Comparison of Supervised Learning Algorithms to Predict on-time Delivery of Packages in an E-Commerce Shipping Data and Airlines Passenger Satisfaction in an Airline Passenger Satisfaction Survey Data

AOL Report Data Mining Group 9

By :

2602168406 - Raphaelle Albetho Wijaya

2602170934 - Christopher Jovison

1. INTRODUCTION

1.1 Background of the Study

Package delays are a problem that is often experienced by almost every individual, where in some cases, this is not a big problem. But in some other cases when the goods needed must be there as soon as possible, the delay in package delivery is very problematic. Late package delivery can reduce user satisfaction with a company that provides delivery services. Therefore, it is important to analyze the probability and risk of delays.

In this research, our group aims to apply Machine Learning techniques to make predictions about package delays. This aims to improve the efficiency and accuracy of package delivery, and can be a reference to improve the performance of the service provider company.

1.2 Problem Definition

As explained in the background of the study, delays in package delivery can be a big deal when an item is needed urgently. There are so many factors that can affect the delivery conditions of a package such as shipping method, package weight, and so on. These factors may be difficult or even impossible for humans to analyze at a glance. If not considered seriously, the problem of delayed package delivery can become a nightmare for the service provider company in the future. One of the problems that can arise and have fatal consequences is such as a decrease in the level of public trust in shipping service providers.

1.3 Solution to the Problem

The purpose of this research is to provide a solution in the form of applying machine learning and or data mining models to predict the timeliness of package delivery. We use supervised learning algorithms such as K-Nearest Neighbor, Naive Bayes, Support Vector Machine, and Random Forest and compare their performance to find which model is the best for predicting package delivery timeliness.

In this study, we will also consider factors that can directly affect the timeliness of package delivery. By considering these factors, our model will be built to be able to learn and predict accurately.

1.4 Other Studies

Research titled “Predicting On-time Delivery in the Trucking Industry” conducted by Rafael Duarte Alcoba and Kenneth W. Ohlund said that predictive models can be tailored to

optimize other, more specific metrics. For a company that focuses on achieving a high service level, the minimization of missed delays is critical. Although it is possible to improve overall misclassification by adding some extra variables, the researcher avoid it. The researcher find that adding variables without very high explanatory power adds complexity, reduces robustness, and can lead to overfitting. The model, developed using six explanatory variables with statistical significance, results in a 76.4% resource reduction while incurring an impactful error of 2.4%

In another research titled “Predicting Delivery Time and Estimating Shipment Delays with Machine Learning (Supply Chain and Logistics Series)” conducted by Lyron Foster said that predicting delivery time and estimating shipment delays with machine learning can be a valuable tool for businesses in the logistics industry. It allows them to make data-driven decisions, optimize their operations, and provide better service to their customers.

2. DATA

2.1 Dataset Used

The dataset we use is the "E-Commerce Shipping Data" dataset from kaggle. The dataset consists of 10999 data rows and 12 features (see table below for details).

Feature	Description
ID	ID Number of Customers.
Warehouse block	The Company have big Warehouse which is divided in to block such as A,B,C,D,E.
Mode of shipment	The Company Ships the products in multiple way such as Ship, Flight and Road.
Customer care calls	The number of calls made from enquiry for enquiry of the shipment.
Customer rating	The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best).
Cost of the product	Cost of the Product in US Dollars.
Prior purchases	The Number of Prior Purchase.
Product importance	The company has categorized the product in the various parameter such as low, medium, high.
Gender	Male and Female.
Discount offered	Discount offered on that specific product.
Weight	It is the weight in grams.
Reached on time	It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

By considering these parameters, the model will be able to provide more accurate predictions regarding the timeliness of package arrival. Each feature is selected based on factors that directly affect the timeliness of package delivery.

2.2 Pre-Processing

Here are the details of the initial to final stages of pre-processing that our group applied to our dataset:

A. Drop Unused Column : Since the ID and Gender column is not needed, we decided to drop the ID and Gender column.

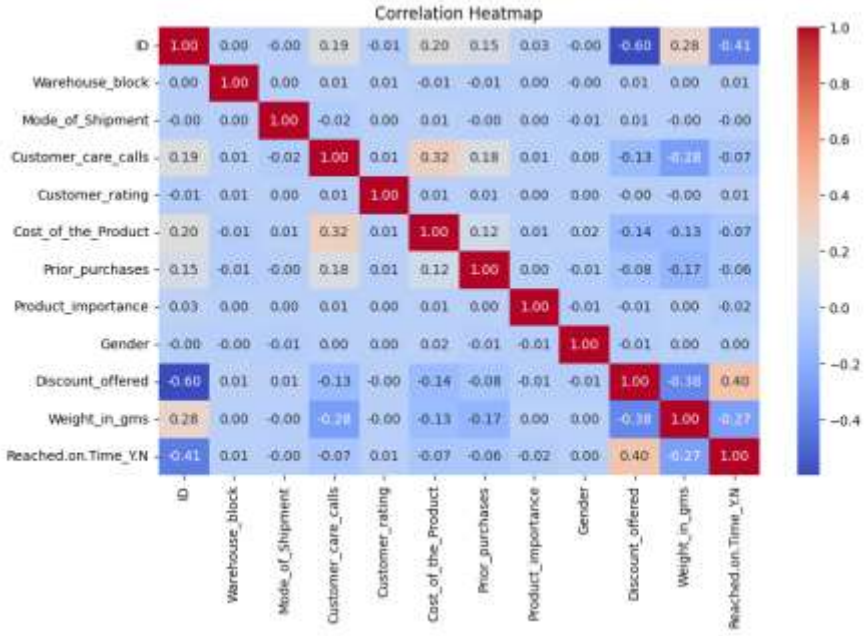
B. Handle Missing Value : Since the dataset we use is clean and has no missing values, we do not fill in the missing values.

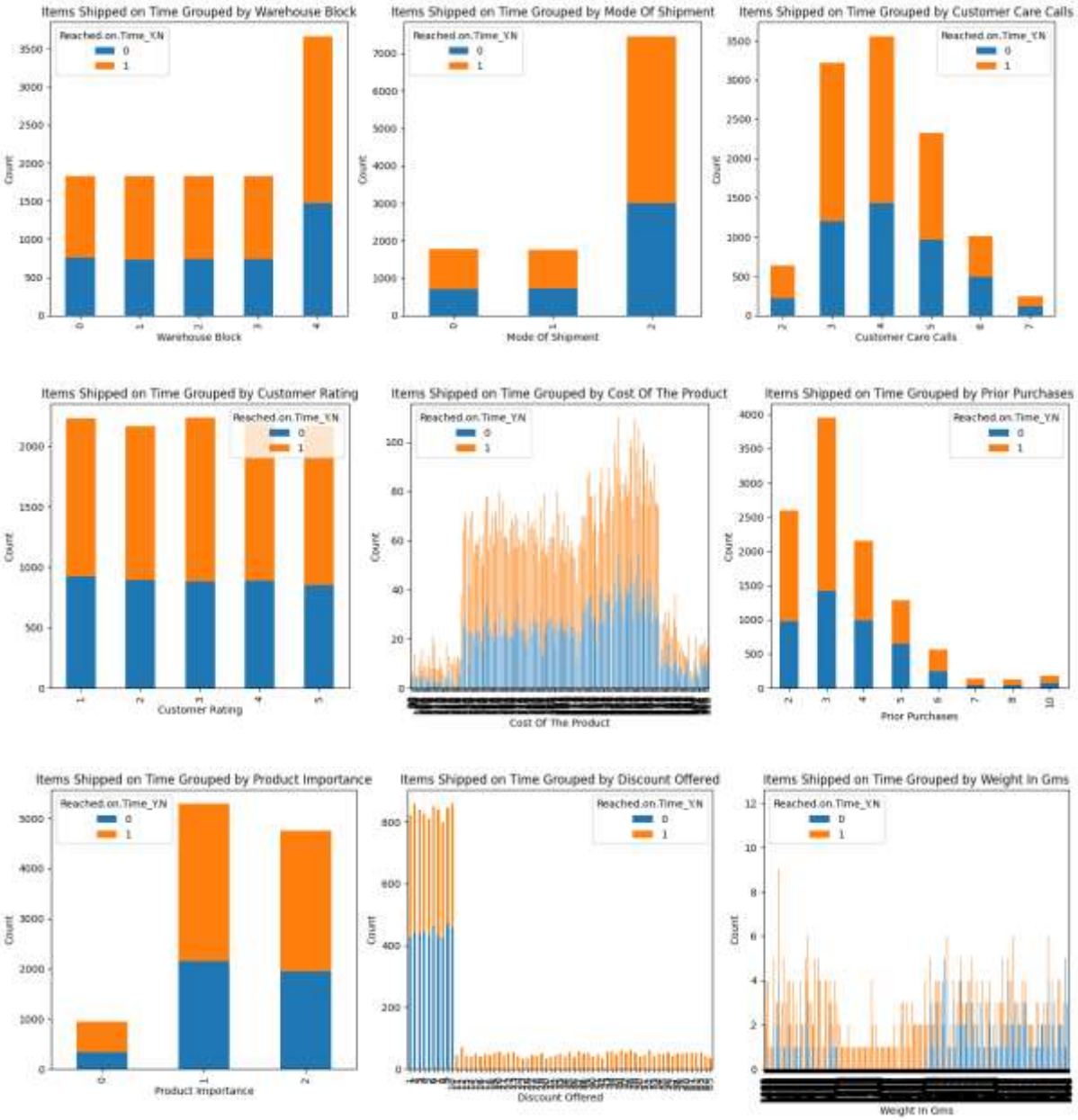
C. Encoding : For each categorical data in the dataset, we encode it to convert its data type into numerical data. We encode the Warehouse block, Mode of shipment, Product importance, and Gender features.

