# Project Step 0:

Team member: Data Miners Unearthed Members:

* Kenny Aranda (n01513222)



* Zeke Geiger (n01360784)



* Ross Crawford (n01431391)



Project title: Airline Passenger Satisfaction Prediction. Data: Airline passenger satisfaction dataset.

Airline passenger satisfaction ([https://www.kaggle.com/datasets/teejmahal20/airline-passenger-](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction) [satisfaction](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction))

Leaders:

Step 1 leader: Zeke Geiger; Step 2 leader: Kenny Aranda; Step 3 leader: Ross Crawford; Step 4 leader: Zeke Geiger

Communication: We plan to have a group chat through Discord or Microsoft teams. We are going to communicate twice a week to review the assignments for the project steps.

Tools: Weka, Microsoft word, Teams

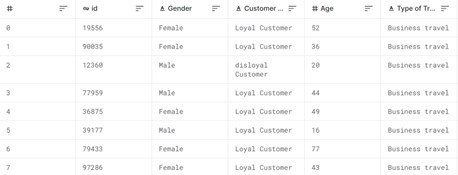
# Project Step 1:

The problem involves analyzing an airline passenger satisfaction dataset to uncover factors that are highly correlated with passenger

satisfaction or dissatisfaction. The primary task is to predict passenger satisfaction based on the

features in Figure 1:

Airlines can use insights from this analysis to identify areas where passenger satisfaction is



*Figure 1*

low and make targeted improvements. This can lead to happier customers and increased loyalty.

Additionally, understanding factors leading to dissatisfaction can help airlines allocate resources more

efficiently to address the most critical issues, potentially reducing costs. Airlines that consistently provide high passenger satisfaction are likely to gain a competitive edge in the industry. Knowing what matters to passengers can help airlines differentiate their services.

Airlines are among the primary audience for this analysis. Airline companies can use this analysis to enhance their services, prioritize improvements, and tailor their offerings to meet passenger

expectations. However, there are other audiences that can benefit. Passengers will benefit from improved airline services and overall satisfaction during their travels. Regulatory authorities can use the insight to monitor and enforce passenger satisfaction standards within the airline industry. Marketing teams can use the findings to develop targeted campaigns that highlight the strengths of their airline's services.

Professionals in the field of data analysis and data science can gain insights into practical applications of data mining and predictive modeling using this real-world dataset.

While one might think humans could solve this problem, there is no feasible approach. Almost three million people fly in and out of U.S. airport every single day1. Additionally, the dataset we chose has over 100,000 datapoints, each with 22 factors that could affect customer satisfaction in varying and seemingly abstract ways. Assuming 1% of people flying in and out of the U.S. respond to a customer satisfaction survey, there are 30,000 new data entries every day that need to be analyzed. A computer- based solution is best; it can analyze the dataset using an algorithm and find the factors that most affect

customer satisfaction. A solution to the problem would be able to analyze the data of a customer and their experiences/opinions of flight details and determine whether they were satisfied with the flight. It would take their data, such as their age and satisfaction with legroom, flight time, flight services, etc. and output if they were satisfied or neutral/dissatisfied with the flight. Figure 2 outlines the Black Box.





Input: Information including passenger information and flight information



Black Box



Output: The satisfaction of the passenger with their flight.

*Figure 2*

The dataset is sourced from Kaggle: [https://www.kaggle.com/datasets/teejmahal20/airline-passenger-](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction) [satisfaction.](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction) The data comes from TJ Klein, who compiled and reformatted information from an

anonymous source.

|  |  |  |
| --- | --- | --- |
| [test.csv](https://d.docs.live.net/a04f41e89d90197f/SCHOOL/Fall 2023/CAP 4770 Data Mining/test.csv) | 3.04mb | 20% of full dataset |
| [train.csv](https://d.docs.live.net/a04f41e89d90197f/SCHOOL/Fall 2023/CAP 4770 Data Mining/train.csv) | 12.19mb | 80% of full dataset |

1 https:/[/w](http://www.faa.gov/air_traffic/by_the_numbers)w[w.faa.gov/air\_traffic/by\_the\_numbers](http://www.faa.gov/air_traffic/by_the_numbers)

