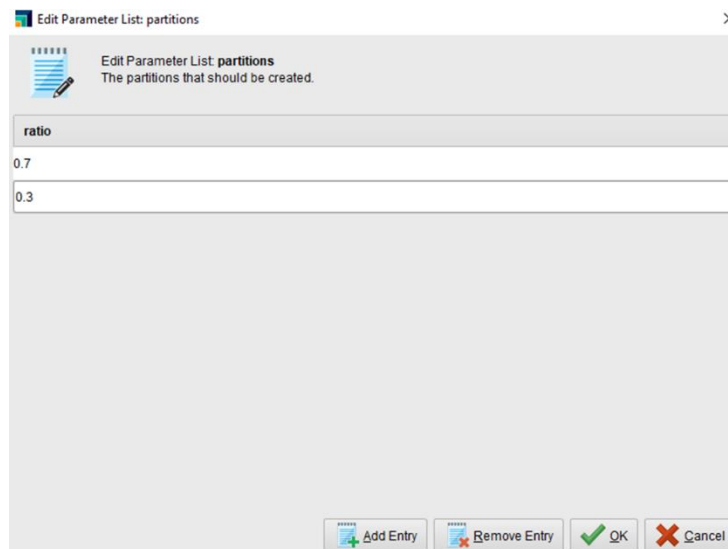
epsilon

rho

[Hide advanced parameters](#)

☒ [Change compatibility \(10.0.000\)](#)

Figure 59 - Deep Learning Model Parameters



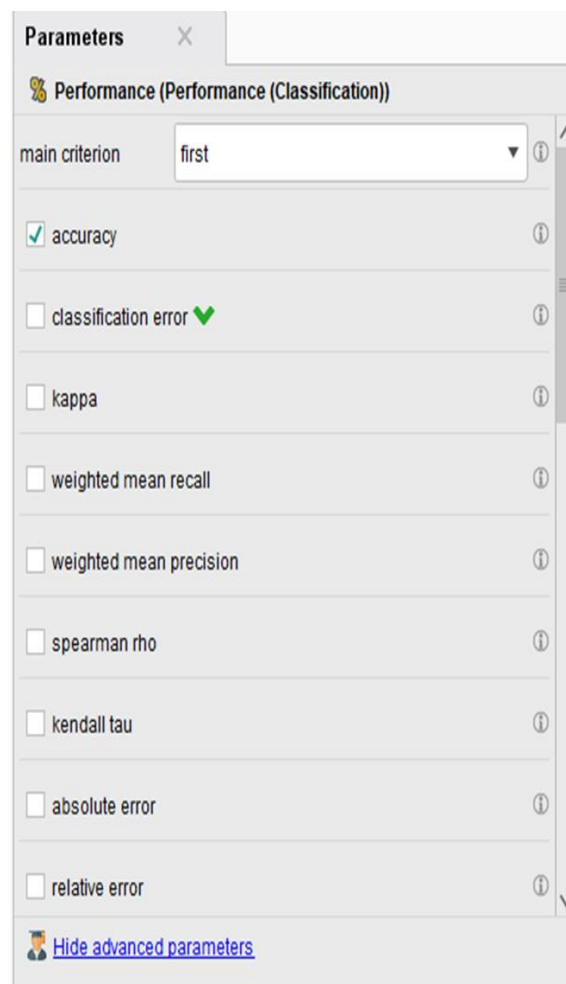
Edit Parameter List: partitions

Edit Parameter List: partitions  
The partitions that should be created.

ratio
0.7
0.3

Add Entry Remove Entry OK Cancel

Figure 60 - Splitting Data for Deep Learning



Parameters

Performance (Performance (Classification))

main criterion first

☒ accuracy

☐ classification error ✓

☐ kappa

☐ weighted mean recall

☐ weighted mean precision

☐ spearman rho

☐ kendall tau

☐ absolute error

☐ relative error

[Hide advanced parameters](#)

Figure 61 - Performance Classification

☒ Table View
 ☐ Plot View

accuracy: 95.50%

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	17012	800	95.51%
pred. satisfied	597	12669	95.50%
class recall	96.61%	94.06%	

*Figure 62 - Performance metrics of Deep Learning*

Based on the result, the accuracy of Deep Learning model is around 95.50% and the deep learning model has achieved an AUC (Area Under the Curve) value of 0.9928912, indicating that the model has high predictive accuracy. The model's mean\_per\_class\_error is 0.03150225, indicating that the model misclassified approximately 3.2% of the samples.

