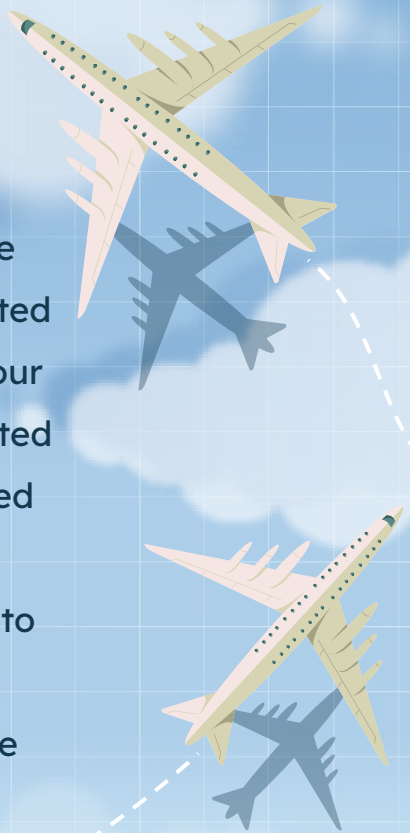
PREDICTING AIRLINE PASSENGER SATISFACTION

Krisha Patel, Rithika Pathuri, Isha Gajera

ABSTRACT

For this project, we used rpart() and naive_bayes() to analyze the "Airplane Satisfaction" dataset through predictions. The dataset contains information related to passenger/flight attributes, which provided a comprehensive foundation for our investigation on determining passenger satisfaction. Utilizing rpart, we constructed a decision tree model that could effectively classify the satisfaction of fliers based on a multitude of factors, and we determined which predictor is the greatest indicator of passenger satisfaction. Additionally, Naive Bayes was implemented to assess the probabilistic relationships between various attributes and predict satisfaction levels. Afterwards, we compared the performance of rpart and Naive Bayes functions in accurately classifying passenger satisfaction in order to determine the strengths and limitations of each algorithm.



WHAT DATA DID WE USE?

you can download the dataset we used here:
https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