# Project Step 2:

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Description** | **Nominal/Numeric** |
| **Gender** | Gender of the passengers (Female, Male) | Nominal |
| **Customer Type** | The customer type (Loyal  customer, disloyal customer) | Nominal |
| **Age** | The actual age of the passengers | Numeric |
| **Type of Travel** | Purpose of the flight of the passengers (Personal Travel, Business Travel) | Nominal |
| **Class** | Travel class in the plane of the passengers (Business, Eco, Eco Plus) | Nominal |
| **Flight distance** | The flight distance of this journey | Numeric |
| **Inflight Wi-Fi service** | Satisfaction level of the inflight Wi-Fi service | Numeric |
| **Departure/Arrival time convenient** | Satisfaction level of Departure/Arrival time convenient (1-5) | Numeric |
| **Ease of Online booking** | Satisfaction level of online booking (1-5) | Numeric |
| **Gate location** | Satisfaction level of Gate location (1-5) | Numeric |
| **Food and drink** | Satisfaction level of Food and drink (1-5) | Numeric |
| **Online boarding** | Satisfaction level of online boarding (1-5) | Numeric |
| **Seat comfort** | Satisfaction level of Seat comfort (1-5) | Numeric |
| **Inflight entertainment** | Satisfaction level of inflight entertainment (1-5) | Numeric |
| **On-board service** | Satisfaction level of On-board service (1-5) | Numeric |
| **Leg room service** | Satisfaction level of Leg room service (1-5) | Numeric |
| **Baggage handling** | Satisfaction level of baggage handling (1-5) | Numeric |
| **Check-in service** | Satisfaction level of Check-in service (1-5) | Numeric |
| **Inflight service** | Satisfaction level of inflight service (1-5) | Numeric |
| **Cleanliness** | Satisfaction level of Cleanliness (1-5) | Numeric |
| **Departure Delay in Minutes** | Minutes delayed when departure | Numeric |
| **Arrival Delay in Minutes** | Minutes delayed when Arrival | Numeric |
| **Satisfaction** | Airline satisfaction level (Satisfaction, neutral or dissatisfaction) | Nominal |

We want to predict what attributes affect customer satisfaction to determine the most effective attributes to change. Therefore, satisfaction is the attribute picked as the class. Since a customer is either satisfied or neutral/dissatisfied, we are dealing with a binary classification problem. Below is a table listing all of the attributes, their descriptions, and their types:

Our data is split into two .csv files; a training file and a test file. They are split as such:

|  |  |  |
| --- | --- | --- |
| **File name** | **# of Instances** | **# of Attributes** |
| **Train.csv** | 103904 | 24 |
| **Test.csv** | 25976 | 24 |

The numeric attributes of the training dataset and test dataset along with the minimum and maximum values, the mean, and the standard deviation are shown in a table below. The values from both datasets are the same in most cases and similar in the rest. For example, the departure delay and arrival delay

attributes have a higher max value in the training dataset than the test dataset. Since the training dataset has 4 times the number of instances compared to the test dataset, it has more room for extreme values that can skew the statistics.