The confusion matrix reveals that the model correctly predicted 5549 of the neutral or dissatisfied passengers out of 5721 (a 0.0301 error rate) and correctly predicted 4091 of the satisfied passengers out of 4364 (a 0.0626 error rate). The total number of misclassified samples was 445 out of 10085, resulting in an overall error rate of 0.0441.

Overall, the Deep Learning model appears to be performing well on this dataset, with high accuracy and low error rates. However, it is always recommended to evaluate the model's performance on a separate testing dataset to ensure its generalizability to new data.



#### 5.4 Model 4: K nearest neighbour - Sia De Long (TP060810)

K-Nearest Neighbours (K-NN) is a non-parametric algorithm that makes no assumptions regarding the data's underlying distribution that used for classification and regression tasks. The "K" in K-NN stands for the number of neighbours that are taken into account when creating a prediction for a brand-new data point. The algorithm determines how far each new data point is from every other data point in the training set before making a forecast. The distance can be defined using a variety of distance metrics, including the Manhattan distance and Euclidean distance. The algorithm then chooses the K-nearest neighbours to the new data point based on the distance metric and predicts the value for the new data point using either the mean of the K-nearest neighbours or the most prevalent class. The algorithm is straightforward to use and has some potential applications, especially when the underlying decision boundary is nonlinear but it can be computationally expensive, especially when dealing with huge datasets, and it necessitates careful consideration of the distance metric and the value of K. Furthermore, the algorithm makes the unwarranted assumption that each attribute is equally significant, which may not always be the case (Rouse, 2017).

From the result, KNN shows that it is easy to understand and implement that does not require any complex mathematical calculations or assumptions. making it a popular choice for beginners in machine learning while it does not require any assumptions about the underlying distribution of the data which make it suitable for a wide range of applications, it can handle any number of features or classes while also be customized by choosing different values of k or different distance metrics to suit different problems or data types. Besides that, K-NN does not assume any specific distribution unlike parametric algorithms such as linear regression which means that K-NN can be used in situations where the underlying distribution is unknown or when the data is highly skewed or has outliers (Soni, 2020).

However, K-NN needs to calculate the distance between the new data point and all the existing data points in the training set to determine the nearest neighbours while it can be a time-consuming process, especially when dealing with large datasets. Not to mentioned, the algorithm needs to be run multiple times if different values of K are to be tried which will further increasing the computational cost. Besides that, K-NN is sensitive to irrelevant features which can negatively impact the accuracy of the algorithm. This is cause by the algorithm considering all the features equally important when calculating the distance

between data points. If some features are irrelevant or noisy, they can introduce unnecessary variability in the distance calculation and lead to inaccurate predictions (Soni, 2020).

In the case of predicting passenger satisfaction levels in the airline industry, KNN can be a useful algorithm as it can classify passengers into satisfied or unsatisfied categories based on their similarity to other passengers in the dataset. KNN works by calculating the distance between the new data point and all the other data points in the dataset. The algorithm then selects the  $k$  closest data points and assigns the new data point to the class that occurs most frequently among the  $k$  neighbors (Sezgen et al., 2019).

One of the benefits of using the KNN algorithm for this task is that it does not require any training phase, which means that it can be applied quickly and efficiently to new data. Additionally, KNN can work well on datasets with non-linear relationships between the variables and can handle missing data by imputing values based on the closest neighbors.

However, one potential drawback of using the KNN algorithm is that it can be sensitive to the choice of the value of  $k$ . If  $k$  is too small, the algorithm may be too sensitive to noise in the data, while if  $k$  is too large, the algorithm may fail to capture local patterns in the data.

