Comparative Analysis of Parametric and Non-Parametric Machine Learning Algorithms for Airline Passenger Satisfaction

Kabir Musa Muhammad
Computer Engineering Department
Binyaminu Usman Polytechnic, Hadejia, Jigawa, Nigeria
ringimkabir@bupoly.edu.ng

Hassan Adamu
Computer Science Department
Binyaminu Usman Polytechnic, Hadejia, Jigawa, Nigeria
hassanadamu@bupoly.edu.ng

***Abstract*— The study aims to solve a classification problem of airline passenger satisfaction using two supervised machine learning algorithms: *parametric (Naïve Bayes) and non-parametric (Decision Tree)*. A new model was created with the help of these two algorithms to predict the customer's satisfaction to see whether the customers are satisfied or dissatisfied based on their previous experience from the services provided by the airlines. However, when a dataset is balanced, an accuracy measure is enough to compare the performance of models but here in this assignment, other performance evaluation metrics such as precision, recall and f1-score were used to measure the results of the models in order to obtain and compare different accuracies. Meanwhile, the results obtained from the experiments shows that the non-parametric classifier (Decision Tree) outperform the parametric classifier with 90% accuracy as well as 89% on both precisions, recall and f1-score using weighted and macro average, while the parametric (Naïve Bayes) achieved 84% accuracy. Conclusively, the results reveals that the customers were satisfied with the airline operations because the overall results of precision, recall and f1-score on the targeted class label (satisfied) surpasses the class label (neutral/dissatisfied).**

Keywords*— Parametric, Non-parametric, Naïve Bayes, Decision Tree

I. INTRODUCTION

A. Naïve Bayes Model (Parametric)

Naïve Bayes is a supervised machine learning algorithm which classifies data using training and testing sets with the help of labelled class as a target feature. Naïve Bayes is a rule-based and a probabilistic model that plays an important role in the field of machine learning classification techniques. The rule-base is the mathematical approach to help in creating models in different methods through data understanding and exploration. Many researchers discovered Naïve Bayes model as proficient and effective on both small and huge dataset as it usually produces better classification results (Badugu and Kolikipogu (2020).

A recent study by Filippakis et al. (2020), Sahoo et al. (2017) and Vanhoenshoven et al. (2016) summarized the concept of Naïve Bayes algorithm as a model based on the Bayesian Theorem. The theorem of Bayes is used to calculate the chances of an incident if the chances of an incident that has already happened is given. This model believes that each feature is independent and contributes similarly to the result. The probability of a record for each class is calculated by the chances of all the values of the function. The record is expected to be the class that results in the greatest probability.

The reason behind chosen this supervised machine learning algorithm in this assignment is because of its efficiency, easy and very fast in terms of training model as well as classification and prediction of the testing data. Also, Naïve Bayes performance is highly incredible when dealing with dataset for both categorical and non-categorical input variables because of its representation of probabilistic information.

B. Decision Tree Model (Non-parametric)

Decision tree is another supervised machine learning algorithm that falls under non-parametric models. The classifier follows conditional statement approach of if-then-else rules. The model creates a decision tree in which the leaf nodes represent the class name, while the node without any of the leaf in the tree structure represents the decision point Hussain and Morah (2020) as for the leaf node type, a specific record is predicted, and the training dataset is divided into many subgroups until all the subgroups belongs to a single class. However, different features or techniques are used to choose the decision nodes by reducing the tree in order to suit the test data. In their recent study.

Abdelmegeid et al. (2020) describe decision tree as a technique that can be quickly understood because of the tree structure where it could be graphically represented, which therefore takes less time to understand and define the data based on the input variables at the bottom of the tree branches.

The reason why this model was selected in this assignment as non-parametric classifier is because of its simplicity and easiness in terms of training and classifications irrespective of the number of features or attributes in a dataset. Also, the following decision tree algorithms are the most common and widely used under the family of this model which are highly efficient, the algorithms are J48 algorithms, ID3 (Iterative Dichotomiser) and C4.5 algorithms (Keeley, C., 2017).