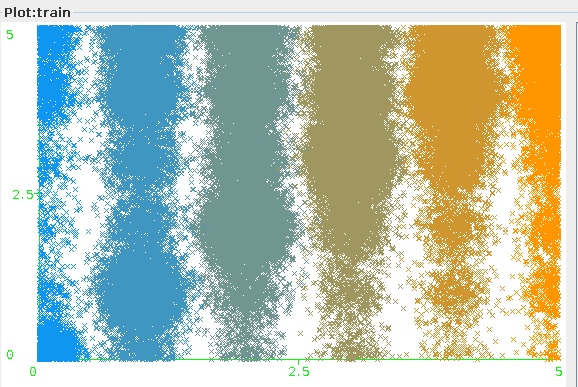
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Attribute*** | **Min**  **(train | test)** | **Max**  **(train | test)** | **Mean**  **(train | test)** | **StDev**  **(train | test)** |
| *Flight Distance* | 31 | 31 | 4983 | 4983 | 1189.4 | 1193.8 | 997.1 | 998.7 |
| *Inflight Wi-F****i*** *Service* | 0 | 0 | 5 | 5 | 2.73 | 2.73 | 1.33 | 1.34 |
| *Departure/Arrival Time*  *Convenience* | 0 | 0 | 5 | 5 | 3.06 | 3.05 | 1.53 | 1.53 |
| *Ease of Online Booking* | 0 | 0 | 5 | 5 | 2.76 | 2.76 | 1.40 | 1.41 |
| *Gate Location* | 0 | 0 | 5 | 5 | 2.98 | 2.98 | 1.28 | 1.28 |
| *Food and Drink* | 0 | 0 | 5 | 5 | 3.20 | 3.21 | 1.33 | 1.33 |
| *Online Boarding* | 0 | 0 | 5 | 5 | 3.25 | 3.26 | 1.35 | 1.36 |
| *Seat comfort* | 0 | 0 | 5 | 5 | 3.43 | 3.45 | 1.31 | 1.32 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Inflight entertainment* | 0 | 0 | 5 | 5 | 3.36 | 3.36 | 1.33 | 1.34 |
| *On-board service* | 0 | 0 | 5 | 5 | 3.38 | 3.39 | 1.29 | 1.28 |
| *Leg room service* | 0 | 0 | 5 | 5 | 3.35 | 3.35 | 1.31 | 1.31 |
| *Baggage handling* | 0 | 0 | 5 | 5 | 3.63 | 3.63 | 1.18 | 1.17 |
| *Check-in service* | 0 | 0 | 5 | 5 | 3.30 | 3.31 | 1.27 | 1.27 |
| *Inflight service* | 0 | 0 | 5 | 5 | 3.64 | 3.65 | 1.18 | 1.18 |
| *Cleanliness* | 0 | 0 | 5 | 5 | 3.29 | 3.29 | 1.31 | 1.32 |
| *Departure Delay (in minutes)* | 0 | 0 | 1592 | 1128 | 14.82 | 14.31 | 38.23 | 37.42 |
| *Arrival Delay (in minutes)* | 0 | 0 | 1584 | 1115 | 15.18 | 14.74 | 38.70 | 37.52 |

The class distribution is also similar for both the training dataset and test dataset. While the amount is different for each dataset, they still retain the same proportion of satisfied-to-neutral/dissatisfied.

### STEP 4: Explore the training data:

The three attributes most correlated with the class are online boarding, inflight wifi service, and class, respectively.



ZeroR Naive Bayes J48 KSTAR LMT

56% 86% 72% 82% 73%

Naive Bayes performed suprisingly well with the base line correctly classifying just over half of the dataset

### Step 3 report

**1a.** The Random Forest classifier is an ensemble learning method that combines multiple

Decision Trees to improve prediction accuracy. It works by building several Decision Trees on subsets of the data and then combining their predictions. This reduces overfitting and enhances accuracy, making it a popular choice for various machine learning tasks.

**1b.** Random Forest differs from other ensemble classifiers in that it builds multiple Decision Trees with feature selection randomness, reducing overfitting and enhancing accuracy. Other

ensemble methods like bagging, boosting, voting, and stacking use different strategies and may combine diverse base models for different purposes.

### 2a.

RandomForest 1: No changed settings (Baseline) RandomForest 2: bagSizePercent changed to 95

