CONTINOUS ASSIGNMENT 2 COVER SHEET

Name and student number:

Karan Koundinya Janakiram
10602768

Course Title: Master of Science in Data Analytics

Lecturer Name: Terri Hoare

Module/Subject Title: B9DA110 Advanced Data and Network Mining

Assignment Title: Continuous assessment lessons learned report for the application of data mining techniques for the prediction of airline satisfaction (CA2)

By submitting this assignment, I confirm that I am aware of DBS's policy regarding cheating, plagiarism and all other forms of academic impropriety. The coursework submitted is my own or my group's work, and all other sources consulted have been appropriately acknowledged.

I am aware that in the case of doubt an investigation will be held

Name: Karan Koundinya Janakiram Date: 23/4/2023

INDEX

CONTINUOUS ASSESSMENT LESSONS LEARNED REPORT FOR THE APPLICATION OF DATA MINING TECHNIQUES FOR THE PREDICTION OF AIRLINE SATISFACTION

INTRODUCTION:

In the current business ecosystem leveraging the power of data to make a sizeable contribution on an organizational level is of highest importance considering the added revenue which can obtained by adopting these methods. In this age of divergent technologies, data has become the cornerstone of innovation, so to capitalize on this becomes necessary for organizations and stakeholders. Analysing data to gain actionable insights became an important step in leveraging the power of data, but the traditional methods of data analysis were unable to cope with the vast amounts of data being generated and failed to provide valuable insights. This created a need for automated techniques that could discover hidden patterns and relationships in the data

This led to the development of data mining algorithms and tools, Data mining is the process of discovering patterns, relationships, and insights from large datasets. It involves using statistical and machine learning techniques to extract valuable information that can be used for business intelligence, decision-making, and other applications. Data mining was introduced as a way to address the challenge of extracting useful information from the growing amount of data that was being collected by organizations. In the 1980s, the amount of data being generated started to exceed the capacity of traditional methods of data analysis, such as manual data entry and simple statistical analysis. This led to the development of automated techniques for discovering patterns and relationships in data, which became known as data mining.

The introduction of data mining was driven by the need to make better use of the large and complex data sets that were being generated by businesses, government agencies, and other organizations. By automating the process of discovering patterns and relationships in data, data mining allowed organizations to gain valuable insights that could be used to improve decision-making, identify new opportunities, and gain a competitive edge in their respective industries.

With the use of data mining the need to turn data into information is not just unidimensional as it cuts across various domains across the globe and helps take critical decisions in the marketplace were providing decision support systems have led to increased revenues, mitigation of potential risks and improving bottom-line.

Given below are the examples of the how data mining is used in multiple domains:

A manufacturing company can use Data mining to analyse production data to identify patterns in machine performance, which can be used to improve efficiency and reduce downtime

A financial institution uses Data Mining to analyse transaction data to detect fraudulent activity, such as unauthorized credit card usage, to prevent financial losses

A healthcare provider uses Data Mining to analyse patient data to identify trends in diagnoses and treatments, which can help improve patient outcomes and reduce healthcare costs

Data mining can be used in sports analytics to unearth player performance patterns and with that help in improving the weak aspects of a player's game

APPLICATION OF DATA MINING FOR OUR CA:

For our CA, we have considered the "AIRLINE PASSENGER SATISFACTION" dataset from Kaggle which primarily deals with the deals with airline passenger satisfaction. It includes data on various features of the airline experience, such as the airline type, customer type, inflight services, and onboard amenities. The goal of the dataset is to predict whether or not a passenger is satisfied with their airline experience based on these features

The main reasons why we as a team opted for this dataset is because the dataset deals with a common issue in the airline industry that is customer satisfaction, this makes it meaningful for researchers like us to analyse and study the underlying patterns and behavioural aspect of the data associated with this topic

To analyse the dataset at hand we have used the two tools majorly

- Google Colab for scripting in python
- Rapid Miner

Since the main question we have picked from the dataset is to classify whether the passenger's experience was satisfactory or dissatisfactory/neutral, we have used the Logistic Regression as our base algorithm followed by Decision Tree and K Nearest Neighbour, i have implemented the same in both rapid miner and on Google Colab.