D. Pre-Modeling : At this stage, we perform a train-test split with a ratio of 70% data for training and 30% for testing. We also standardize using a standard scaler. We do not do undersampling or oversampling because the dataset we use is already quite balanced.

3. DATASET

Here are the details of the dataset we used, below we present the Correlation Heatmap and plot the number of on-time and off-time packet arrivals based on each feature:





4. METHOD

Our research aims to develop an effective predictive model for package delivery timeliness. We used four machine learning algorithms to identify the best model with the highest accuracy such as :

A. K-Nearest Neighbors : K-Nearest Neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

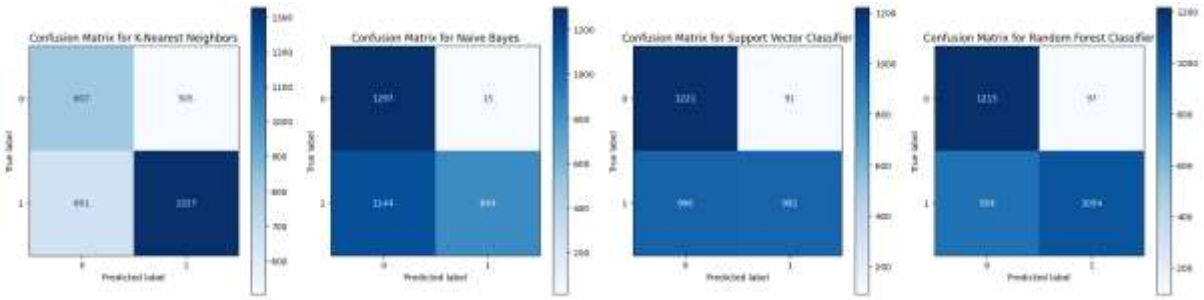
B. Naïve Bayes : Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

C. Support Vector Classifier : SVC is a powerful algorithm for binary classification tasks. It works by finding the hyperplane that best separates the classes in the feature space. SVC is effective in high-dimensional spaces and is particularly useful when the number of features exceeds the number of samples.

D. Random Forest : Random forest is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

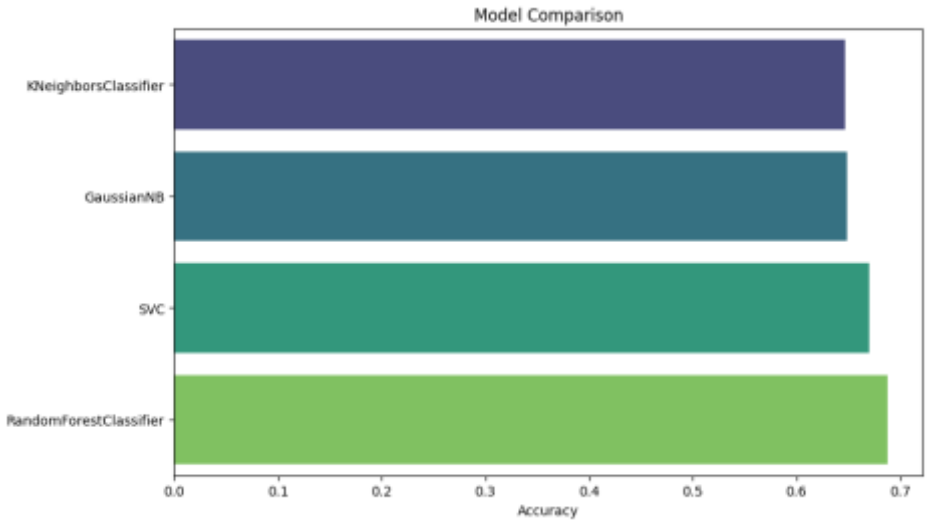
5. RESULT & ANALYSIS

Here we present the confusion matrix of the prediction model that we applied (K-Nearest Neighbors, Naïve Bayes, Support Vector Classifier, and Random Forest) :



From the results seen in the confusion matrix, two things can be concluded, namely between the model that has not been able to predict well or the dataset we use is indeed a poor dataset.