RandomForest 3: bagSizePercent changed to 90

RandomForest 4: numExecutionSlots changed to 2

RandomForest 5: maxDepth changed to 10

RandomForest 6: maxDepth changed to 20

RandomForest 7: batchSize changed to 150

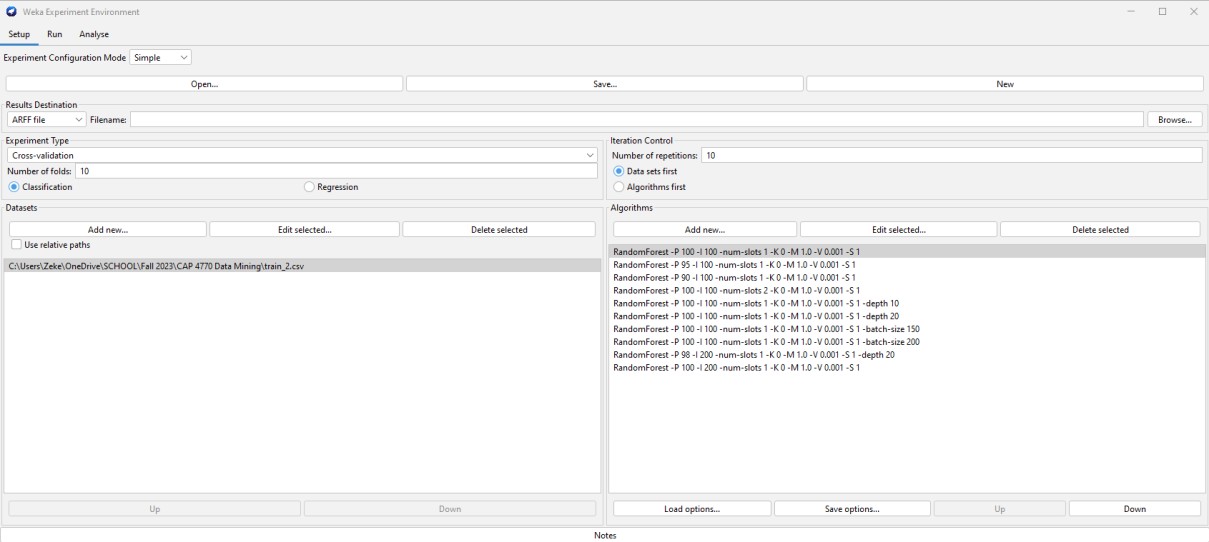
RandomForest 8: batchSize changed to 200

RandomForest 9: bagSizePercent changed to 98, numIterations changed to 200, and depth to 20 RandomForest 10: numIterations changed to 200.

**2b.** To validate the RandomForest algorithms, we are using 10-fold cross validation. This

separates the data into “folds” or sections, where nine folds are used to train the algorithms on the supplied dataset and the final fold is used as testing. This is then repeated nine more times; each fold will be a testing fold at least once. The results are then averaged at the end. We are using this validation technique because it allows us to test the algorithms on ten test sets for each algorithm. If we did a training/testing dataset split, we would only test the algorithm parameters once; doing cross-validation allows us to test it ten times.

**2c.** To test our performance, we used the corrected paired t-tester. The Paired T-tester lets us

compare the mean of the baseline RandomForest algorithm cross validation with the means of the other RandomForest algorithm cross validations. Each run of the cross validation on an algorithm is paired with the run on the baseline. After this, we use an equation comparing the sample size, the sample mean of the differences, and the sample standard deviation of the differences. If the number given is higher or lower than the significance defined (in our case, we used 0.05,) then the algorithm is deemed to be statistically better or worse than the baseline algorithm. We used the corrected paired t-tester because we are using the same dataset; the algorithms can be paired one-to-one because they have the same sample size and the same objective. Additionally, we are trying to prove that changing the parameters on the algorithm can have a higher performance; using the paired t-tester allows this.

**3A.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **Percenta ge P** | **Iteratio ns I** | **Num**  **-**  **slots** | **K** | **M** | **V** | **S** | **Dept h** | **Batch**  **-size** | **Percent\_corre ct** | **Significan ce** |
| **1** | **100** | **100** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **-** | **-** | **96.42** | **-** |
| **2** | **95** | **100** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **-** | **-** | **96.42** | **-** |
| **3** | **90** | **100** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **-** | **-** | **96.40** | **\*** |
| **4** | **100** | **100** | **2** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **-** | **-** | **96.42** | **-** |
| **5** | **100** | **100** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **10** | **-** | **94.86** | **\*** |
| **6** | **100** | **100** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **20** | **-** | **96.38** | **\*** |
| **7** | **100** | **100** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **-** | **150** | **96.42** | **-** |
| **8** | **100** | **100** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **-** | **200** | **96.42** | **-** |
| **9** | **98** | **200** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **20** | **-** | **96.41** | **-** |
| **1**  **0** | **100** | **200** | **1** | **0** | **1.**  **0** | **0.00**  **1** | **1** | **-** | **-** | **96.45** | **v** |

**Significance Key:**

* v: Significantly better
* \*: Significantly worse
* -: Not significantly different

The 10th combination, with **P=100**, **I=200**, and **num-slots=1**, achieved the highest accuracy of 96.45% and is marked as significantly better (**v**) compared to the default parameters. The results show that increasing the number of iterations (**I**) from 100 to 200 (as in the 9th and 10th combinations) improves the performance, which is intuitive since more iterations allow the Random Forest to learn better.