I have picked logistic regression because it is a statistical technique used for binary classification problems, where the goal is to predict the probability of an event occurring and, in our case, to check whether the experience was satisfactory or dissatisfactory/neutral

we have implemented the same algorithms in both the tools as according to the CRISP DM methodology

CRISP-DM has helped businesses identify patterns and trends in their sales data, allowing them to make informed decisions about pricing strategies, package positioning, and customer targeting. By applying the CRISP-DM methodology, businesses can analyse their data in a systematic and repeatable way, ensuring that their revenue decisions are based on reliable insights and not just guesswork

The six phases of the CRISP DM methodology are mentioned below along with steps of our research along with each phase

- **BUSINESS UNDERSTANDING:**

  The steps in this phase are
  Define the problem: The problem is to predict passenger satisfaction with airline services based on various factors such as flight distance, gender, inflight WIFI service, etc.
  Determine the goals: The goal is to develop a logistic regression model that can accurately predict whether a passenger will be satisfied or dissatisfied with the airline services.
  Define the success criteria: The success criteria are to achieve a high accuracy in predicting passenger satisfaction out of all the algorithms from the tools considered, we need to pick which is the best performing algorithm from both the tools

  To aid our understanding of the data and to help the airlines organizations from a business and brand point of view, we have created four metrics which will us a clear picture on how the passengers are feeling about their flight journeys and the negative aspects of it can be taken into account by the organization and improved so that it can help their brand and business.

- **UNDERSTANDING THE DATA:**

  The main steps in this stage are loading the dataset onto the tools and then analyse the data to identify patterns and relationships between variables. This includes examining the distribution of each variable, checking for missing values, and identifying potential outliers The below screenshots are from google Collab which depicts the steps above

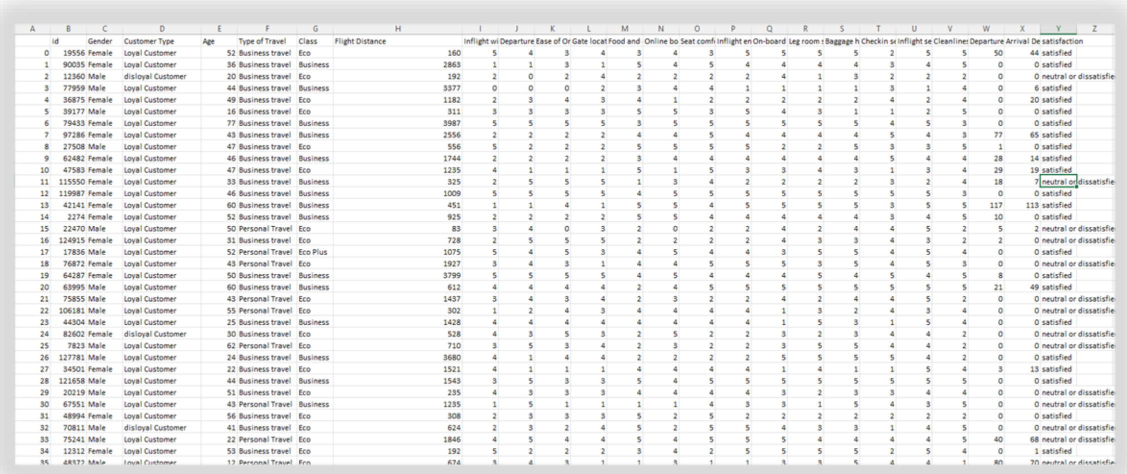


FIG 1.1 Snapshot of the Dataset

BY PRINTING THESE TWO PIECES OF INFORMATION, THE CODE PROVIDES AN OVERVIEW OF THE STRUCTURE OF THE DT DATAFRAME, INCLUDING ITS SHAPE AND COLUMN NAMES.

```
from numpy.core.arrayprint import printoptions
print(dt.shape)
print(dt.columns)
```

```
(25976, 25)
Index(['Unnamed: 0', 'id', 'Gender', 'Customer Type', 'Age', 'Type of Travel',
       'Class', 'Flight Distance', 'Inflight wifi service',
       'Departure/Arrival time convenient', 'Ease of Online booking',
       'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort',
       'Inflight entertainment', 'On-board service', 'Leg room service',
       'Baggage handling', 'Checkin service', 'Inflight service',
       'Cleanliness', 'Departure Delay in Minutes', 'Arrival Delay in Minutes',
       'satisfaction'],
      dtype='object')
```