6. EVALUATION



From the table above, we can see that the accuracy of each model is :

- K-Nearest Neighbor : 64.66%
- Naive Bayes : 64.87%
- SVC: 67.06%
- Random Forest: 68.75%

The overall model we use is still not able to produce good accuracy. We have tried to do dimension reduction such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Hyperparameter Tuning but the score does not change much, even worse.

7. CONCLUSION

In summary, the research aims to address package delivery delays by applying machine learning techniques. Despite evaluating various models (including K-Nearest Neighbor, Naive Bayes, Support Vector Machine, and Random Forest), the overall accuracy remains suboptimal. Further improvements are necessary to enhance the precision of predicting package delivery timeliness.

8. IMPLICATION

The study on predicting package delivery timeliness through machine learning models has significant implications for the logistics industry. Accurate predictions directly impact a service provider’s reputation. When packages arrive promptly, customers perceive the company as reliable and trustworthy. Conversely, consistent delays can erode public trust and harm the brand image. Therefore, maintaining a positive reputation is crucial for sustained success.

Operational efficiency and cost savings are closely tied to accurate predictions. By optimizing resource allocation such as rerouting packages based on predictions companies can reduce inefficiencies. Fewer re-deliveries, minimized customer complaints, and better resource utilization all contribute to cost savings. Accurate predictions provide a competitive advantage. Customers prioritize reliability when choosing service providers. Companies known for timely deliveries stand out in a crowded market. Moreover, proactive risk management addressing potential delays before they occur mitigates disruptions and positions companies for growth.

Customer satisfaction drives business success. Satisfied customers are more likely to recommend the service and remain loyal. As e-commerce continues to expand, reliable package delivery becomes a critical factor in customer choice.

In summary, while the research shows promise, further improvements are necessary to enhance accuracy in predicting package delivery timeliness. Service providers must leverage these findings to optimize operations, build trust, and thrive in a competitive landscape.

9. NEW_INTRODUCTION

9.1 Background of the Study

After conducting research and applying machine learning models to the E-Commerce Shipping Data dataset, our group realized that there was a mistake in the poor dataset we chose. We have tried many things such as dimensionality reduction method, undersampling or oversampling, even hyperparameter tunning but the model performance is even worse. After further research, we decided to apply our model to a different dataset.

In this research, our group will try to prove that the error is in the dataset and also at the same time give more insight to airlines or airports to improve their performance to increase passenger satisfaction through our research results.

9.2 Problem Definition

The main mistake was the poor dataset we had. To prove that the error was in our dataset, our group conducted a re-study by applying the model we created before to the new dataset without changing the model we created and only adjusting some of the preprocessing steps such as encoding.

Airline Passenger Satisfaction is very important. Customer satisfaction can depend on many factors such as good service at the airport, as well as service while on the plane until leaving the airport. We conducted a study to predict the level of customer satisfaction with machine learning models. We use a dataset that is equipped with features that directly affect customer satisfaction.

From this research, airline companies and airports can use the prediction results of our model to evaluate and improve their performance.

9.3 Solution to the Problem

Our group takes a new dataset that is similar to the old dataset and uses it for our machine learning model. The new dataset we used was the Airline Passenger Satisfaction dataset that we took from Kaggle just like the previous data. We did not change most of the preprocessing

and the model we used was exactly the same. We will compare the performance of each model against both datasets.

9.4 Other Studies

Research entitled "Predicting Airline Passenger Satisfaction with Classification Algorithms" conducted by B. Herawan Hayadi et al. states that the classification model is very appropriate to help airline service providers and airports to increase passenger satisfaction. Factors such as Wi-Fi service and ease of online ticket booking are factors that greatly affect passenger satisfaction.

In another study, Juliet Namukasa in her research entitled "The influence of airline service quality on passenger satisfaction and loyalty: The case of Uganda airline industry" states that airline service quality has a significant influence on passenger satisfaction and loyalty, where not only in-flight services, but pre-flight and post-flight services also have a significant influence on passenger satisfaction.

10. NEW_DATA

10.1 Dataset Used

The dataset we use is the "Airline Passenger Satisfaction" dataset from kaggle. The dataset consists of 25976 data rows and 24 features (see table below for details).

Feature	Description
Gender	Gender of the passengers (Female, Male)
Customer Type	The customer type (Loyal customer, disloyal customer)
Age	The actual age of the passengers
Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance	The flight distance of this journey
Inflight wifi service	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
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 (Satisfaction, neutral or dissatisfaction)

By considering these parameters, the model will be able to provide more accurate predictions regarding the airline passengers satisfaction. Each feature is selected based on factors that directly affect the airline passengers satisfaction.