The dataset used in this assignment is an Airline Passenger Satisfaction downloaded from Kaggle and uploaded by TJ Klein (2020). The dataset was used purposely to classify the airline passengers' satisfaction based on their previous experiences.

1.2 Dataset Class Distribution

The dataset contains 103,905 instances from 0 to 103,904 with total number of 25 columns (features). Also, the dataset is divided into separate CSV files as training and testing set in which 80% of the data goes for training and 20% for testing. The table 1 shown represent the dataset description, while table 2 shows the total number of the features that the dataset contains including their data type and their brief description. Meanwhile, the total number of instances of the training set is 103,904 while that of the testing set is 25,976. However, the dataset was labeled into two classes (satisfied and neutral/dissatisfied), these labeled classes are under the target attribute "satisfaction".

Table 1. Dataset description for both training and testing CSV files

Characteristics	Multivariate
Attributes/Features	Continuous and Nominal
Number of Instances	103904
Number of Features	25
Target attribute	1
Attribute data type	Integer (19), Float (1), Object (5)
Missing Values	There were 310 (0.3%) missing values on training set and 83 (0.32%) on testing set for "Arrived Delay in Minutes" feature. The remaining 24 features has no missing value.

1.3 Dataset Characteristics (Balanced or Unbalanced)

As mentioned in the previous section that the dataset was separated into different CSV files as training and testing set. Therefore, table 3 described the sample size of the training set on the target attribute (satisfaction).

Table 2. Sample size description of the training set

S/N	Class category	Observations/instances	Percentage
1	Satisfied data	45,025	43.3%
2	Neutral/ dissatisfied	58,880	56.7%
	Differences	13,855	13.3%
	Total	103,905	100%

Table 3 indicated that the total number of observations or instances assigned to the class category "satisfied" is 45,025 equivalents to 43.3% while that of neutral or dissatisfied is 58,880 (56.7%) out of the total number of the instances in the dataset which is 103,094 (100%). This means that the category of neutral or dissatisfied have the highest number of observations with 13,855 (13.3%) differences. Table 4 described the sample size of the testing set on target attribute (satisfaction).

Table 3. Attributes (features) description for both training and testing CSV files

S/No	Features Column)	Data Type	Description
1	Unnamed column	int64	Serial number from 0 to 24
2	id	int64	Flight identification number for each passenger
3	Gender	object	Gender of the passengers (Female, Male)
4	Customer Type	object	The customer type (Loyal customer, disloyal customer)
5	Age	int64	The actual age of the passengers
6	Type of Travel	object	Purpose of the flight of the passengers (Personal Travel, Business Travel)
7	Class	object	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
8	Flight distance	int64	The flight distance of this journey
9	Inflight Wi-Fi service	int64	Satisfaction level of the inflight Wi-Fi service (0: Not Applicable; 1-5)
10	Departure/Arrival time convenient	int64	Satisfaction level of Departure/Arrival time convenient
11	Ease of Online booking	int64	Satisfaction level of online booking
12	Gate location	int64	Satisfaction level of Gate location
13	Food and drink	int64	Satisfaction level of Food and drink
14	Online boarding	int64	Satisfaction level of online boarding
15	Seat comfort	int64	Satisfaction level of Seat comfort
16	Inflight entertainment	int64	Satisfaction level of inflight entertainment
17	On-board service	int64	Satisfaction level of On-board service
18	Leg room service	int64	Satisfaction level of Leg room service
19	Baggage handling	int64	Satisfaction level of baggage handling
20	Check-in service	int64	Satisfaction level of Check-in service
21	Inflight service	int64	Satisfaction level of inflight service
22	Cleanliness	int64	Satisfaction level of Cleanliness
23	Departure Delay in Minutes	int64	Minutes delayed when departure
24	Arrival Delay in Minutes	float64	Minutes delayed when Arrival
25	Satisfaction	object	Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