FIG 1.2 Snapshot of the code

CHECKING FOR NULL VALULES

```
[ ]  print(dt.isnull())
```

```
25972          False              False
25973          False              False
25974          False              False
25975          False              False

       Departure/Arrival time convenient  ...  Inflight entertainment  \
0                                  False  ...                   False
1                                  False  ...                   False
2                                  False  ...                   False
3                                  False  ...                   False
4                                  False  ...                   False
...                                  ...  ...                     ...
25971                              False  ...                   False
25972                              False  ...                   False
25973                              False  ...                   False
25974                              False  ...                   False
25975                              False  ...                   False
```

FIG 1.3 Snapshot of the code

- **DATA PREPARATION:**

  The main steps in this phase are cleaning and transforming the categorical variables into numerical values, normalize numerical variables

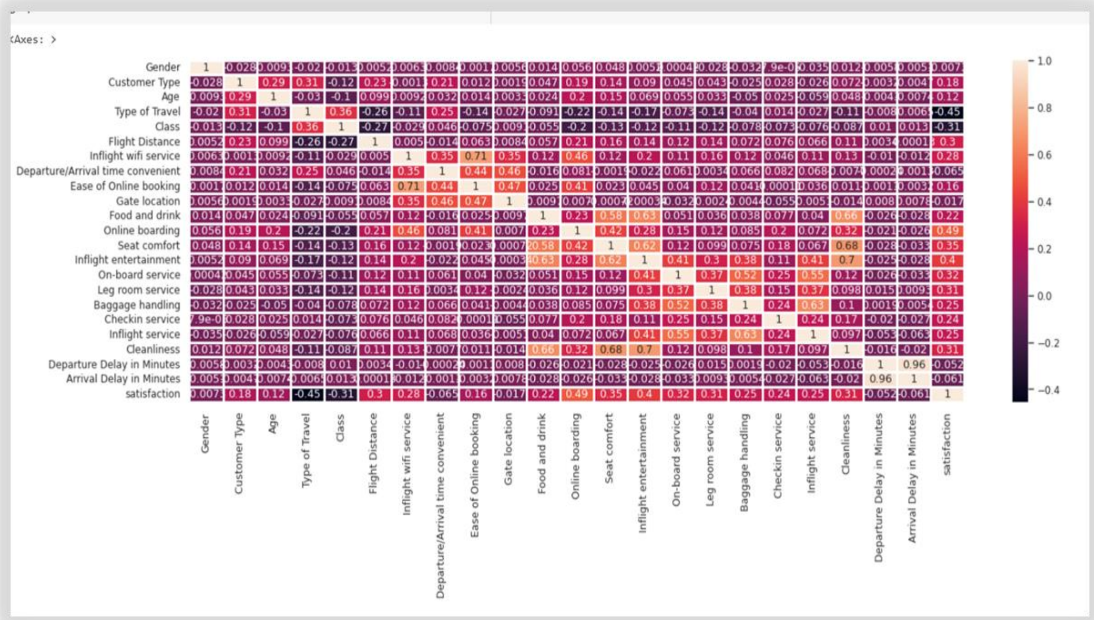


FIG 1.4 Snapshot of the correlation matrix

Here in the correlation matrix the values under 0.32 are considered to be dropped to progress with the logistic regression

DataFrame dt is being used to predict the satisfaction of airline customers. The drop method is used to remove several columns that are not being used as features in the prediction, such as Gender, Customer Type, Age, Type of Travel, Class, Flight Distance, and various ratings related to the airline service.

The resulting DataFrame x contains only the relevant features, while the satisfaction column is stored separately in y.