To use the KNN algorithm for predicting passenger satisfaction levels in the airline industry, the below steps can be followed:

1. Load the dataset: Use the Read CSV operator to load the airline passenger satisfaction dataset.
2. Preprocess the dataset: Use operators such as Replace Missing Values and Nominal to Numerical to preprocess the data and convert categorical variables to numerical values.
3. Split the dataset: Use the Split Data operator to split the dataset into training and testing data, with a split ratio of 70:30.
4. Scale the data: Use the Normalize operator to scale the data and ensure that each feature contributes equally to the distance calculation.
5. Train the KNN model: Use the  $k$ -NN operator to train the KNN model on the training data.
6. Choose the value of  $k$ : Experiment with different values of  $k$  using the Cross Validation operator to find the value that provides the best performance on the testing data.

7. Make predictions: Use the Apply Model operator to make predictions on the testing data.
8. Evaluate the model: Use the Performance operator to evaluate the performance of the KNN model using metrics such as accuracy, precision, recall, and F1 score.

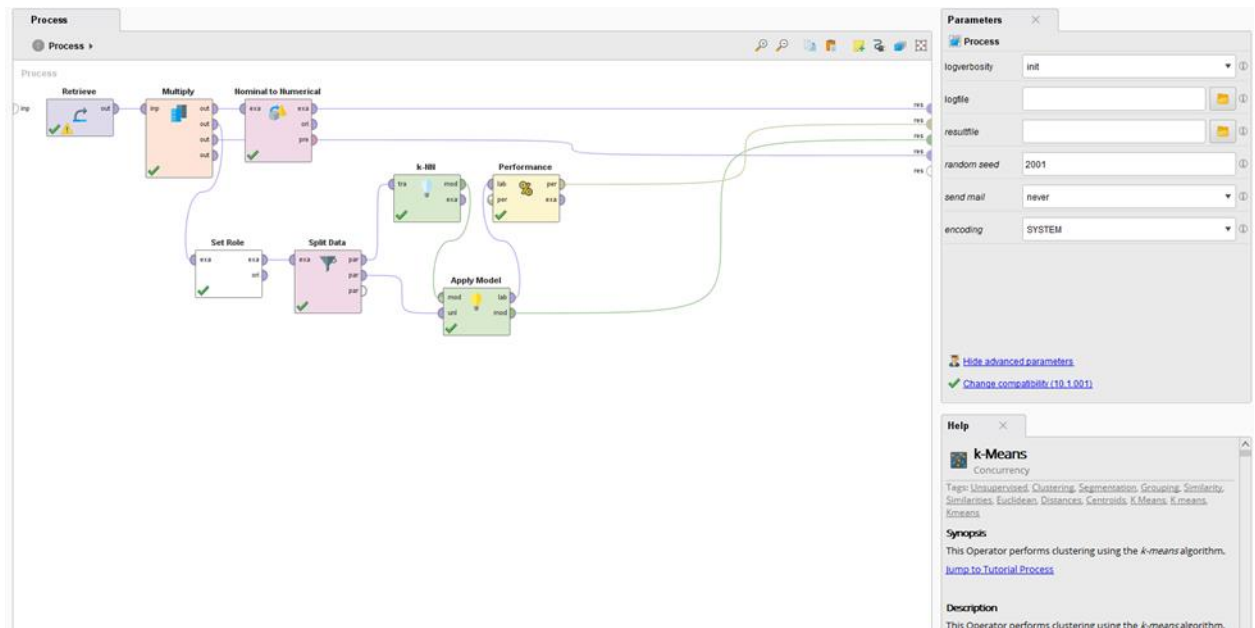


Figure 63 - KNN Model

The 'Parameters' dialog box for the 'k-NN' operator is shown. The parameters are:

- k: 5
- weighted vote: ☒ (checked)
- measure types: MixedMeasures
- mixed measure: MixedEuclideanDistance