Table 4. Sample size description of the testing set

S/N	Class category	Observations/Instances	Percentage
1	Satisfied data	11,403	43.9%
2	Neutral/ dissatisfied	14,574	56.1%
	Differences	3,171	12.2%
	Total	25,977	100%

Table 4 indicates that the total number of instances assigned to the class category "satisfied" is 11,403 equivalents to 43.9% while that of neutral or dissatisfied is 14,574 (56.1%) out of the total number of the instances in the dataset which is 25,977 (100%). This means that the category of neutral or dissatisfied have the highest number of observations with 13,855 (12.2%) differences.

However, comparing the differences of the sample size between the two CSV files which is training and testing set on the target attribute (satisfaction), the variation is just 1.1% thus makes the dataset to be in the ratio of balanced dataset. Meanwhile, a balanced dataset always yields a better results accuracy compared to unbalanced dataset because, for an unbalanced dataset to produce good results accuracy, it has to undergo different sampling techniques such as under-sampling and over-sampling techniques (Yıldırım, P., 2016).

II. FEATURE SELECTION

In machine learning, feature selection is a common pre-processing technique which is a procedure of finding and choosing a subset of features related to a target feature in a dataset that eventually helps in producing a better model (Miao, J., and Niu, L., 2016) redundant and irrelevant features in a dataset could make a classifier to have a lesser good performance in results accuracy, while a well-organized dataset with relevant features helps classifiers for easy learning during training and prevent them from overfitting and underfitting so as to achieve better generalization ability which would eventually help in producing a good and accurate results (Shuyang, et al., 2016).

There are three most popular and common feature selection techniques widely used in different machine learning research, these techniques are: filter-based approach, wrapper-based approach and embedded approaches. The first approach is filtering method which is use at a pre-processing stage before model classification, and the main aim of this approach is to find out which feature or attributes are important and dependent to the classifier. The second method is wrapper-based approach which is used to search and identify most important subset of attributes or features that could produce a better and good results accuracy of a model classification. The third method is an embedded approach which is used to perform a feature selection procedure within the machine learning model itself, it is called embedded because the techniques perform the selection at the time of model training (Chuan, 2017). The above-mentioned approaches usually applied based on the nature of the dataset and the machine learning research domain. Therefore, a filter-based feature selection technique is used in this assignment because the models selected are supervised machine learning models. Figure 1 shows supervised feature selection technique.

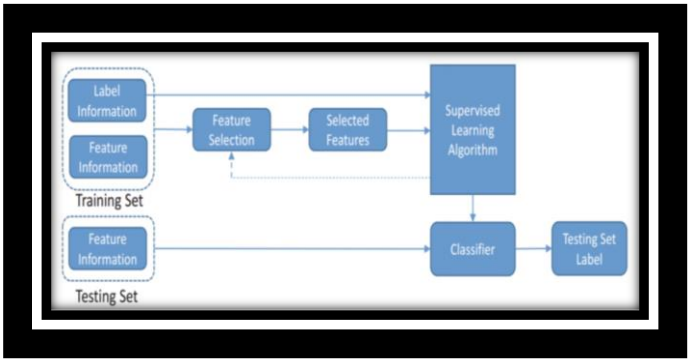

Figure 1. Supervised feature selection technique

As mentioned earlier in Section 1, the airline passenger satisfaction dataset was divided into two separate CSV files for training and testing. Therefore, the training data would be used to train the learning classifiers and create a new model while the testing set is an unseen data that would be used to evaluate the performance of the classifiers. However, considering the characteristics of the dataset, there are a lot of categorical attributes/features that need to be encoded and transform into numerical data including the target attribute (satisfaction).

III. PERFORMANCE EVALUATION METRIC

The performance evaluation metrics to be used in this assignment are four, these are: Accuracy, Precision, Recall and F1-measure to help in identifying and comparing the best classifier amongst selected parametric and non-parametric models. Table 5 described them with their respective formula. Also, the results and discussion would be presented in section 4.

