1.0 Dataset Background

Customer loyalty is an important source of competitive advantage that can be provided by the company to keep their loyal customers. One of the most effective ways to keep the customer loyal is to provide a good customer experiences and satisfaction during the service or for the product provided. A previous report have confirmed that a firm's most profitable customers have a strong tendency to be attracted by a good quality alternatives [1]. That's why maintaining a good customer in-flight satisfaction is very important to keep an airline company be competitive.

1.1 Dataset Introduction

In this report, a dataset from Kaggle have been retrieve to examine the customer in-flight satisfaction [2]. The customer in-flight satisfaction will be evaluate and discuss by building several models to determine the customers in-flight satisfaction.

The nature of this dataset is a survey provided to the airline passenger satisfaction according to their opinion towards several airline services and their satisfaction to the airline service. There are total of 24 attributes in the dataset which are stated as below:

- Id: Customer Id
- Gender: Gender of the passengers
- Customer Type: The customer type
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers
- Class: Travel class in the plane of the passengers
- Flight distance: The flight distance of this journey
- Inflight Wi-Fi service: Satisfaction level of the inflight Wi-Fi
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of Gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of Seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment
- On-board service: Satisfaction level of On-board service
- Leg room service: Satisfaction level of Leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of Check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure
- Arrival Delay in Minutes: Minutes delayed when Arrival
- Satisfaction: Airline satisfaction level

1.2 Dataset Class Distribution

The dataset attribute can be distributed to several types and class. The dataset class can be distributed by using its dataset characteristic such as age and gender. In this dataset, there are four main distribution that can be done to the dataset class which are shown in the below

- Gender (Male or Female)
- Customer Type (Loyal or Disloyal)
- Type of Travel (Personal or Business)
- Flight Class (Eco, Eco Plus or Business)

Since the age of the passenger have a wider range and difficult to be determined, the variables age will not be consider as a class in this dataset. The balancing of each class will be done in the next section of the report.

Two of the main data types are Numerical Variables and Categorical Variables. In terms of numerical variables are those variables that should be treated as they are in mathematics and the mathematics ways such as subtraction of summation are meaningful for them. For example the age should be considered as numerical variables however a postcode address shall not be consider as a numerical variables.

Where categorical variables are the variables that should not be treated like numbers which as mentioned in the above the postcode address. However, some of the categorical variables may have a natural ordering which will determine their effect, that's why inside of categorical variables, there are Nominal Categorical Variables which doesn't have ordering and Ordinal Categorical Variables that have ordering.

In terms of programming, the variables can also be distributed into several variables types which are Integer, Double, Character and Boolean. The below tables shows the attributes types in terms of their data types and their programming types in this dataset.

Types of Variables	**Types of variables (Programming)**	**Attributes**
Numerical (Discrete)	Integer	• Age • Flight Distance • Departure Delay in Minutes • Arrival Delay in Minutes
Categorical(Nominal)	Boolean	• Gender • Customer Type • Type of Travel
	Integer	• Id
Categorical(Ordinal)	Integer	• Inflight Wi-Fi 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 • Checking service • Inflight service • Cleanliness • Satisfaction
	Character	• Class

For the Numerical Discrete attributes, all the integer in the type of variables such as age and flight distance act as a whole number. The age of passenger is in years, the flight distance is in KM, where both of the Departure delay and Arrival delay is in minutes.

In terms of nominal categorical variables, there are both Boolean and integer types of variables. The reasons why gender, customer types and type of travel is in Boolean is because they only have two possible outcome which are male or female, business or personal travel and loyal or disloyal customers. The only integer type of attributes in Nominal Categorical variable is Id. The Id act as the user Id to identify the user identity and their experiences during the flight.

Where the ordinal categorical variables have been divided into two main programming variable types which are Integer and Characters. Most of the evaluation attribute done in the survey by the passenger is in ordinal categorical type and in Integer form. As the attribute is identify by the range of 1 to 5 that the higher number stand for the higher satisfaction. Where the flight class is considered as Ordinal Categorical Type in character form that have its ordinal characteristic. There are three class which are Eco, Eco Plus and Business class, that the Eco class is the lowest class where the Business class is the highest class.