STEP: 1

UNDERSTANDING THE DATA

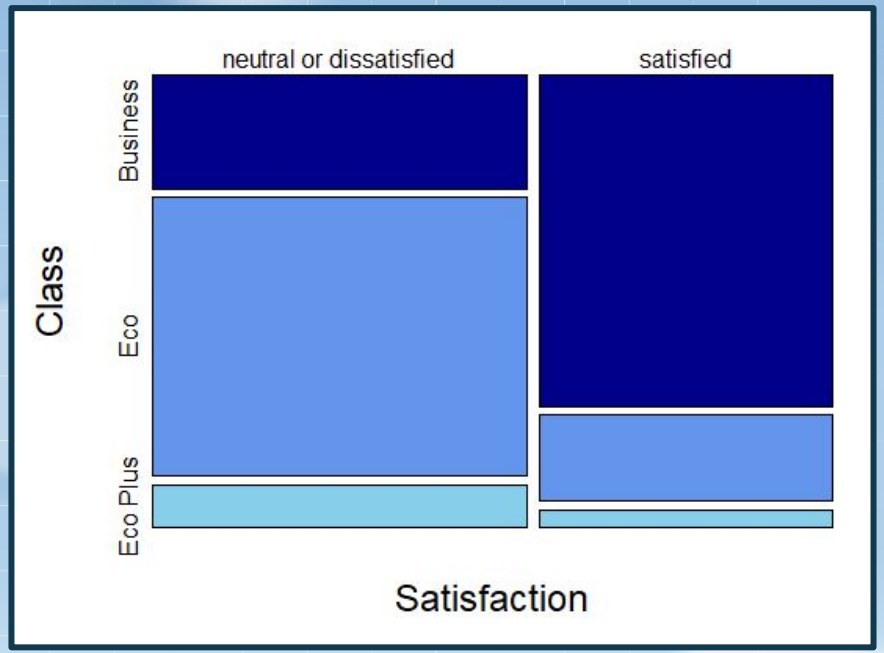
using box & mosaic plots in R

```
> str(AirlineTrain)
'data.frame':   103904 obs. of  27 variables:
 $ X                                : int  0 1 2 3 4 5 6 7 8 9 ...
 $ id                               : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
 $ Gender                           : chr  "Male" "Male" "Female" "Female" ...
 $ Customer.Type                    : chr  "Loyal Customer" "disloyal Customer" "Loyal Customer" "Loyal Customer" ...
 $ Age                              : int  13 25 26 25 61 26 47 52 41 20 ...
 $ Type.of.Travel                   : chr  "Personal Travel" "Business travel" "Business travel" "Business travel" ...
 $ Class                            : chr  "Eco Plus" "Business" "Business" "Business" ...
 $ Flight.Distance                  : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
 $ Inflight.wifi.service            : int  3 3 2 2 3 3 2 4 1 3 ...
 $ Departure.Arrival.time.convenient: int  4 2 2 5 3 4 4 3 2 3 ...
 $ Ease.of.Online.booking           : int  3 3 2 5 3 2 2 4 2 3 ...
 $ Gate.location                    : int  1 3 2 5 3 1 3 4 2 4 ...
 $ Food.and.drink                   : int  5 1 5 2 4 1 2 5 4 2 ...
 $ Online.boarding                  : int  3 3 5 2 5 2 2 5 3 3 ...
 $ Seat.comfort                     : int  5 1 5 2 5 1 2 5 3 3 ...
 $ Inflight.entertainment           : int  5 1 5 2 3 1 2 5 1 2 ...
 $ On.board.service                 : int  4 1 4 2 3 3 3 5 1 2 ...
 $ Leg.room.service                 : int  3 5 3 5 4 4 3 5 2 3 ...
 $ Baggage.handling                 : int  4 3 4 3 4 4 4 5 1 4 ...
 $ Checkin.service                  : int  4 1 4 1 3 4 3 4 4 4 ...
 $ Inflight.service                 : int  5 4 4 4 3 4 5 5 1 3 ...
 $ Cleanliness                      : int  5 1 5 2 3 1 2 4 2 2 ...
 $ Departure.Delay.in.Minutes       : int  25 1 0 11 0 0 9 4 0 0 ...
 $ Arrival.Delay.in.Minutes         : num  18 6 0 9 0 0 23 0 0 0 ...
 $ satisfaction                     : chr  "neutral or dissatisfied" "neutral or dissatisfied" "satisfied" "neutral or dissatisfied"
...
 $ neutral or dissatisfied          : num  0.742 0.781 0.742 0.781 0.238 ...
 $ satisfied                        : num  0.258 0.219 0.258 0.219 0.762 ...
>
```

used str() to get an overview of the dataset before plotting and drawing conclusions

Based on this data, we plotted predictors with satisfaction to get a better understanding of the important factors. The following are the variables we deemed important.

CLASS FLOWN VS SATISFACTION



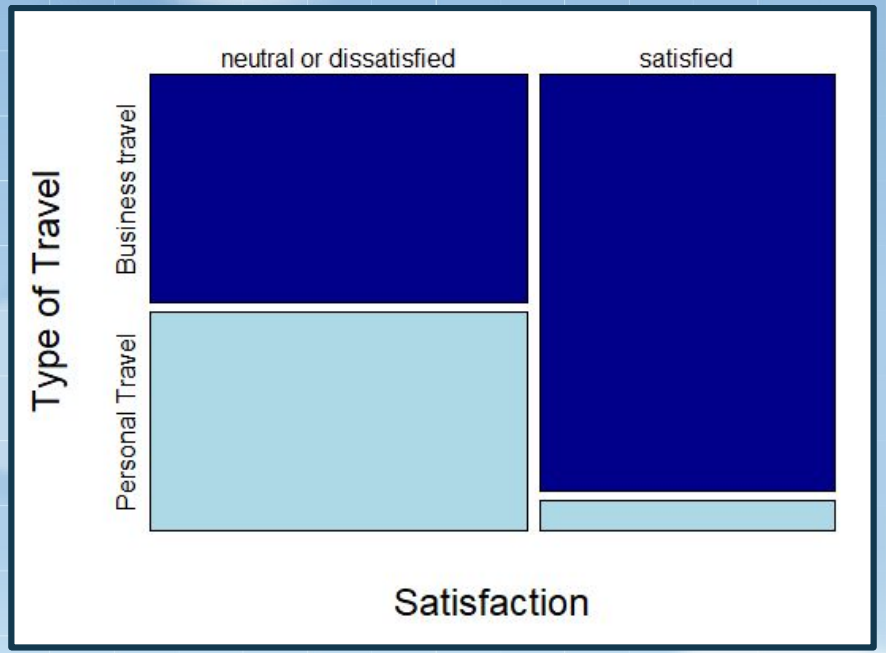
Since **Class** and **Satisfaction** are both **categorical** variables, we plotted a mosaic to show the proportions of passengers in each class and their satisfaction ratings. This allows us to visually highlight any apparent relationships or skew.