However, another important machine learning technique used is confusion matrix. It is one of the powerful techniques used to evaluate the performance of machine learning algorithms with the help of its four terms. These terms are true positive (TP), true negative (TN), false positive (FP) and false negative (FN). All the above-mentioned performance evaluation matrices (accuracy, precision, recall and F1-score) used to be evaluated using this technique. The following Table 6 is an illustration of how confusion matrix is being calculated.

Table 5. Performance Evaluation Metrics

Metrics	Formula	Evaluation Focus
Accuracy (Acc)	$\frac{tp+tn}{tp+fp+tn+fn}$	The accuracy measures the ration of correct classification or predictions over total number of instances evaluated
Precision (P)	$\frac{tp}{tp+fp}$	This would be used to measure the positive patterns that are correctly classified from the total predicted patterns in a positive class
Recall (R)	$\frac{tp}{tp+tn}$	This would be used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2*p*r}{p+r}$	This represents the harmonic mean between precision and recall values

Table 6. Confusion Matrix

Actual Values	Predicted Values: Class Labels	Satisfied	Neutral/Dissatisfied	Total to calculate recall
	Satisfied	102	1	103
	Neutral/Dissatisfied	3	124	127
Total to calculate precision		105	125	230

Key:

Grey color represent the actual values and the predicted values

Blue Represent the target variable which is the labeled class in the dataset

Red represent the correctly classified instances

White represent the incorrectly classified instances or observations

Yellow color represents the calculation of precision and recall which helps in calculating the f1-score as the harmonic mean of the precision and recall

IV. EXPERIMENTS AND ANALYSIS

In this section, the experimental settings, dataset exploration and the experimental results were presented.

4.1 Experimental Settings

The experiments conducted in this assignment is shown in Table 7.

Table 7. Experimental settings

Experiment Number	Algorithms
Experiment 1	Dataset Exploration and visualization
Experiment 2	Naïve Bayes (Parametric classifier)
Experiment 3	Decision Tree (Non Parametric classifier)

4.2 Dataset Exploration (Experiment No. 1)

A. Dataset Class Distribution

This is the structure of the dataset showing only five rows out of the total number of the instances for both training set and the testing set which contain 25 columns (features) as presented in Figure 2.

B. Dataset Missing Values

As part of the pre-processing of the dataset, it was discovered that there are 310 missing values equivalents to 0.3% on the training set as well as 83 (0.2%) for the testing set. These missing values were all from the "Arrival Delay in Minutes" feature. The remaining features were not affected by any missing value.

C. Dataset Transformation

After discovering the missing values on one of the attributes "Arrival Delay in Minutes", the missing values were filled up using median calculation of the entire variables in the column. Also, two features were dropped "Unnamed" and "id" because they are irrelevant attributes in the dataset. However, four features "Gender", "Customer Type", "Type of Travel" and "Class" were transformed from string characters to binary numbers of ranges from 0-2 as shown in figure 3.

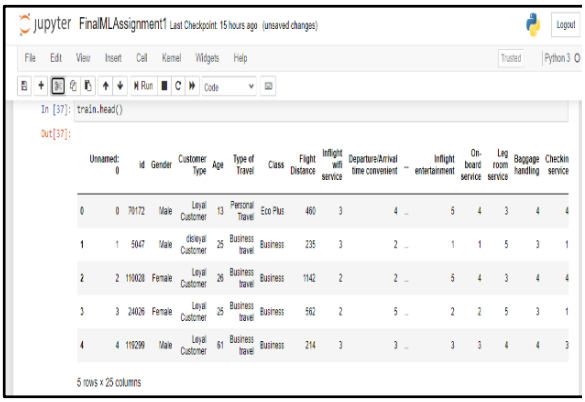

Figure 3. Dataset transformation

D. Features Correlation

The dataset was further explored to identify the features that correlated well on the target attribute "satisfaction". In this regard, heatmap was used as a color-coding technique that shows the graphical representation of values in the dataset. Also, the stronger the color representation, the lower the correlated magnitude while, the lighter or brighter the color representation, the highest the correlation magnitude as shown in Figure 4.