1.3 Dataset Characteristic

Before the data mining process, the dataset have to be clear and to ensure that the dataset class is balanced. A dataset is consider as imbalanced if one class such as age or gender vastly outnumbers the other class [3]. An imbalanced dataset may led to overfitting model and poor performance in terms of generalization. In this part of the report, several method will be used to ensure that the dataset class is balanced.

As mentioned in the previous part, the dataset can be distributed into four main class which are distributed by using their Gender, Customer type, Type of Travel and Flight class. The data class will be considered as imbalanced if one of the value have more than 82% among all of the value. In this case, the value in each of the attribute need to be check and to ensure that none of them occupy more than 82% of all of the value [3].

Firstly, let look at the gender variables. There are only two value in the Gender attribute which are the male and female value. The total rows of the dataset is 103904 where the frequency of each of the value can be retrieve by using python by using the

value_counts() function. From the above function, the frequency for male is 51177 and for female is 52727, which they have occupy for 49.25% for male and 50.75% for female. Since both of the value in gender doesn't exceed 82% of the total data, the gender dataset class is consider as balanced and no further modification is required.

Next, the Customer type class have two value which are loyal and disloyal customer. The frequency for loyal customer is 84923 where the frequency for disloyal customer is 18981, which they have occupy for 81.74% for loyal customer and 18.26% for disloyal customer. Even though the percentage of loyal customer is fairly higher than the disloyal customer, however the percentage of loyal customer is 81.74% which is still lower than the standard 82% of the dataset class, therefore the customer type dataset class is still consider as balanced and no further modification is required.

The type of travel is divided into two values which are the business type and the personal type. The frequency for the business type travel is 71655 where the personal type is 32249 which means the business type travel have occupy for 68.96% and 31.04% for personal type. The business type travel have a higher percentage compare to personal type but since it is still under 82%, the type of travel dataset class is consider as balanced and no further modification is required.

Finally, the flight class have three main values which are Eco, Eco Plus and Business. The frequency of the values are 46745 for Eco, 7494 for Eco Plus and 49665 for Business which the percentage are 44.99% for Eco, 7.21% for Eco Plus and 47.8% for Business. Business class have the highest frequency among these three values where Eco plus have the lowest frequency. Since none of the value exceed 82% the flight class is considered as balanced. The summary of each dataset class with their values percentage and balance status have been shown in the below table:

Dataset Class	**Attribute Values**	**Status**
Gender	• Male (49.25%) • Female (50.75%)	Balanced
Customer Type	• Loyal Customer (81.74%) • Disloyal Customer (18.26%)	Balanced
Type of Travel	• Business Travel (68.96%) • Personal Travel (31.04%)	Balanced
Class	• Eco (44.99%) • Eco Plus (7.21%) • Business (47.8%)	Balanced

2.0 Feature Selection Option

Since the main problem is to examinate the satisfaction for the airline passenger, the main objective attribute for this report will be the satisfaction of the passenger. There are only two results for the passenger satisfaction which are ‘satisfied’ and ‘neutral or dissatisfied’ which will be act as the dependent variable.

For the independent variables, the survey form attribute which been done by the passenger will be used and there is total 14 of them. The machine learning model will be build by these attributes and its performance will be determined. The attributes and attributes that are selected are listed in the below table:

Variables	**Attributes**
Dependent	• Satisfaction
Independent	• Inflight Wi-Fi 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 • Checking service • Inflight service • Cleanliness

Since the dataset is ready and consists of test and train data, the machine learning model that will be the Decision Tree model and the Multiple Linear Regression Model. These two machine learning model will be built and their performance will be compared. Besides, the number of variables used will also be vary each time to determine the effect of these variable numbers to the model performance.

There are several features to identify the classifiers model. The function that will be used in python will be the model.score() function. This function will try to predict the objective variables and determine it score by comparing it with the actual data which is the test data. The higher the score means the higher performance for the model. When examine the performance of decision tree, a confusion matrix will be plotted to observe the performance of this model. However, confusion matrix is not suitable to examinate the performance of multiple linear regression since its variable type is float.