Figure 64 - KNN Parameters Set

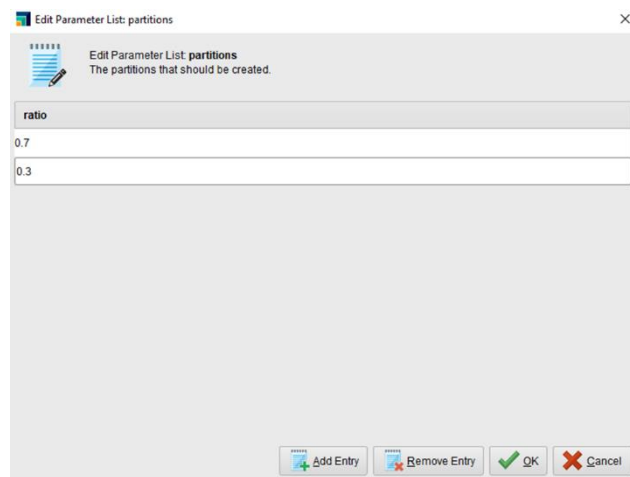


Figure 65 - Splitting Data for KNN Model

Table View Plot View

accuracy: 74.30%

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	14128	4505	75.82%
pred. satisfied	3481	8964	72.03%
class recall	80.23%	66.55%	

Figure 66 - Performance of KNN

Based on the program output, the KNN model achieved an accuracy of 74.30% on predicting passenger satisfaction levels in the airline industry. The confusion matrix shows that out of the 31078 samples, the model correctly predicted 14128 of the neutral or dissatisfied passengers and 8964 of the satisfied passengers. However, the model misclassified 3481 satisfied passengers as neutral or dissatisfied and 4505 neutral or dissatisfied passengers as satisfied.

The performance of the KNN model seems to be lower than the other models that have been used, such as Decision Tree and Random Forest. It is possible that the KNN model is not able to capture the complex relationships between the variables in the dataset, or that the value of  $k$  was not optimized for this specific task.

## 6.0 Evaluation

### 6.1 Comparison

Aspect \ Model	Decision Tree	Random Forest	Deep Learning	K-Nearest Neighbours (K-NN)
Accuracy	88.44%	82.45%	95.5%	74.30%
<b>Precision</b>				
Satisfied	86.35%	81.1%	95.5%	72.03%
Neutral or Dissatisfied	90.06%	83.49%	95.51%	75.82%
<b>Recall</b>				
Satisfied	87.10%	78.98%	94.06%	66.55%
Neutral or Dissatisfied	89.47%	85.24%	96.61%	80.23%
<b>F1 Score</b>				
Satisfied	0.87	0.80	0.95	0.69
Neutral or Dissatisfied	0.90	0.84	0.96	0.78

The highest precision for satisfied prediction is from deep learning model which is 95.5% while neutral and dissatisfied prediction is also from deep learning model with 92.68%. That means result of both satisfied with neutral and dissatisfied predicted from deep learning model is mostly accurate and can be trusted.

The highest recall for satisfied prediction is from decision tree model which is 94.06% while neutral and dissatisfied prediction is also from deep learning model with 96.61%. That means that decision tree is the most effective model to predict out the actual satisfied with neutral and dissatisfied from the data and has the least actual satisfied not predicted correctly.

Overall, deep learning has the highest accuracy and F1 score for both satisfied and neutral or dissatisfied prediction which mean it is the most suitable model to be chosen in this case while K-NN would be the worst model to be chosen because it has the poorest result among the models.

## 6.2 Critical analysis and recommendations

According to the evaluations done on the models in the previous stage, the results were that the deep learning model is the model with highest accuracy and F1 Score for the prediction of satisfaction levels as per our testing. With that knowledge in mind, the airline company will be able to use the deep learning model to perform analysis on existing datasets to find out what are the factors that affect customer satisfaction and come up with solutions to improve the whole airline experience for customers from before they even place a booking for their airline ticket to the end of their journey.

Based on the exploratory data analysis as done by the researcher in Section 4.1 on the dataset, there are a few statistics and analytics that show some erratic data and patterns in customer data such as in the dataset, it is observed that the age distribution of customers with age above 70 is comparatively lower compared to the customer distribution of ages below 70. This may be due to the fact that people of ages around 70 start to be more susceptible to illnesses and their bodies may be significantly frailer than that of the younger customers and they will have trouble handling some long flight hours and the economy seats that are not optimized for the elderly hence there is an observed lower customer distribution of customers above the age of 70. The airline company may use this information to provide services that are primarily targeted to attract customers of that age such as providing a special pricing on customers above the age of 70 while also including seats that are more comfortable and supportive to be suitable for senior citizens. This way they can improve customers in that age group and potentially improve sales in the airline market as the airline company can be the first amongst other companies to be offering such a service package. This ultimately can not only increase the revenue of the airline company but also increase customer satisfaction as customers are being offered services that fit what they need. The airline company can also use the best performing deep learning model to perform predictions as to what are the other services that would be needed and wanted by customers of the age group of interest.