This graph shows:

- typically, the people in Business were more satisfied.
- Eco Plus passengers were rarely satisfied and rarely neutral/dissatisfied

This could mean that there is a disproportionate amount of data in this dataset for people from Business vs. Eco Plus.

TYPE OF TRAVEL VS SATISFACTION



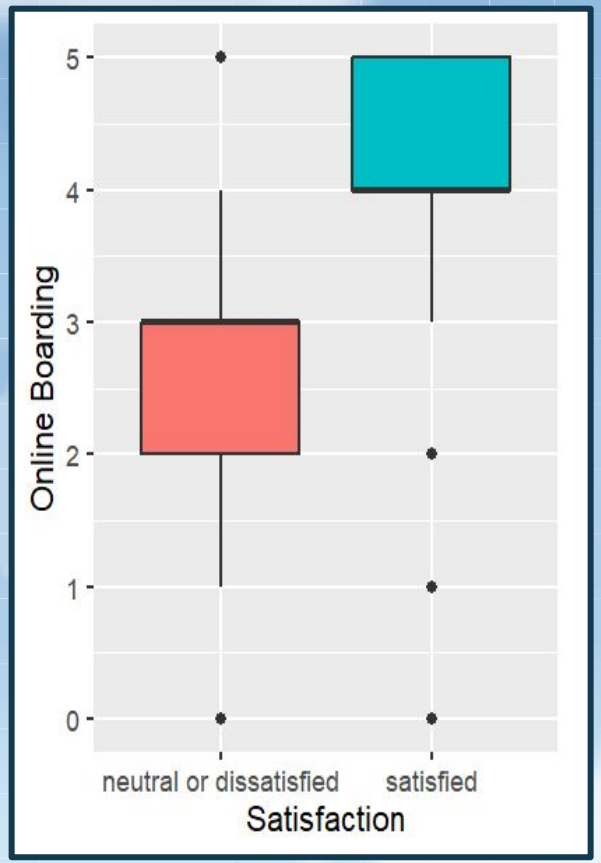
Since **Travel Type** and **Satisfaction** are both **categorical** variables, we plotted a mosaic to show the proportions of passengers in each category and their satisfaction ratings.

This graph shows:

- typically, the people traveling for Business were more satisfied than not.
- People traveling for personal travel were rarely satisfied and proportionately similar to those n/d travelling for business.

This could mean that there is a disproportionate amount of data in this dataset for people flying for Business vs Personal reasons. This skew will significantly impact our prediction.

ONLINE BOARDING VS SATISFACTION



Since **Satisfaction** is **categorical** and the rest of our variables from now on are **numerical**, we plotted them with boxplots. This clearly shows the distribution of the data, especially highlighting means and outliers.