## 4A.

ZeroR is the simplest classification method which relies on the target and ignores all predictors. OneR is slightly more complex than ZeroR. it generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule".

RandomForest

OneR

ZeroR

0

20

40

60

80

100

120

Series 1Series 3

Baseline Performances: ZeroR, being the simplest classifier that predicts the majority class, has the lowest performance at 56.67%. OneR, which considers one rule for prediction, performs significantly better than ZeroR with a performance of 79.04%.

RandomForest's Superiority: The RandomForest classifier, with the specified parameters, outperforms both baselines by a significant margin, achieving a performance of 96.45%. This showcases the power of ensemble methods like RandomForest, which combine multiple

decision trees to produce a more accurate and stable prediction.

Comparison with Baselines: It's crucial to compare sophisticated models with simple

baselines like ZeroR and OneR. The vast difference in performance between RandomForest and the baselines emphasizes the effectiveness of the RandomForest model for this particular dataset.

Potential Overfitting: While the RandomForest's performance is impressive, it's close to 100%, which might indicate potential overfitting to the training data. It would be essential to validate the model's performance on a separate test set to ensure its generalization

capabilities.

**Project step 4**

**4.** Analyze the detailed results of your "top" model.

Our top model from the selected few was random forest. CVParameterSelection inside AttributeSelectedClassifier using random forest selected the following as the top 5 attributes and can be explained intuitively: 1. Type of Travel: Business travel will likely be higher class and is seen as a must so customers are more satisfied than personal travel, 2.Class: First class customers will have a better experience and will therefore be more satisfied, 3. Inflight wifi service: Planes with wifi will have more entertainment value and therefore more satisfaction, 4. Online boarding: An accessible and easy booking process leads to satisfied customers, 5. Inflight entertainment: Like Inflight wifi but less personable however still provides an entertaining and satisfying flight. Comparing this to a model to search for cancer (must have high standards), we see our model performs just slightly less at 92.24% compared to 94% of the cancer classification ([1]). This means our model performs very well and these attributes have a high link to airline satisfaction

[1]: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6858312/>

**5. Conclusions and future work**

1. **What are some of the key high-level takeaways from your project findings?**

The analysis identified key attributes strongly correlated with passenger satisfaction. These included the type of travel, class, inflight WiFi service, online boarding, and inflight entertainment. Understanding the impact of these factors can help airlines prioritize improvements. Enhancing these aspects might lead to increased passenger satisfaction and loyalty. The type of travel and class might impact satisfaction differently for various passenger segments. For instance, business travelers might prioritize different amenities compared to leisure travelers within the same class. Inflight WiFi service and online boarding emerged as influential factors. Investing in better connectivity and streamlined boarding processes could significantly enhance passenger experiences. Highlighting these strengths in marketing and service offerings might attract more passengers. Insights gleaned from this analysis can guide future decision-making processes for airlines. It provides a roadmap for focusing resources and efforts on areas that matter most to passengers.

**Top of Form**

1. **What are some of the limitations of your work? In other words, if you had more time and/or resources, what improvements would/could you try?**One potential limitation of our work was the exclusive use of some classifiers instead of exploring other machine learning algorithms, such as TensorFlow classifiers. TensorFlow allows for building complex neural network models that might capture intricate patterns within the data, potentially leading to more nuanced insights. Time and resource constraints limited our ability to explore a wider array of models that could potentially yield different insights or more accurate predictions.

**Appendix**

**Member 1 name:** Kenny Aranda

**Member 2 name:** Zeke Geiger

**Member 3 name:** Ross Crawford

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Member 1** | **Member 2** | **Member 3** |
| **Step 1** | **33.33%** | **33.33%** | **33.33%** |
| **Step 2** | **33.33%** | **33.33%** | **33.33%** |
| **Step 3** | **33.33%** | **33.33%** | **33.33%** |
| **Step 4** | **33.33%** | **33.33%** | **33.33%** |