Aside from age groups, the data analysis in stage 1 also shows that the airline company has a significantly higher loyal customer base compared to disloyal customers hence it highly suggests that the airline company look into ways and methods to retain those loyal customers and have their continued patronage. This can be achieved by the use of the best evaluated deep learning model to perform analysis on a separate dataset to learn about what are the characteristics or services that the airline company provides to the customers that differentiates a loyal customer from a disloyal one. As the analysis also shows that 68.88% of

the airline's customers use their services for business travels, the airline can offer services and packages that will appeal to the customers that use their airline for business travels such as offering seats with charging ports, wi-fi service in the air, a mini table that can support a laptop, some privacy for sensitive document confidentiality and much more which the deep learning model can be used to further refine the services that can be provided. The deep learning model can also be used to predict and suggest any changes that the airline company can make to increase or at least retain the amount of loyal customers in general in the best interests of the airline company in the long run.

As previously mentioned, where the airline company should provide services and packages that appeal more to their customers who use them for business travels, they can opt to replace the eco-plus class offerings in some of their planes with the seats and facilities instead of the low performing sales of the eco-plus sales. This way the airline company can turn a low demand sales option into an option that would help them further boost sales for their main sales market. These new seat offerings can replace the eco-plus seats in planes that will be used for long distance flights as analysis suggests that eco-plus customers and eco-class customers mainly fly short distances only and those seats should remain as an option to maintain customer satisfaction. The airline company can then use the deep learning model to predict what are the connections and relationships between customer satisfaction and the classes as there is significantly more satisfied customers in the business class compared to the eco-class and eco-plus class. If the airline can overcome this problem with the deep learning model, they will greatly increase their overall customer retention, customer satisfaction, revenue, and sales.



## References

- Agrawal, R. (2014). K-Nearest Neighbor for Uncertain Data . *International Journal of Computer Applications*, 11.
- AnalytixLabs. (2022, October 1st). *AnalytixLabs*. Decision Tree Algorithm in Machine Learning: Advantages, Disadvantages, and Limitations: <https://www.analytixlabs.co.in/blog/decision-tree-algorithm/>
- Baswardono, W., Kurniadi, D., Mulyani, A., & Arifin, D. M. (2019). Comparative analysis of decision tree algorithms: Random forest and C4. 5 for airlines customer satisfaction classification. *Journal of Physics: Conference Series*, 1402(6), 066055.
- Batista, G., & Silva, D. (2009). How k-Nearest Neighbor Parameters Affect its Performance. *Argentine symposium on artificial intelligence*, 1-12.
- Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114, 24-31.
- Brownlee, J. (2019, August 16th). *Machine Learning Mastery*. What is Deep Learning?: <https://machinelearningmastery.com/what-is-deep-learning/>
- Chen, T.-Y., Chang, Y.-H., Yang, M.-C., & Chen, H.-W. (2020). How to Cultivate a Green Decision Tree without Loss of Accuracy? *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 1-6. <https://doi.org/10.1145/3370748.3406566>
- ENAS , G., & CHO! , S. (1986). Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. *Statistical methods of discrimination and classification*, 235-244.
- Geeks For Geeks. (n.d.). *Geeks For Geeks*. Advantages and Disadvantages of Deep Learning: <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-deep-learning/>
- Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, 4(2), 13-19.
- IBM. (n.d.). *IBM*. Decision Trees: <https://www.ibm.com/topics/decision-trees>