3.0 Classifiers and Test Evaluation

The test and train data have already been separated in the dataset source which are 25976 items for test data and 103904 items for the train data. The train data is 80% of the total data where the test data is 20% of the total data, this makes the train to test ratio to be 1:4. The machine learning model will be built by using the train data and it performance will be determine by predicting the test data.

As mentioned in the previous part, two machine learning model which are Decision Tree and Multiple Linear Regression will be used to build the model and their performance will be evaluate. Decision Tree is a non-parametric classifiers where the multiple linear regression is a parametric classifiers.

In generally, since the independent variables are in whole number and have a lower range of value, decision tree will act more effective in prediction as compare to the multiple linear regression method [4]. Therefore, the hypothesis for the report will be as shown in below:

H_0 : Decision Tree Model have a better performance then Multiple Linear Regression Model in passenger satisfaction prediction.

H_1 : Multiple Linear Regression Model have a better performance then Decision Tree Model in passenger satisfaction prediction.

Besides, there are total number of 14 independent variables in the model. Assume that as more independent variables involved in the model, the model will provides a higher performance and result for the prediction. Therefore, the second hypothesis in this report will be:

H_0 : As more independent variables used, the model will have a better performance.

H_1 : As more independent variables used, the model will not have a better performance.

4.0 Experiments

To deal with the second hypothesis, several independent variables have to be select to be use when the variables have a lower number. The variable selected will be the same when comparing both Decision Tree and Multiple Linear Regression model and the selection of variables are stated in the below table.

Variables Number (n)	**Variables selected**
n = 14 (All)	• Inflight Wi-Fi 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

	• Checking service • Inflight service • Cleanliness
n = 8	• Inflight wifi service • Departure/Arrival time convenient • Ease of Online booking • Gate location • Food and drink • Online boarding • Seat comfort • Inflight entertainment
n = 6	• Inflight wifi service • Departure/Arrival time convenient • Ease of Online booking • Gate location • Food and drink • Online boarding
n = 4	• Inflight wifi service • Departure/Arrival time convenient • Ease of Online booking • Gate location
n = 2	• Inflight wifi service • Departure/Arrival time convenient

To build the models, Python have been used in this report to build both of the machine learning model. Firstly, several package have to be included to the program to build the machine learning model and to plot the graph. The packages needed are pandas, numpy, sklearn and seaborn as shown as below.

```
import pandas as pd
import numpy as np
from sklearn import linear_model
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
import seaborn as sns

#Package for Plotting
import matplotlib.pyplot as plt # library for plotting
%matplotlib inline
from sklearn import metrics
```

After that, both test and train dataset have been read and assigned to a variable. Since the objective attribute ‘satisfaction’ is in string type, label encoding will be needed to change the value ‘neutral or dissatisfied’ to 0 and ‘satisfied’ to 1. The process of reading dataset and label encoding is performed by using python as shown as below:

```
#Reading the Dataset
df = pd.read_csv('D:/train.csv')
df_test = pd.read_csv('D:/test.csv')

#Convert the Satisfaction to 0 for neutral of dissatisfied and 1 for satisfied
df['satisfaction_cleaned'] = np.where(df['satisfaction']=="neutral or dissatisfied",0,1)
df_test['satisfaction_cleaned'] = np.where(df_test['satisfaction']=="neutral or dissatisfied",0,1)
```

The independent variables from train dataset have been assigned to 'x_train' and from test dataset have been assigned to 'x_test', where the dependent variable from train dataset have been assigned to 'y_train' and from test dataset have been assigned to 'y_test'. After that, the multiple linear regression model have been built and its performance score have been recorded, the below diagram shows the building of multiple linear regression model and it score when using 2 independent attributes:

```
##Multiple Linear Regression Method
model = linear_model.LinearRegression()
model.fit(x_train,y_train)
print("The prediction accuracy is: {0:2.2f}{1:s}".
      format(model.score(x_test,y_test)*100,"%"))
```

```
The prediction accuracy is: 10.84%
```