This graph shows:

- the average Online Boarding ratings for passengers who were:
  - neutral/dissatisfied = 3
  - satisfied = 4

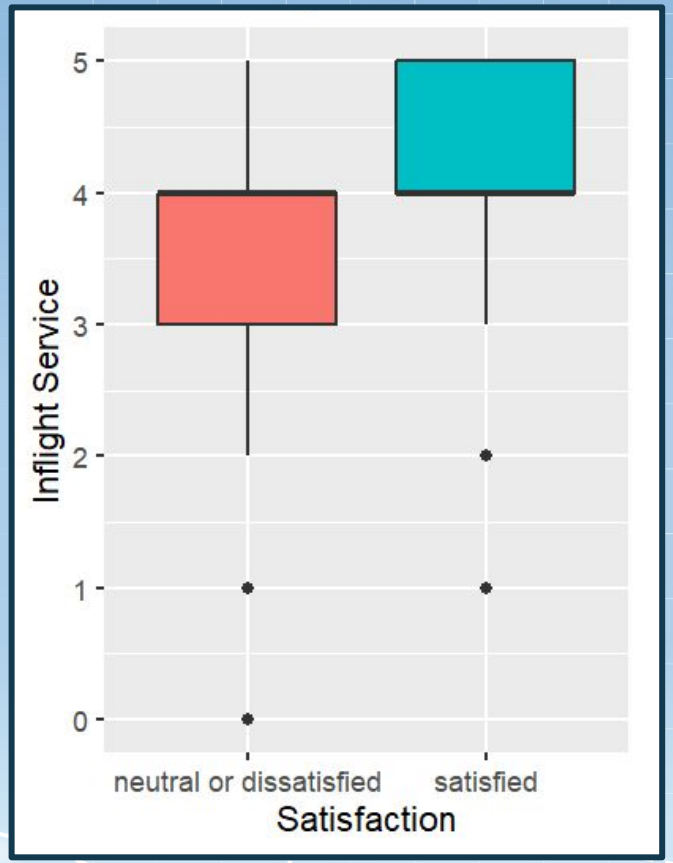
Scores for Online Boarding were typically lower for neutral/dissatisfied passengers compared to satisfied passengers. The ranges differ a lot as well (excluding outliers). The difference in distribution implies significance.

INFLIGHT SERVICE VS SATISFACTION

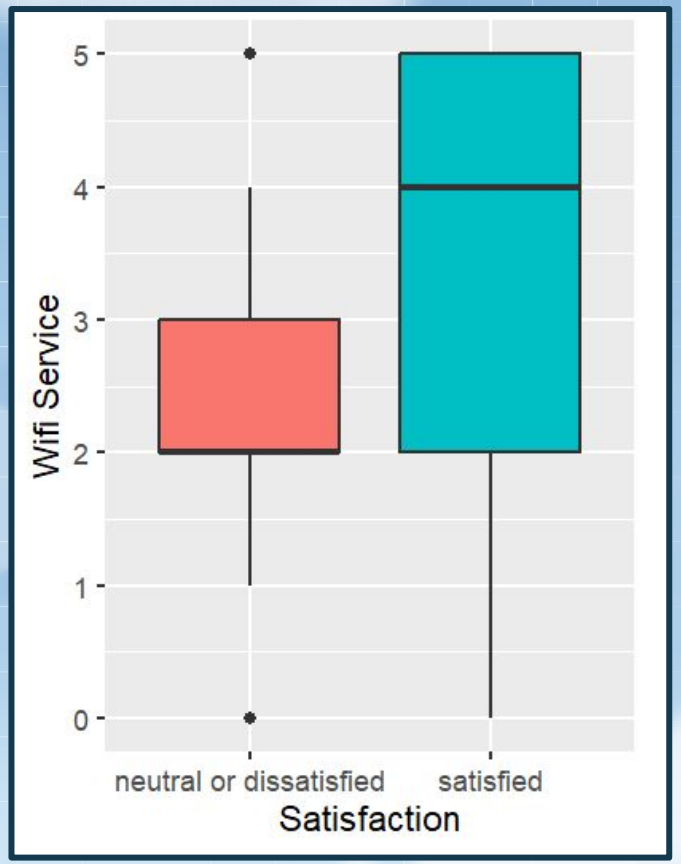
This graph shows:

- the average Inflight Service ratings for passengers who were:
  - neutral/dissatisfied = 4
  - satisfied = 4

The mean scores for Inflight Service for neutral/dissatisfied passengers are the same as the satisfied passengers. There is less variation in range when comparing 2-5 and 3-5. The outliers are similar as well. We don't believe this distribution implies significance as much as other variables.