10.2 Pre-Processing (Same as number 2.2)

Here are the details of the initial to final stages of pre-processing that our group applied to our dataset:

A. Drop Unused Column : Since the ID and Gender column is not needed, we decided to drop the ID and Gender column.

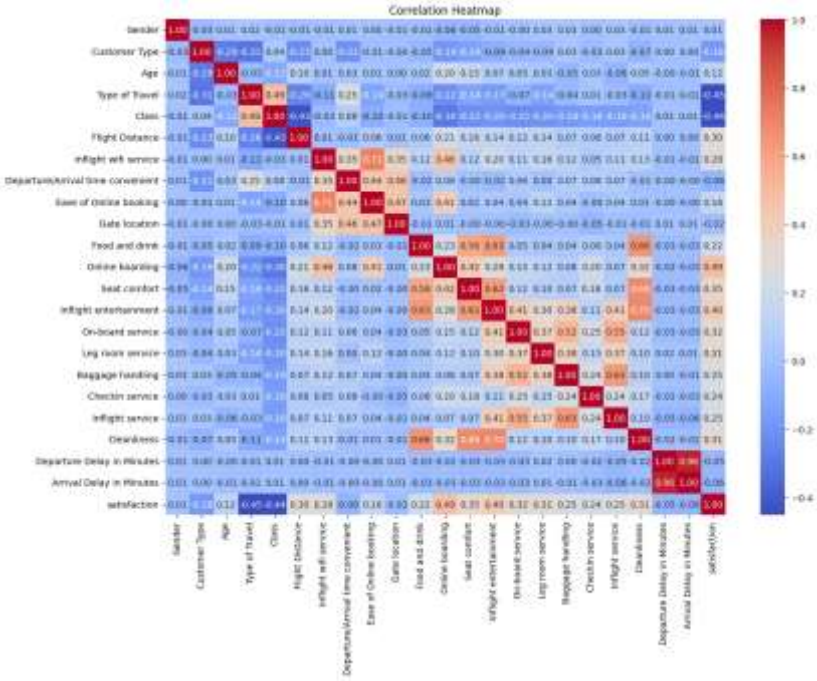
B. Handle Missing Value : Since the dataset we use is clean and has no missing values, we do not fill in the missing values.

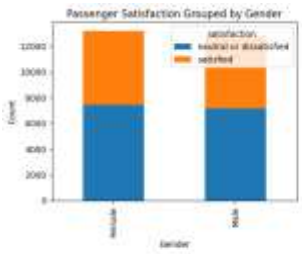
C. Encoding : For each categorical data in the dataset, we encode it to convert its data type into numerical data. We encode the Warehouse block, Mode of shipment, Product importance, and Gender features.

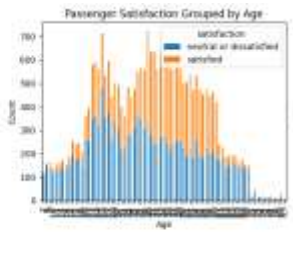
D. Pre-Modeling : At this stage, we perform a train-test split with a ratio of 70% data for training and 30% for testing. We also standardize using a standard scaler. We do not do undersampling or oversampling because the dataset we use is already quite balanced.

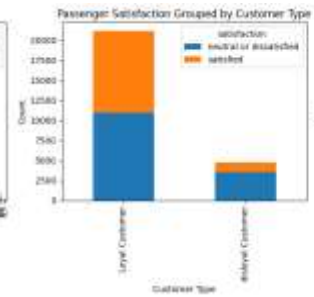
11. NEW_DATASET

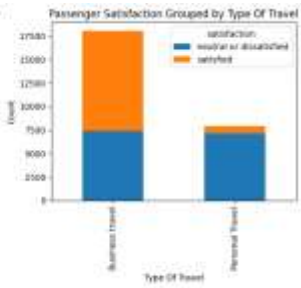
Here are the details of the dataset we used, below we present the Correlation Heatmap and plot the number of on-time and off-time packet arrivals based on each feature:

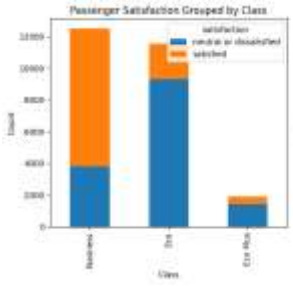

Correlation Heatmap
Gender
Customer Type
Age
Type of Travel
Class
Flight Distance
Inflight wifi service
Departure/Arrival time convenient
Ease of Online booking
Gate location
Food and drink
Online boarding
Seat comfort
Inflight entertainment
On-board service
Leg room service
Baggage handling
Checkin service
Inflight service
Cleanliness
Departure Delay in Minutes
Arrival Delay in Minutes
satisfaction