- Jagandeep, S. (2020, December 18th). *Medium*. Random Forest: Pros and Cons: <https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04>
- Javatpoint. (n.d.). *KDD- Knowledge Discovery in Databases*. <https://www.javatpoint.com/kdd-process-in-data-mining>
- Kowsari, K., Brown, D. E., Heidarysafa, M., Jafari Meimandi, K., Gerber, M. S., & Barnes, L. E. (2017). HDLTex: Hierarchical Deep Learning for Text Classification. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 364–371. <https://doi.org/10.1109/ICMLA.2017.0-134>
- Kranz, G. (n.d.). *data classification*. <https://www.techtarget.com/searchdatamanagement/definition/data-classification>
- Li, S., Xia, G., & Zhang, X. (2023). Customer Churn Combination Prediction Model Based on Convolutional Neural Network and Gradient Boosting Decision Tree. *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*. <https://doi.org/10.1145/3579654.3579666>
- Manapragada, C., Webb, G. I., & Salehi, M. (2018). Extremely Fast Decision Tree. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1953–1962. <https://doi.org/10.1145/3219819.3220005>
- Perera, P. (2019). Decision Tree Approach for Predicting the Credit Risk of Leasing Customers in Sri Lanka. *Proceedings of the 3rd International Conference on Business and Information Management*, 65–68. <https://doi.org/10.1145/3361785.3361797>
- Rasool Fakoor, & Azade Nazi. (n.d.). *Using deep learning to enhance cancer diagnosis and classification*.
- Rouse, M. (2017, March 14th). *Technopedia*. K-Nearest Neighbor: <https://www.techopedia.com/definition/32066/k-nearest-neighbor-k-nn>
- Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1), 3-29.
- Sezgen, E., Mason, K. J., & Mayer, R. (2019). Voice of airline passenger: A text mining approach to understand customer satisfaction. *Journal of Air Transport Management*, 77, 65-74.

- Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, 5(4), 2094-2097.
- Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 6308-6325.
- Soni, A. (2020, July 3). *Medium*. Advantages And Disadvantages of KNN: <https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336>
- Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, 134, 93-101.
- Stancin, I., & Jovic, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. *42nd International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 977-982.
- Veres, M., & Moussa, M. (2020). Deep Learning for Intelligent Transportation Systems: A Survey of Emerging Trends. *IEEE Transactions on Intelligent Transportation Systems*, 21(8), 3152–3168. <https://doi.org/10.1109/TITS.2019.2929020>
- What is Deep Learning? (n.d.). Retrieved April 30, 2023, from <https://www.mathworks.com/discovery/deep-learning.html>
- Wu, S., Xia, N., Ren, Y., & Wang, Z. (2022). A Classification Prediction Method Using Rough Set and Decision Tree. *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 552–557. <https://doi.org/10.1145/3532213.3532297>
- Xu, Q., & Yin, J. (2021). Application of random forest algorithm in physical education. *Scientific Programming*, 1-10.
- Yiu, T. (2019, Jun 12nd). *Medium*. Understanding Random Forest: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
- Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11). Introduction to machine learning: k-nearest neighbors.

Zhu, B., Xie, G., Yuan, Y., & Duan, Y. (2018). Investigating Decision Tree in Churn Prediction with Class Imbalance. *Proceedings of the International Conference on Data Processing and Applications*, 11–15. <https://doi.org/10.1145/3224207.3224217>

Hou, Y. C., Baharuddin, M. Z., Yussof, S., & Dzulkifly, S. (2020). Social Distancing Detection with Deep Learning Model. *2020 8th International Conference on Information Technology and Multimedia (ICIMU)*, 334–338. <https://doi.org/10.1109/ICIMU49871.2020.9243478>

## Workload matrix

	Yan Mun Kye (TP056066)	Sia De Long (TP060810)	Hor Shen Hau (TP061524)	Tan Sheng Jeh (TP056267)
Introduction	25%	25%	25%	25%
Methodology	25%	25%	25%	25%
Literature review	25%	25%	25%	25%
Data preparation	25%	25%	25%	25%
Decision tree	100%	N/A	N/A	N/A
Random forest	N/A	N/A	N/A	100%
Deep learning	N/A	N/A	100%	N/A
KNN	N/A	100%	N/A	N/A
Evaluation	25%	25%	25%	25%