WIFI SERVICE VS SATISFACTION



This graph shows:

- the average Inflight Wifi Service ratings for passengers who were:
  - neutral/dissatisfied = 2
  - satisfied = 4

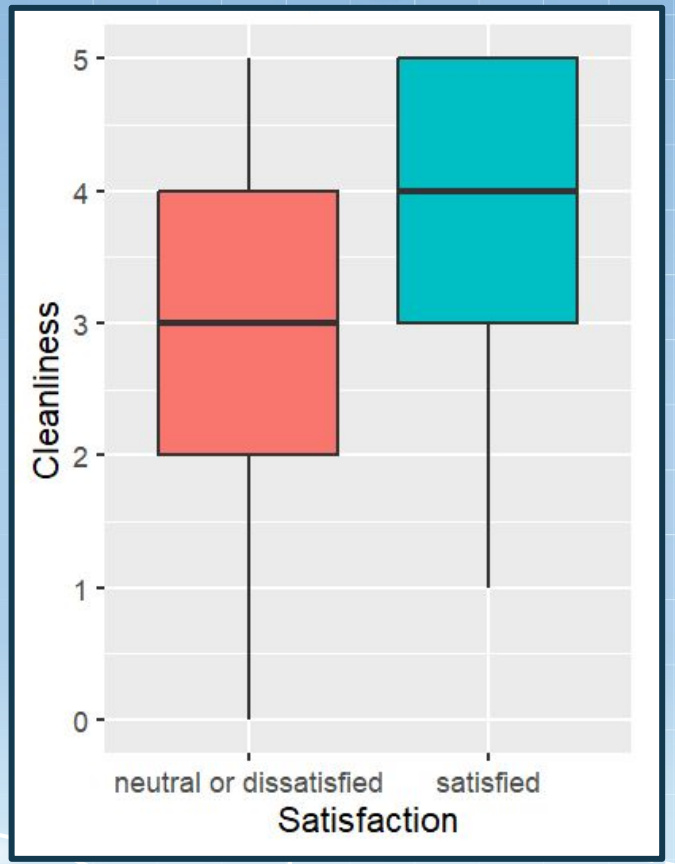
Scores for Wifi Service were typically lower for neutral/dissatisfied passengers compared to satisfied passengers. The ranges are very different as well. The difference in means may imply significance.

CLEANLINESS VS SATISFACTION

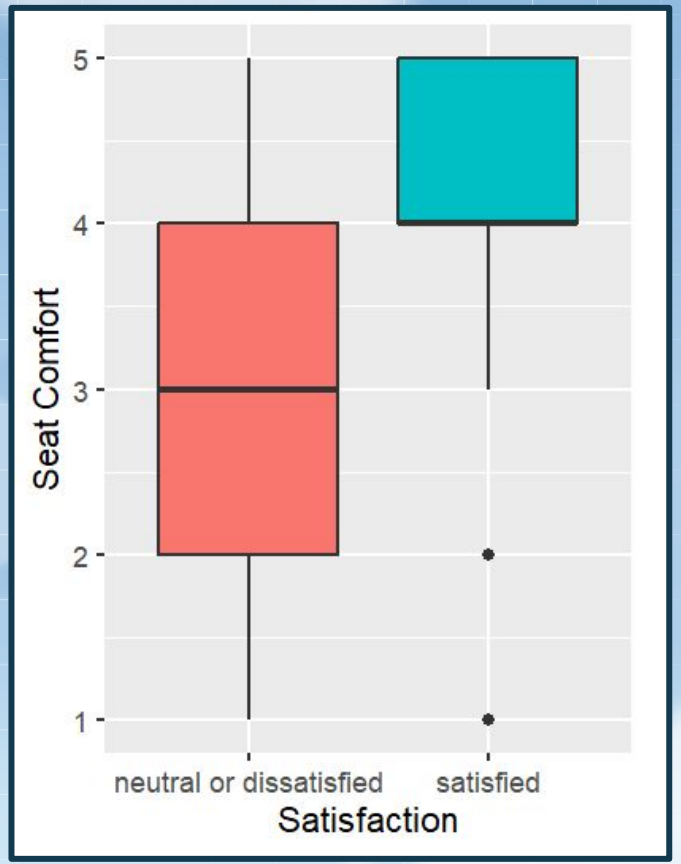
This graph shows:

- the average Cleanliness ratings for passengers who were:
  - neutral/dissatisfied = 3
  - satisfied = 4

Scores for Cleanliness were typically lower for neutral/dissatisfied passengers compared to satisfied passengers. The range is larger than that of the satisfied customers. We don't believe this distribution implies significance as much as other variables.



SEAT COMFORT VS SATISFACTION



This graph shows:

- the average Seat Comfort ratings for passengers who were:
  - neutral/dissatisfied = 3
  - satisfied = 4

Although scores for Seat Comfort were typically higher for satisfied passengers, they are evenly distributed among neutral/dissatisfied passengers. This suggests that seat comfort may not play a significant role in their satisfaction ratings. It appears as though the satisfied customers may have been biased and thus scored seat comfort higher on average.

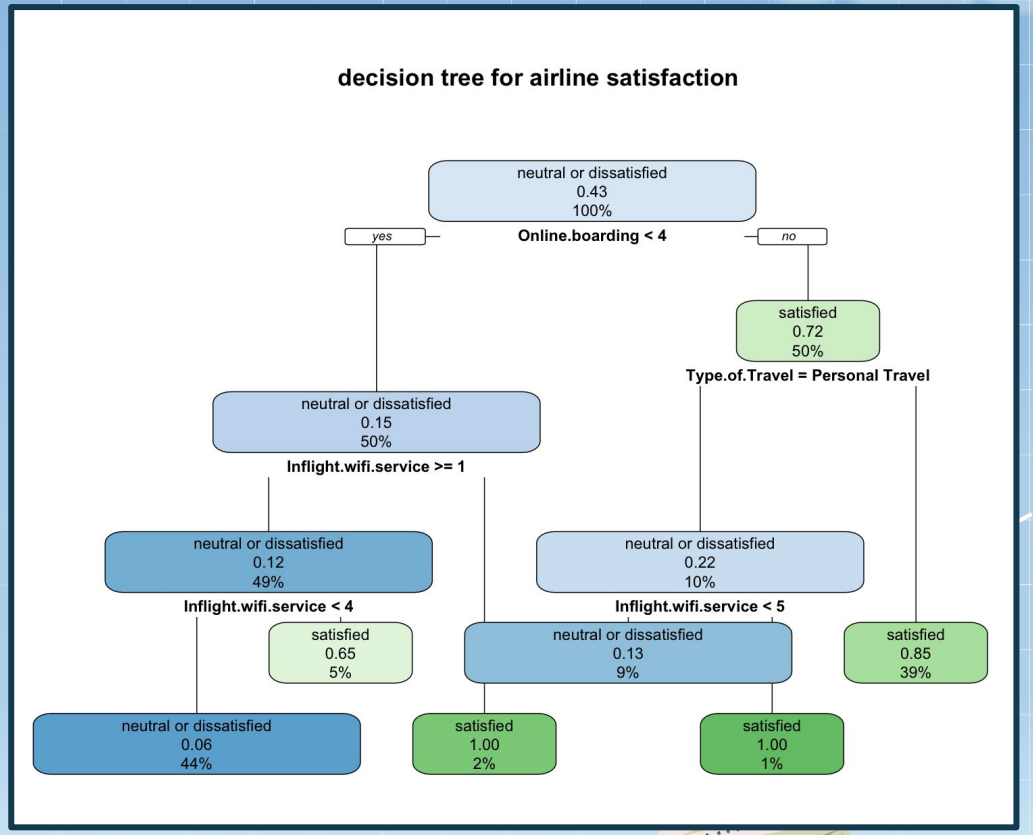
STEP: 2

CREATE PREDICTION MODELS

Using rpart() and naive_bayes()