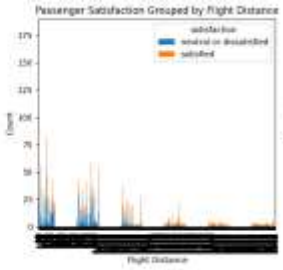

Passenger Satisfaction Grouped by Gender
satisfaction
neutral or dissatisfied
satisfied
Count
Female
Male
Gender

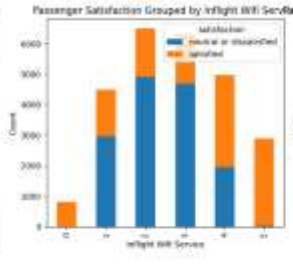

Passenger Satisfaction Grouped by Age
satisfaction
neutral or dissatisfied
satisfied
Count
Age

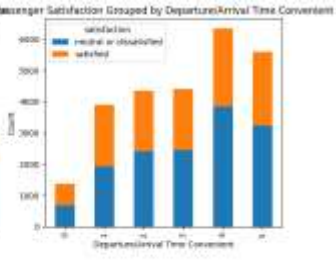

Passenger Satisfaction Grouped by Customer Type
satisfaction
neutral or dissatisfied
satisfied
Count
Loyal Customer
disloyal Customer
Customer Type


Passenger Satisfaction Grouped by Type Of Travel
satisfaction
neutral or dissatisfied
satisfied
Count
Business travel
Personal Travel
Type Of Travel

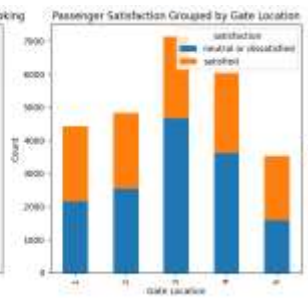

Passenger Satisfaction Grouped by Class
satisfaction
neutral or dissatisfied
satisfied
Count
Business
Eco
Eco Plus
Class

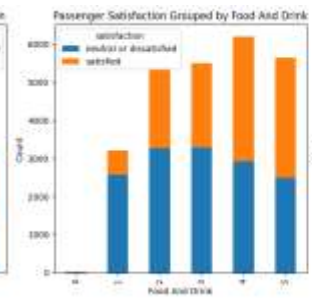

Passenger Satisfaction Grouped by Flight Distance
satisfaction
neutral or dissatisfied
satisfied
Count
Flight Distance

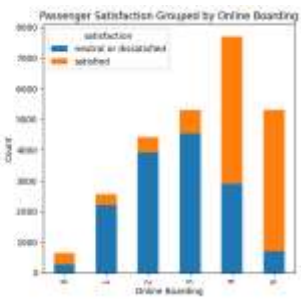

Passenger Satisfaction Grouped by Inflight Wifi Service
satisfaction
neutral or dissatisfied
satisfied
Count
Inflight Wifi Service