By adding more independent variables from 2 to all the variables into the multiple linear regression model, each of the model build's performance have been recorded and listed in the below table:

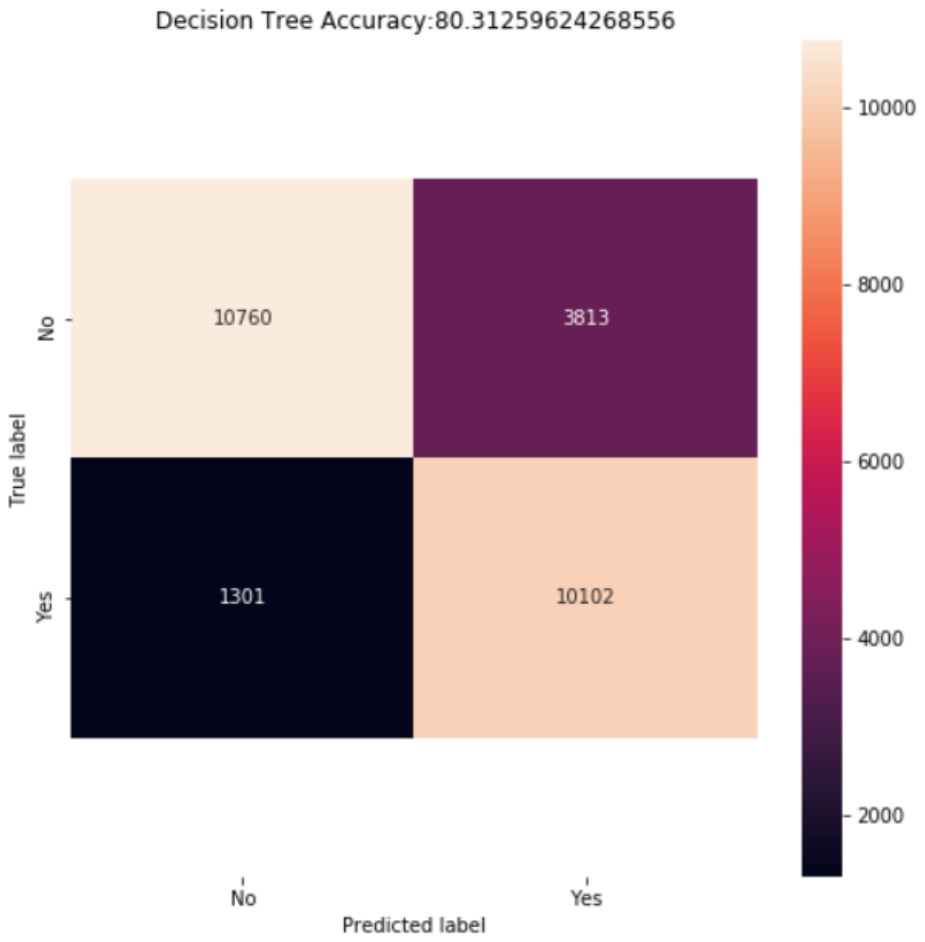
Number of Variables	N = 15 (all)	N = 8	N = 6	N = 4	N = 2
Model Performance (%)	39.23	33.92	28.03	11.18	10.84

From the above table we can observed that as the number of variables used increase, the multiple linear model performance have also been increase accordingly. The greatest improve of performance occurs between 4 numbers of independent variables and 6 numbers of independent variables. This data proved that the model performance does have a better performance when more independent variables are used in this dataset, and so the null hypothesis for second hypothesis is accepted for Multiple Linear Regression model.