This suggests that a machine learning algorithm will be used to train a model to predict the satisfaction level of airline customers based on the remaining features in x

- **MODELLING:**
  This mainly deals with the use of statistical modelling methods based on the business problem at hand, we often try to look at a model which gives us a better rate of accuracy over a model which aims at perfection
  The steps in this phase are Build the model: Train the following models
  logistic regression, Decision Tree and K Nearest Neighbour on the training data.
  Evaluate the model: Evaluate the performance of the model on the testing data using evaluation metrics such as accuracy, precision, recall, and F1 score. Down below are the screenshots of the algorithm applied on both Colab and Rapid Miner

LOGISTIC REGRESSION:

These two screengrabs depict the logistic regression algorithm in Google colab and Rapid Miner

LOGISTIC REGRESSION

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.30, random_state = 101)

model = LogisticRegression(max_iter=10000)
model.fit(x_train, y_train)
model.score(x_test, y_test)
```

```
0.8037982805081484
```

```
y_pred = model.predict(x_test)
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test, y_pred))
```

```
[[3663  684]
 [ 845 2601]]
              precision    recall  f1-score   support

           0       0.81      0.84      0.83      4347
           1       0.79      0.75      0.77      3446

    accuracy                           0.80      7793
   macro avg       0.80      0.80      0.80      7793
weighted avg       0.80      0.80      0.80      7793
```

FIG 1.5 LR in Google Colab

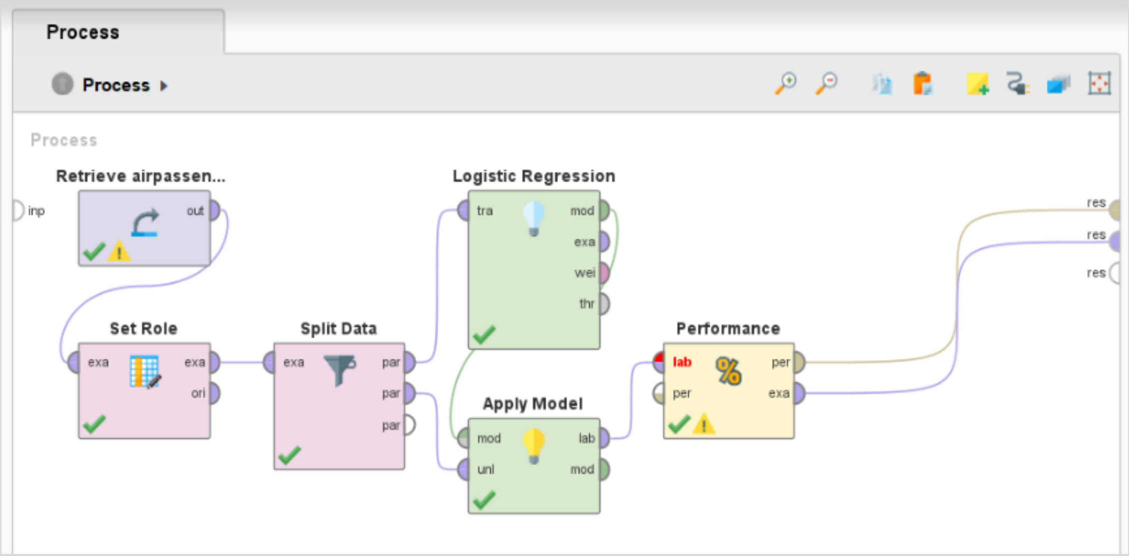


FIG 1.6 LR in Rapid Miner

DECISION TREE:

These two screengrabs depict the Decision Tree algorithm in Google colab and Rapid Miner

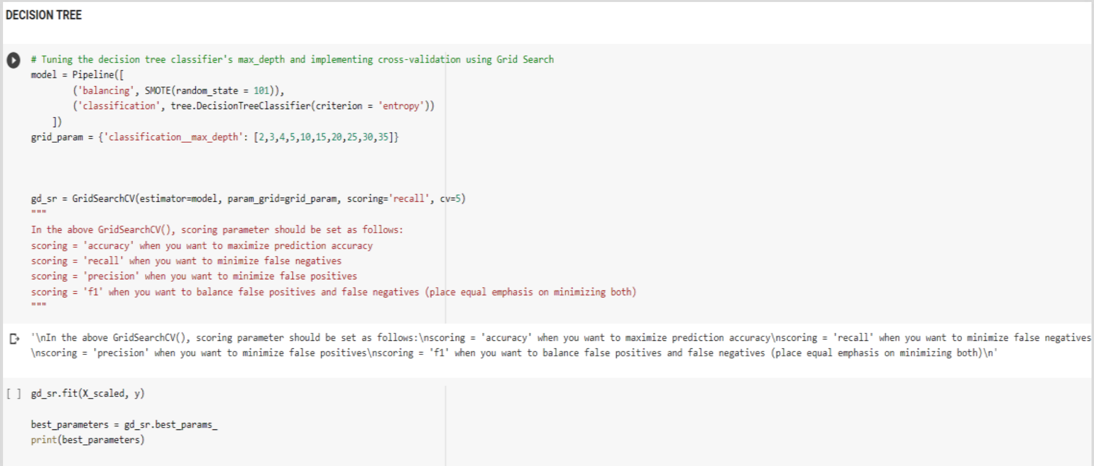


FIG 1.7 DT in Google Colab

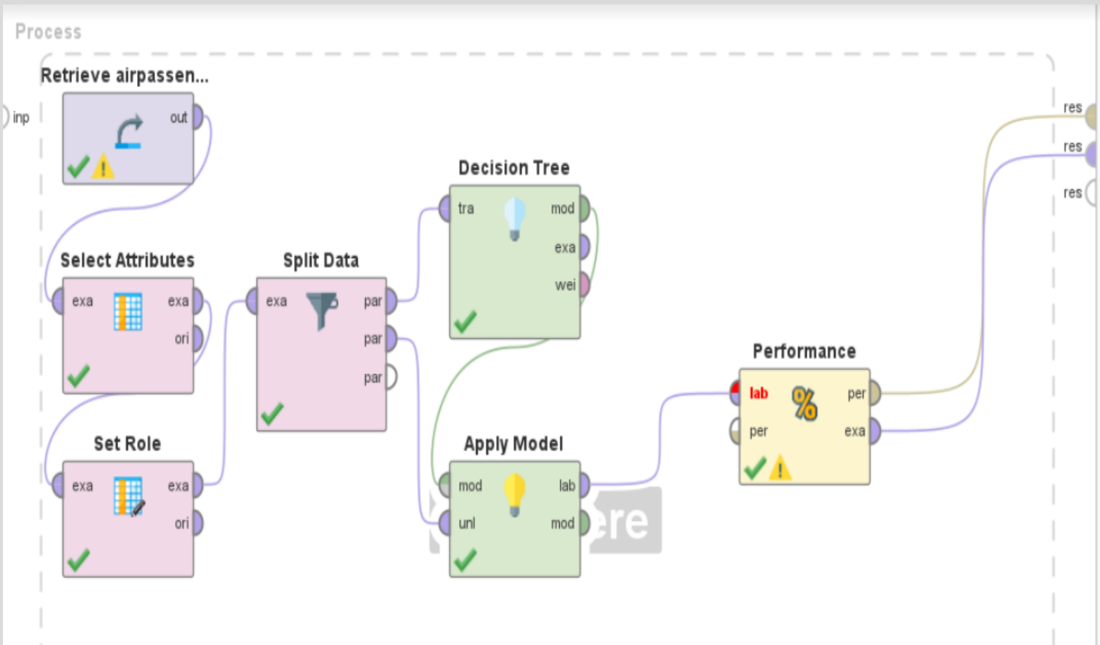


FIG 1.8 DT in Rapid Miner

K NEAREST NEIGHBOUR:

These two screengrabs depict the Decision Tree algorithm in Google colab and Rapid Miner

K NEAREST NEIGHBOUR

```
#As we can see 197 is an odd number and Neighbours should be chosen odd.
#define Model: init K-NN
classifier=KNeighborsClassifier(n_neighbors=197,p=2,metric='euclidean')

#Fit the model

classifier.fit(x_train,y_train)
```

```
KNeighborsClassifier
KNeighborsClassifier(metric='euclidean', n_neighbors=197)
```