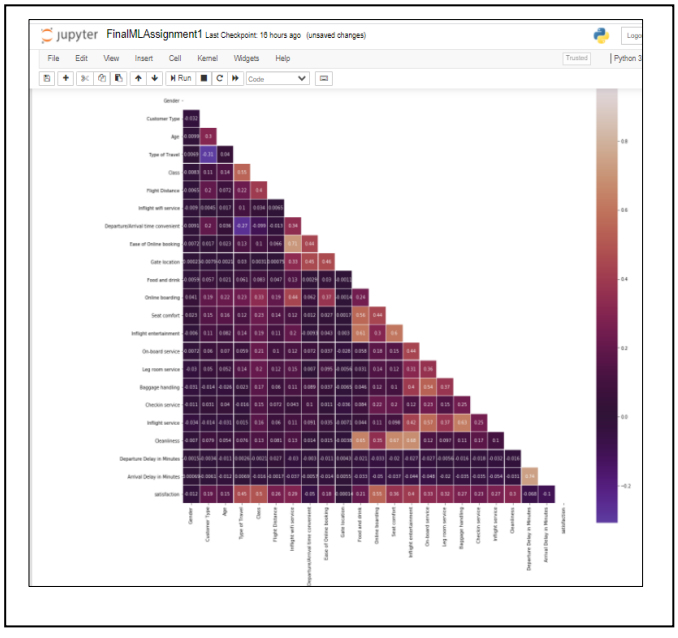

Figure 4. Feature Correlation

Looking at the figure 5, there are three best correlated features on the target attribute (satisfaction), and these features are “Type of Travel”, “Class” and “Online boarding”. Meanwhile, there are five worst correlated features on the targeted attributes, these are “Gender”, “Departure/Arrival time convenient”, “Gate location”, “Departure Delay in Minutes”, and “Arrival Delay in Minutes” as shown in figure 7.

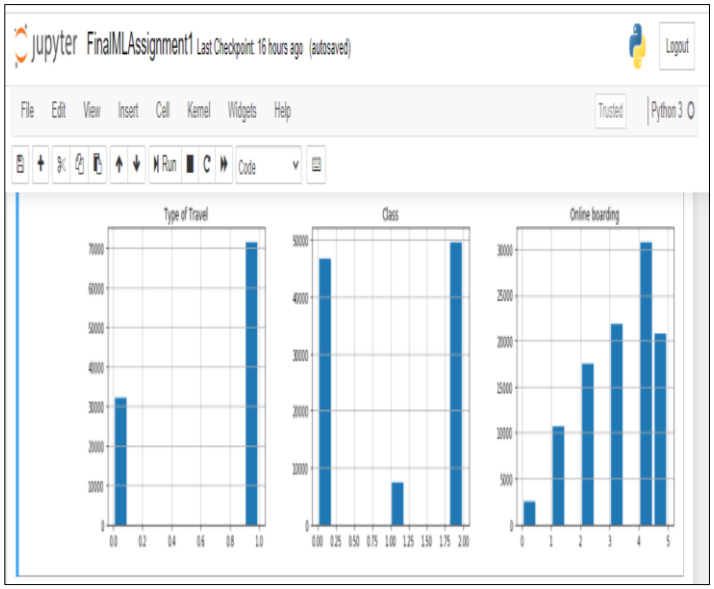

Figure 5. Best correlated features

However, the worst correlated features as shown in figure 6 were also dropped in the dataset to help in improving the performance of the classifiers. Therefore, after dropping the features from the dataset, the shape of the training set becomes (103904 instances and 18 features) so also the shape of the testing set becomes (25976 instances and 18 features).