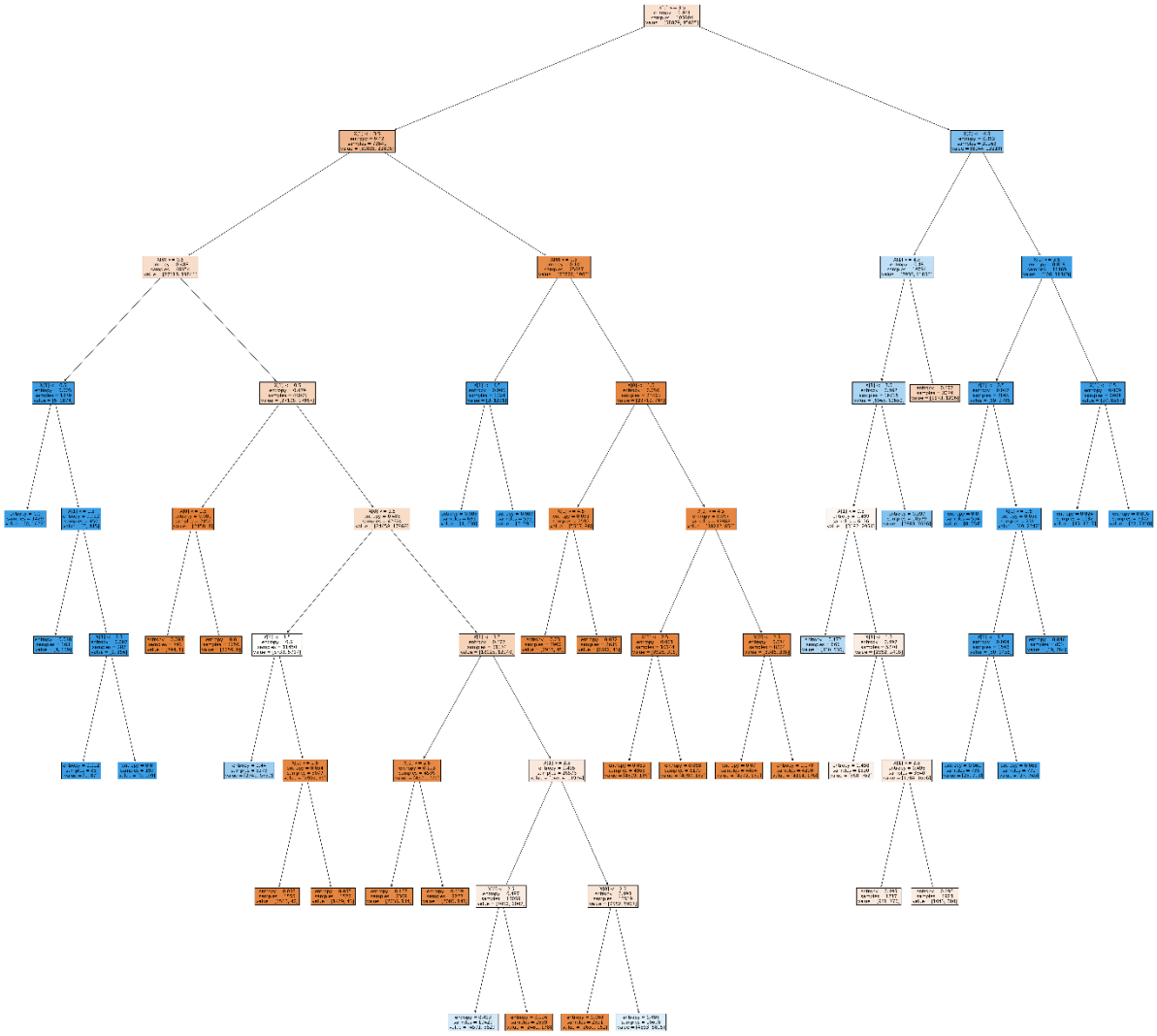
To build the decision tree, there two type of criterion can be used, whiare are the Gini criterion and the Entropy criterion. After test to build the model using these two different criterion, it seems that both criterion will produce the same model performance and no differences between them, therefore the entropy criterion have been used in this part. The below shows the performance of decision tree model and its confusion matrix by using 2 number of variables in building the model.

```
##Decision Tree
# input the decision tree classifier using "entropy" & train the model
dtree = DecisionTreeClassifier(criterion = 'gini').fit(x_train, y_train)
#The acccuracy will be different for each time

# predict the classes of new, unseen data
predict = dtree.predict(x_test)

print("The prediction accuracy is: {0:2.2f}{1:s}".format(dtree.score(x_test,y_test)*100,"%"))

# Creates a confusion matrix for predicted Item and actual data
cm = confusion_matrix(y_test, predict)

# Transform to dataframe for easier plotting
cm_df = pd.DataFrame(cm, index = ['No','Yes'],
     columns = ['No','Yes'])

# plot the confusion matrix
plt.figure(figsize=(8,8))
ax= sns.heatmap(cm_df, annot=True, fmt='g')
bottom, top = ax.get_ylim()
ax.set_ylim(bottom + 0.5, top - 0.5)
plt.title("Decision Tree Accuracy:" + str(dtree.score(x_test,y_test)*100))
plt.ylabel('True label')
plt.xlabel('Predicted label')
```

```
The prediction accuracy is: 80.31%
```