```
#Predict the test result

y_pred=classifier.predict(x_test)
y_pred
```

```
array([1, 0, 0, ..., 0, 0, 0])
```

FIG 1.9 KNN in Google colab

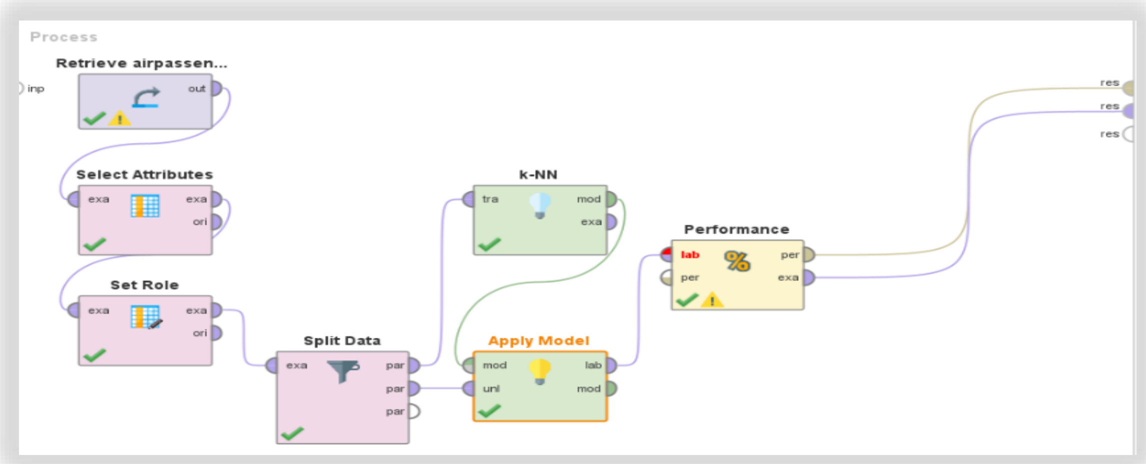


FIG 1.9 KNN in Rapid Miner

- **EVALUATION:**

  This mainly deals with the outcome of the model selected and what is its business implication with respect to the question at hand
  So, in our research we have to compare the results of the three algorithms under consideration and out of them compare which one of them is best suitable from a business point of view

  To check whether data was unbalanced or not, I tried out the smote up sampling, even after that the accuracy and precision percentages remained the same as before it was unsampled

LOGISTIC REGRESSION:

```
[[3663  684]
 [ 845 2601]]
              precision    recall  f1-score   support

           0       0.81      0.84      0.83      4347
           1       0.79      0.75      0.77      3446

    accuracy                           0.80      7793
   macro avg       0.80      0.80      0.80      7793
weighted avg       0.80      0.80      0.80      7793
```

FIG 1.10 Performance metrics in LR(COLAB)

ExampleSet (Apply Model) | PerformanceVector (Performance)

Table View | Plot View

accuracy: 87.01%

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	19874	2918	87.20%
pred. satisfied	2144	14028	86.74%
class recall	90.26%	82.78%	

FIG 1.11Performance metrics in LR (RAPID MINER)

DECISION TREE

```
gd_sr.fit(X_scaled, y)

best_parameters = gd_sr.best_params_
print(best_parameters)

best_result = gd_sr.best_score_ # Mean cross-validated score of the best_estimator
print(best_result)
```

```
{'classification__max_depth': 2}
0.8579323473084288
```

FIG 1.12 Performance metrics in DT (GOOGLE COLAB)

Table View   Plot View

accuracy: 76.31%

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	17715	4928	78.24%
pred. satisfied	4303	12018	73.64%
class recall	80.46%	70.92%	

FIG 1.13 Performance metrics in DT (RAPID MINER)

KNN:

```
#evaluate the model

cm=confusion_matrix(y_test,y_pred)
cm

array([[3852,  495],
       [ 821, 2625]])

print(f1_score(y_test,y_pred))

0.7995735607675907

print(accuracy_score(y_test,y_pred))