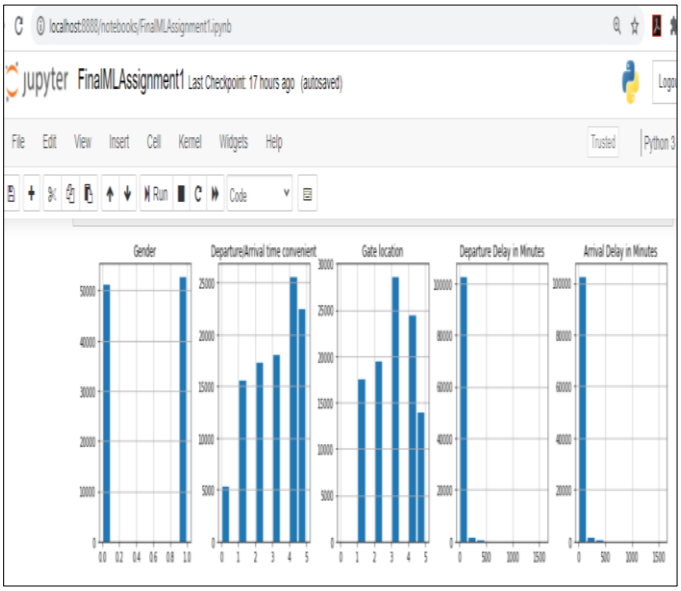

Figure 6. Worst correlated features

4.3 Models Evaluation, Results and Discussion

A. Experiment No. 2 (Naïve Bayes)

This experiment was conducted to evaluate the performance of Naïve Bayes Model based on airline passenger satisfaction using binary classification of 0 and 1 (satisfied and neutral/dissatisfied). Table 8 depicted the results obtained from the experiment where the model achieved an accuracy of 83% with 0.845 precision on satisfied class label and 0.807 precision on neutral/dissatisfied class label. Also, the model, achieved 0.847% recall for satisfied class label and 0.815 recall on neutral/dissatisfied class label. The f1-score is the harmonic mean of precision and recall in which the model scored 0.851 on satisfied class label and 0.811 on neutral/dissatisfied class label. Also, figure 7 graphically reveals the results and figure 8 shows the confusion matrix.

Table 8. Performance Evaluation of Multinominal Naïve Bayes Model

Class Label		**Precision**	**Recall**	**F1-Score**
Satisfied (labelled as 0)		0.854	0.847	0.851
Neutral/Dissatisfied (labelled as 1)		0.807	0.815	0.811
	Macro Average	0.830	0.831	0.831
	Weighted Average	0.834	0.834	0.834
	Model Accuracy	**83%**		