DECISION TREE USING RPART()

- ★ Each node description starts with the most persistent factor in regards to satisfaction (Online boarding in this case)
- ★ As you follow the node to the leaf node, smaller partitions that are dependent on other variables are created to make a pathway for satisfaction. This included factors like inflight wifi and the type of travel.
- ★ These predictors reflect our initial assumptions found through plotting



RPART() RESULTS!

```
> #confusion matrix for training data
> (matrix <- table(satisfactionPrediction, AirlineTrain$satisfaction))

satisfactionPrediction neutral or dissatisfied satisfied
                     1                   51094      4211
                     2                    7785     40814
> 1-sum(diag(matrix))/sum(matrix)
[1] 0.1154527                ← misclassification rate
> #calculation for accuracy
> correct_predictions <- sum(diag(matrix))
> total_predictions <- sum(matrix)
> accuracy <- correct_predictions / total_predictions
> (accuracy) #is 88.45473%
[1] 0.8845473                ← accuracy rate! not the best, but not bad
>
```

NAIVE BAYES CLASSIFICATION

```
> bayesMatrix <- table(satisfactionBayesPrediction, AirlineTrain$satisfaction)
> (bayesMatrix)

satisfactionBayesPrediction neutral or dissatisfied satisfied
    neutral or dissatisfied                   49717      9246
    satisfied                                  9162     35779
> 1 - sum(diag(bayesMatrix)) / sum(bayesMatrix)
[1] 0.1771635
>
```

RESULTS:

Misclassification percentage of 17.71635% = **accuracy of 82.28365%**

STEP: 3

RESULTS

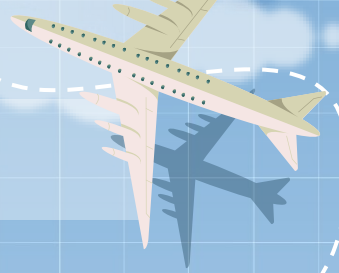
Our findings from the prediction models

WHAT DID WE FIND OUT?

To summarize our findings:

- The prediction model using **Rpart** resulted in an **accuracy** of **88.45** percent with a **misclassification** of **11.55** percent.
- The model using **Naive Bayes Classification** resulted in an **accuracy** of **82.28** percent with a **misclassification** of **17.72** percent.

RPART VS NAIVE BAYES CLASSIFICATION



Rpart → typically builds decision trees by splitting up values of our variables used for predictions

- Splits into homogenous sections
- Recursive partitioning → able to capture complex relationships → higher accuracy
- Decision trees are more difficult to interpret

Naive Bayes Classification → tends to assume the independence of predictors

- Simplifies computations → straightforward interpretations
- Able to capture true independence of variables
- Accuracy may be lower when predictors are related

Both Rpart and Naive Bayes Classification are **useful** depending on the **nature** and **specific characteristics** of the data set being analyzed. For **our use**, we found **Rpart** to be more **useful**, as we were able to **visualize** and better **capture** the complex relationship between airplane satisfaction and numerous predictors from our data set through the **decision tree** unique to rpart().

WHAT DOES THIS MEAN?

Since we discovered that the **Naive Bayes Classification** accuracy was **lower** than the accuracy found using rpart, we can assume that the predictors are **not as independent** of each other and that there **exists a correlation** between the predictors we used and airplane satisfaction. Additionally, the **relatively higher accuracy** from the **rpart** model combined with rpart's ability to display complex relationships indicates yet again how the predictors we used and airplane satisfaction must have a **complex relationship.**

AI USAGE

PLOTTING/GRAPHING

To increase efficiency, we asked ChatGPT to graph box plots for all of our variables. This reduced the amount of repetitive tasks we had to do and streamlined our process. We edited the code to make sure the plots highlighted what we wanted, and made them more visually appealing.

DETERMINING ACCURACY

To double check our calculations, we asked ChatGPT to create a template to verify our accuracy and misclassification results.

LITERATURE USED:

https://www.r-bloggers.com/2021/04/naive-bayes-classification-in-r/

THANK YOU