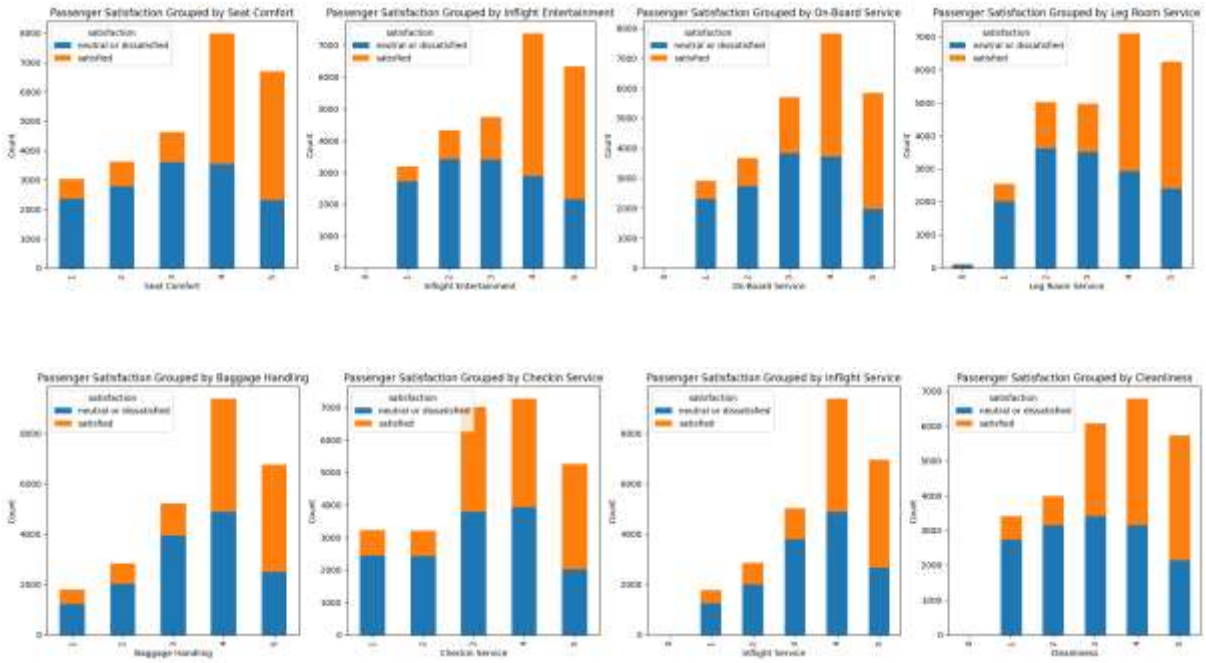

Passenger Satisfaction Grouped by Departure/Arrival Time Convenient
satisfaction
neutral or dissatisfied
satisfied
Count
Departure/Arrival Time Convenient


Passenger Satisfaction Grouped by Ease Of Online Booking
satisfaction
neutral or dissatisfied
satisfied
Count
Ease Of Online Booking


Passenger Satisfaction Grouped by Gate Location
satisfaction
neutral or dissatisfied
satisfied
Count
Gate Location


Passenger Satisfaction Grouped by Food And Drink
satisfaction
neutral or dissatisfied
satisfied
Count
Food And Drink


Passenger Satisfaction Grouped by Online Boarding
satisfaction
neutral or dissatisfied
satisfied
Count
Online Boarding



12. METHOD (Same as number 4)

Our research aims to develop an effective predictive model for package delivery timeliness. We used four machine learning algorithms to identify the best model with the highest accuracy such as :

A. K-Nearest Neighbors : K-Nearest Neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

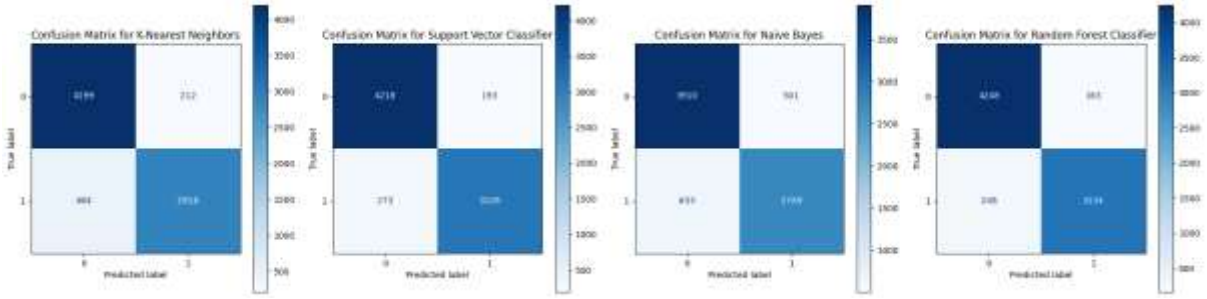
B. Naïve Bayes : Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

C. Support Vector Classifier : SVC is a powerful algorithm for binary classification tasks. It works by finding the hyperplane that best separates the classes in the feature space. SVC is effective in high-dimensional spaces and is particularly useful when the number of features exceeds the number of samples.