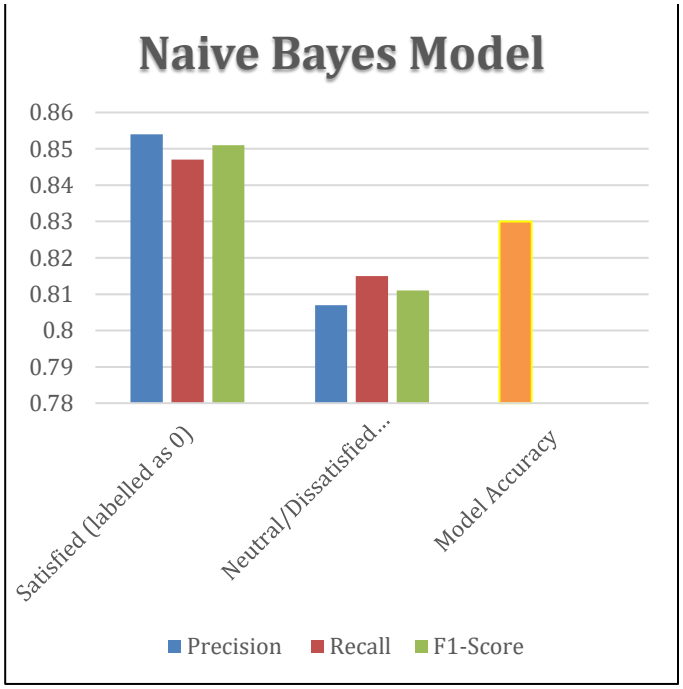

Figure 7. Naïve Bayes experimental results

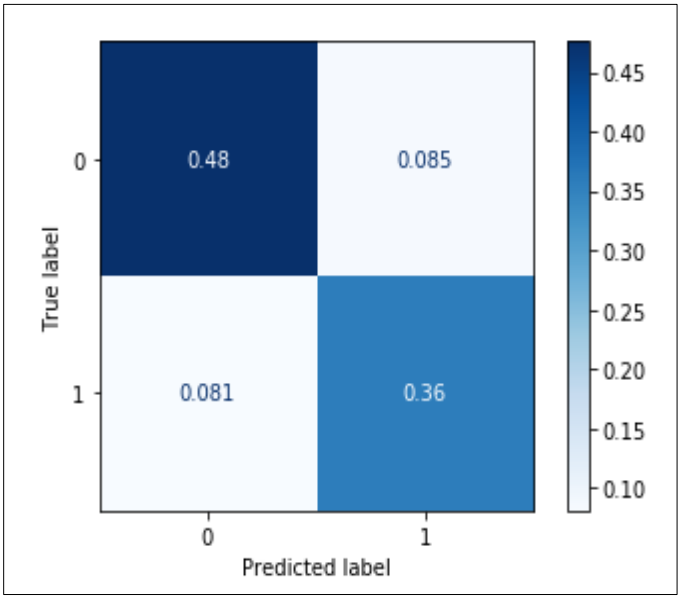

Figure 8. Confusion Matrix of Naïve Bayes

The above confusion matrix reveals that the corrected classified instances is 84% and the total number of the incorrect instances is 16%. This shows that the model performance was good for this classification.

B. Experiment No. 2 (Decision Tree)

This experiment was conducted to evaluate the performance of Decision Tree Model based on airline passenger satisfaction using binary classification of 0 and 1 (satisfied and neutral/dissatisfied). Table 9 depicted the results obtained from the experiment where the model achieved an accuracy of 83% with 0.900 precision on satisfied class label and 0.894 precision on neutral/dissatisfied class label. Also, the model, achieved 0.919 recall for satisfied class label and 0.896 recall on neutral/dissatisfied class label. The f1-score is the harmonic mean of precision and recall in which the model scored 0.909 on satisfied class label and 0.881 on neutral/dissatisfied class label. Therefore, Figure 9 shows the confusion matrix.

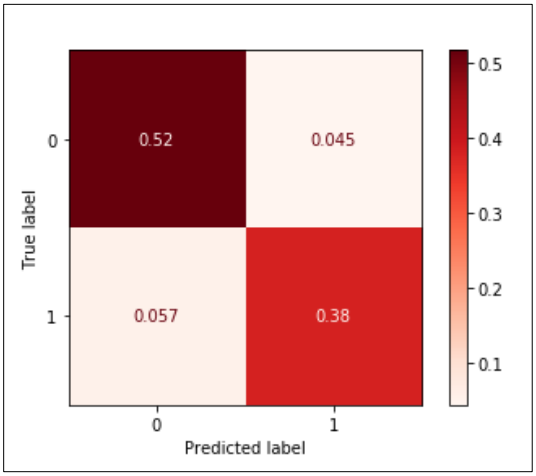

Figure 9. Confusion Matrix of Decision Tree

The above confusion matrix reveals that the corrected classified instances is 90% and the total number of the incorrect instances is 10%. This shows that the model performance was good for this classification.