0.8311305017323238
```

FIG 1.14 Performance metrics in KNN (GOOGLE COLAB)

Table View Plot View

accuracy: 83.63%

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	19069	3429	84.76%
pred. satisfied	2949	13517	82.09%
class recall	86.61%	79.77%	

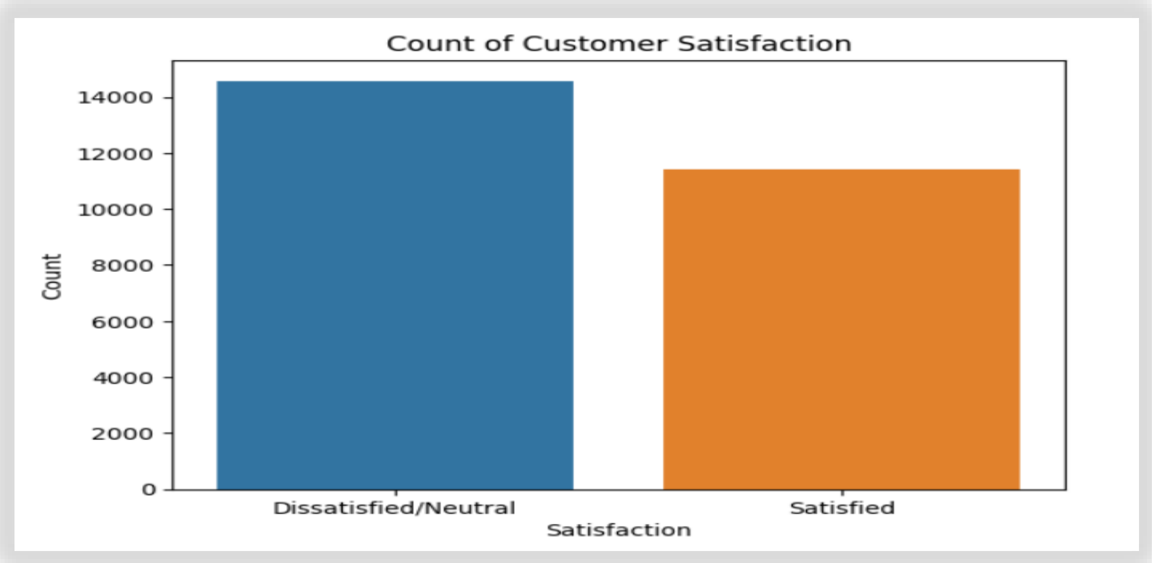
FIG 1.15 Performance metrics in KNN (RAPID MINER)

From the algorithms applied it is crystal clear that out of the algorithms constructed on colab, Decision tree seems to have the best accuracy with 85.79%
While on rapid miner, the logistic regression algorithm has the best accuracy with 87%
Even if the decision tree and logistic regression models were trained on the same dataset, the difference in the performance metrics (accuracy) may be due to several reasons such as differences in the pre-processing steps, hyperparameters tuning, and evaluation metrics used in the two environments (Colab and RapidMiner). Therefore, it may not be appropriate to directly compare the accuracy of these models without considering other factors

- **DEPLOYMENT:** This is the final stage in the CRISP DM methodology, this mainly deals with the prerequisite which was specified when the projected was initiated to take off using the CRISPDM methodology
  To aid the deployment of the model in a real-world scenario and to give a descriptive idea of the performance of the model we have created few Key Performance Indicators

  We have created metrics so that it can help us gauge how the model would be useful in a real-world scenario, we have created

COUNT OF CUSTOMER SATISFACTION: This can help us get a clear understanding of what is the share of satisfied and dissatisfied passengers, it is evident that the percentage of dissatisfied customers is greater than satisfied customer which is an area of concern for the airline companies



ON TIME ARRIVAL PERCENTAGE: This is used to give an indication that only 56.18% which is again a cause of concern because this indicates that out of 100 passengers only 56.18 of them arrive on time

METRIC 2 PERCENTAGE OF CUSTOMERS ARRIVING ON TIME

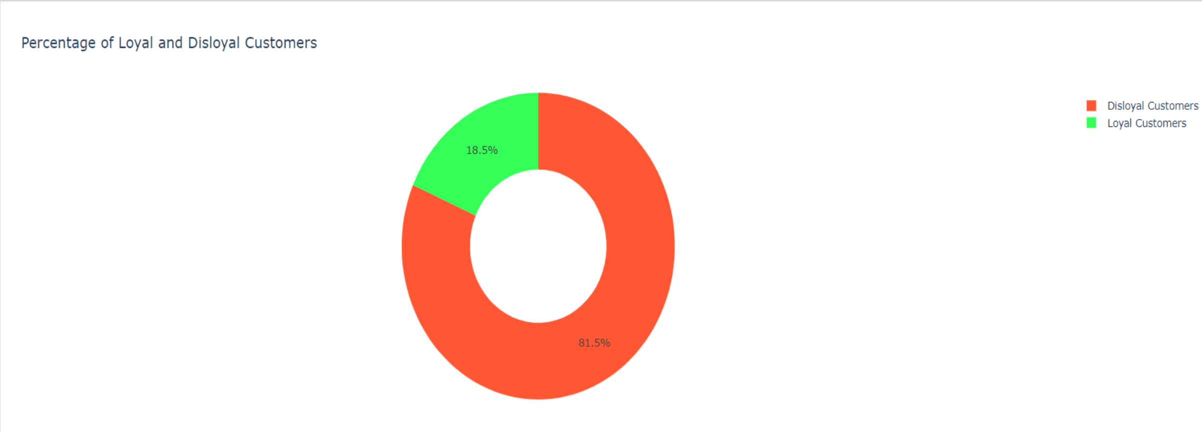
```
[ ]  on_time = dt[dt['Arrival Delay in Minutes'] <= 0]
```