D. Random Forest : Random forest is a commonly used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

13. NEW_RESULT & ANALYSIS

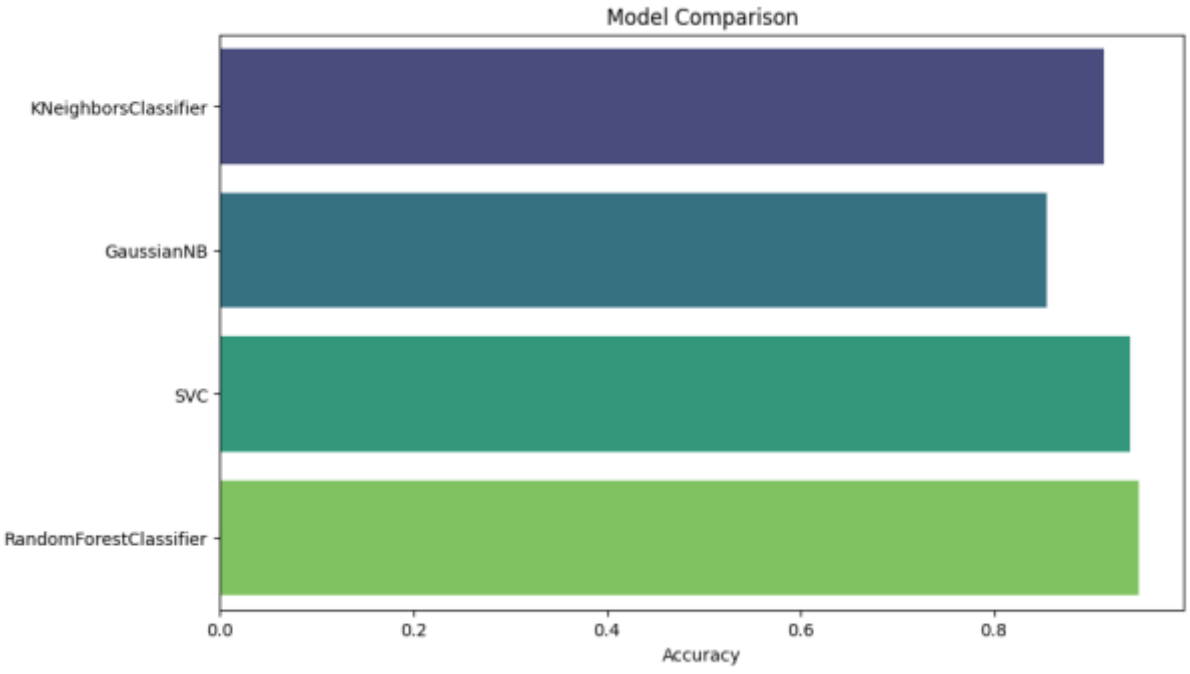
Here we present the confusion matrix of the prediction model that we applied (K-Nearest Neighbors, Naïve Bayes, Support Vector Classifier, and Random Forest) :



From the confusion matrix results, we can conclude that our model can predict airline customer satisfaction. The performance of each model is very high.

Random Forest has the best prediction performance among the other three models. Random Forest correctly predicts 7854 predictions out of a total of 8243 predictions, followed by Support Vector Classifier which correctly predicts 7327 predictions out of a total of 7793 predictions, then followed again by K-Nearest Neighbors which correctly predicts 7117 predictions out of a total of 7797 predictions and finally Naive Bayes which correctly predicts 6659 predictions out of a total of 7793 predictions.

14. NEW_EVALUATION



From the table above, we can see that the accuracy of each model is :

- K-Nearest Neighbors : 91.33%
- Naive Bayes : 85.45%
- SVC: 94.02%
- Random Forest: 94.72%

The overall model we use already able to produce good accuracy. We have tried to do dimension reduction such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Hyperparameter Tuning but the score much worse.

Random Forest has the best performance among the other three models with an accuracy score of 94.72% and Support Vector Classifier has the second best performance with an accuracy score of 94.02%. K-Nearest Neighbors also performed quite well with an accuracy score of 91.33% and Naïve Bayes performed quite well with an accuracy score of 85.45% although not better than the other three models.

15. NEW_CONCLUSION

Here we present the model performance comparison data on different datasets, namely dataset 1 “E-Commerce Shipping Data” and dataset 2 “Airline Passenger Satisfaction”:

	Model	Accuracy
Dataset 1	K-Nearest Neighbors	64.66%
	Naïve Bayes	64.87%
	Support Vector Classifier	67.06%
	Random Forest	68.75%
Dataset 2	K-Nearest Neighbors	91.33%
	Naïve Bayes	94.02%
	Support Vector Classifier	85.45%
	Random Forest	94.72%

After changing the dataset, it can be concluded that in our previous study there was a problem with the dataset selection. Because in the second dataset, with the same preprocessing and modeling steps we can get a much better accuracy value with Random Forest as the best predicting model. So, our second study is more in line with the purpose of the title, which is to predict the level of passenger satisfaction.

16. IMPLICATION

The study of predicting airline passenger satisfaction through machine learning models has significant implications for the airline industry. Accurate predictions directly impact the reputation of the service provider. When many passengers are satisfied, passengers will perceive the service provider company as reliable and trustworthy. Conversely, passenger dissatisfaction can erode public trust and damage the company's image. Therefore, maintaining a positive reputation is essential for continued success.

The addition of internet network services for economy passengers and optimizing the way tickets are booked online are closely related to accurate predictions. By improving these and other services, the company can gain greater customer satisfaction and trust. Customer satisfaction drives business success. Satisfied customers tend to recommend services and remain loyal. Along with the development of the airline business, airline services that provide the best service become an important factor in customer choice.

In summary, in this study we successfully built a predictive machine learning model that is good at predicting passenger satisfaction. Airline service providers can utilize these findings to improve service performance and evaluate minor errors that may decrease satisfaction.