Table 9. Performance Evaluation of Decision Tree Model

Class Label		**Precision**	**Recall**	**F1-Score**
Satisfied (labelled as 0)		0.900	0.919	0.909
Neutral/Dissatisfied (labelled as 1)		0.894	0.896	0.881
	Macro Average	0.897	0.895	0.896
	Weighted Average	0.898	0.897	0.897
	Model Accuracy	**90%**		

C. Models Comparisons

This section aims to make comparison of the results obtained from the previous experiments regarding classification of airline passenger satisfaction on Naïve Bayes as parametric and Decision Tree as non-parametric classifiers. The results of the performance evaluation metrics in this section are based on the weighted and macro average of all the models. Therefore, table 10 shows the comparison of the models.

Table 10. Models Comparison

Models	**Evaluation Metrics on Weighted Average**			
	Precision	**Recall**	**F1-Scores**	**Accuracy**
Naïve Bayes	0.830	0.831	0.831	83%
Decision Tree	0.898	0.897	0.897	90%
Evaluation Metrics on Macro Average				
Naïve Bayes	0.834	0.834	0.834	83%
Decision Tree	0.897	0.895	0.896	90%

From the experimental results of the above table 10, Decision Tree as non-parametric classifier outperform Naïve Bayes with 90% accuracy as well as 89% on both precisions, recall and F1-score on weighted and macro average. The reason why Decision Tree (non-parametric) outperform Naïve Bayes (parametric) model is that; non-parametric machine learning classifiers does not make an assumptions about the output variables though the models mostly require more data and it is slow in training data yet they produced higher results accuracy, unlike parametric models which is faster in training data and does not require much data but mostly produce less results accuracy.

V. CONCLUSION

This assignment focused on the classification of airline passenger satisfaction using Naïve Bayes as parametric classifier and Decision Tree as non-parametric classifier. The dataset was collected from Kaggle which contains two separate CSV files with total number of 103,904 instances or observations for the training set, and 25,976 for the testing set. Both training and

testing set contains 25 columns (features) which drastically reduced to 18 features by removing some irrelevant and unwanted features to help in better performance of the classifiers.

However, the experimental results reveal that Decision Tree as a parametric classifier outperform the non-parametric classifier (Naïve Bayes) with 90% accuracy as well as 89% on both precisions, recall and F1-score on weighted and macro average. Conclusively, the overall results of all the experiments conducted indicated that the level of satisfaction from the passengers based on their previous experiences surpasses the level of dissatisfaction on both precisions, recall and f1-score across all the models.

REFERENCES

[1] TJ Klein (2020). Airline Passenger Satisfaction. Accessed on 21 October 2020. Available [online] at https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

[2] Pınar Yıldırım (2016). Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes. The International Workshop on Data Mining for Decision Support (DMDMS 2016). Procedia Computer Science 83 (2016) 1013 – 1018. doi: 10.1016/j.procs.2016.04.216

[3] Miao, Jianyu and Lingfeng Niu (2016), "A survey on feature selection." Procedia Computer Science, 91, 919–926.

[4] Shuyang, Gao, Greg Ver Steeg, and Aram Galstyan (2016), "Variational information maximization for feature selection." In Advances in Neural Information Processing Systems, volume 29, 487–495.

[5] Chuan, Liu, Wenyong Wang, Qiang Zhao, Xiaoming Shen, and Martin Konan (2017), "A new feature selection method based on a validity index of feature subset." Pattern Recognition Letters, 92, 1–8.

[6] Keeley, Crockett, Annabel Latham, and Nicola Whitton (2017), "On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees." International Journal of Human-Computer Studies, 97, 98–115.

[7] Sahoo, D., Liu, C., Hoi, S.C.H.(2017). Malicious URL detection using Machine Learning: a survey. Retrieve from: arXiv:1701.07179v2 (2017)

[8] Vanhoenshoven, F., Napoles, G., Falcon, R., Vanhoof, K., Koppen, M. (2016). Detecting malicious URLs using machine learning techniques. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI). Retrieved from: https://doi.org/10.1109/ssci.2016.7850079