The decision tree diagram when dealing with 2 of the independent variables have also been built and show as the below diagram.



By adding more independent variables from 2 to all the variables into the decision tree model, each of the model build's performance have been recorded and listed in the below table:

Number of Variables	N = 15 (all)	N = 8	N = 6	N = 4	N = 2
Model Performance (%)	92.72	92.57	91.30	88.49	80.31

From the above table we can observed that as the number of variables used increase, the decision tree performance have also been increase accordingly. The greatest improve of performance occurs between 2 numbers of independent variables and 4 numbers of independent variables. This data proved that the model performance does have a better performance when more independent variables are used in this dataset, and so the null hypothesis for second hypothesis is accepted for Decision Tree models.

5.0 Result and Analysis

By building both decision tree and multilinear regression model and including different numbers of independent variable, their performance have been recorded and summarize as in the below table:

Number of Variables/ Performance (%)	**N = 15 (all)**	**N = 8**	**N = 6**	**N = 4**	**N = 2**
Decision Tree	92.72	92.57	91.30	88.49	80.31
Multiple Linear Regression	39.23	33.92	28.03	11.18	10.84

From the above tables, the performance of Decision Tree models are much more better than the multiple linear regression model in any circumstances. The reason behind this is because the independent variables are in whole number and have a lower range of value, this have led to independent variable perform like categorical variable instead of numeric variable, and therefore decision tree will act more effective in prediction as compare to the multiple linear regression method.

By referring to the first hypothesis, since Decision Tree Model have a better performance then Multiple Linear Regression Model in this dataset, therefore the H_0 hypothesis is accepted. The machine learning method suggest to be used in this dataset will be the Decision Tree Model.

Besides, as mentioned in the experiment part the performance of model increase for both Decision Tree and Multiple Linear regression as the number of independent variables uses increases. This comes to a conclusion that all the independent variables in survey form should be used as the model performance does have a better performance when more independent variables are used in this dataset. Therefore, the H_0 for the second hypothesis is accepted.

6.0 Conclusion

In conclusion, two machine learning model which are the decision tree and the multiple linear regression model have been built to examine the passenger satisfaction level. Since more independent variables in the survey form led to a higher performance level for the machine learning model, all of the survey attribute should be used in building the machine learning model. Besides, decision tree model also have outperformed the multiple linear regression model since the independent variable have a lower range and is a whole number.

7.0 References

[1] Hongwei Jiang, Yahua Zhang, "An investigation of service quality, customer satisfaction and loyalty in China's airline market," *Elsevier: Journal of Air Transport Management,* vol. 57, pp. 80-88, 2016.

[2] T. Klein, "Airline Passenger Satisfaction: What factors lead to customer satisfaction for an Airline?," 11 02 2020. [Online]. Available: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction.

[3] T. Ryan Hoens, Nitesh V. Chawla, "IMBALANCED DATASETS: FROM SAMPLING TO CLASSIFIERS," Department of Computer Science and Engineering, The University of Notre Dame, Notre Dame, 2013.

[4] D. Varghese, "Comparative Study on Classic Machine learning Algorithms: Quick summary on various ML algorithms," Towards Data Science, 7 Dec 2018. [Online]. Available: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=only%20linear%20solutions.-,When%20there%20are%20large%20number%20of%20features%20with%20less%20data,are%20better%20than%20linear%20regression.. [Accessed 16 11 2020].