```
[ ]  on_time_percentage = len(on_time) / len(dt) * 100
```

```
print(f"On-time arrival percentage: {on_time_percentage:.2f}%")
```

```
On-time arrival percentage: 56.18%
```

PERCENTAGE OF LOYAL AND DISLOYAL CUSTOMERS: This gives us the percentage of loyal and disloyal customers; the share of disloyal customers is greater than the loyal passengers



COUNT OF CUSTOMER COMPLAINTS: The count of customers complaints is categorized into the range of severity from 0-5, with zero being the least and five being the most on the severity scale, the greatest number of complaints are from 4 and 5

The above factors indicate that the customers have not been enjoying their flight experiences due a variety of factors, these factors can be used by airline organizations to improve their flight services in ways that would keep the passengers satisfied.

LESSONS LEARNT FROM THIS CA:

While working on this CA has helped me understand how the CRISP DM provides a structured framework that helps organizations to effectively plan, execute, and manage their data mining projects

With the phases of CRISP DM methodology, it emphasizes the importance of understanding the business problem or opportunity at hand before delving into data analysis. This ensures that the analytics efforts are aligned with business objectives and have a greater likelihood of generating meaningful insights.

Also, by following this methodology it has helped me understand the how businesses to maintain a structured approach to data mining and analytics. This not only increases the chances of success but also enables businesses to reuse the methodology for future projects, leading to more efficient and effective analytics initiatives over time.

With regards to the current dataset, we had considered by working on this project it helped me understand how to construct machine learning algorithms on the basis of the problem at hand, it also gave me a clear understanding of how the algorithms can be evaluated in order to pick which one of the algorithms is performing better when put in a state of comparison with other algorithms under consideration.

While in the deployment phase, deploying the analytics model into production helped me understand we can use the insights gained from the model to make more informed decisions

Adding to the previous point it can provide businesses with a competitive advantage over their peers, by continuously monitoring and refining the analytics model over time, businesses can further optimize their operations and gain even more insights from their data

By constructing the key performance indicators in our CA, it helped me gauge how these can be picked up the by airline giants to recalibrate and rethink their business strategies based on the acquired results. From our research it is clear the currently we have more dissatisfied customers, disloyal customers and the number of complaints on the severity list, this is an area of concern for the airline giants

FUTURE WORK:

For the current research we have in our hands we can predict which factors out of the list contributes the most in all the negative aspects of the customer feedback, if we can zero down on a couple of factors, we can leverage the power of data with research and development to reduce the effect of that factor, that would in turn bring in a change in the customer satisfaction levels.

INDUVIDUAL CONTRIBUTION:

I have built the algorithms in python

I have built the KPI's considered

I have formulated the presentation.

this was like a win win situation for me as I got to express what I had learnt through analyzing the data through another medium that is technical writing medium. This has ignited this spark to hone my skills with respect to technical writing and also it gave me an opportunity to express what I had learnt while analyzing the data. Having also formulated the presentation document that's the one is which will be presented by my team on the day of the presentation. To add on to the above-mentioned points, this has helped me master many of the main aspects of working on projects, those are aspects like Teamwork, time management and communication.