Airline

SC1015 Mini-Project

Predicting Flight Delays

Members:

Tan Leong Jun Joseph (U2321339H)
Tan Zhe Kai (U2322419A)

BOARDING PASS

● **GROUP**
FCS2

● **VENUE**
HWLAB 3

● **DATASET**
KAGGLE

Airline

TABLE OF CONTENTS

01 | Problem

02 | EDA

03 | Models

04 | Conclusion

Airline

01 Problem Formulation

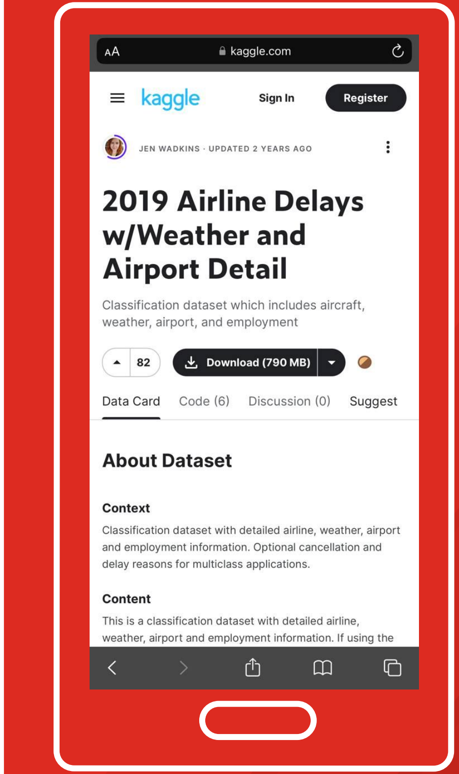
Flight Delay:	>15 MINS
On-time:	~80%
Average length:	~50 MINS

Airline

Dataset

- 8 data files
- Overlapping variables
- All data files connected
- Extensive data preparation and cleansing techniques needed

Source



JEN WADKINS · UPDATED 2 YEARS AGO

2019 Airline Delays w/Weather and Airport Detail

Classification dataset which includes aircraft, weather, airport, and employment

About Dataset

Context

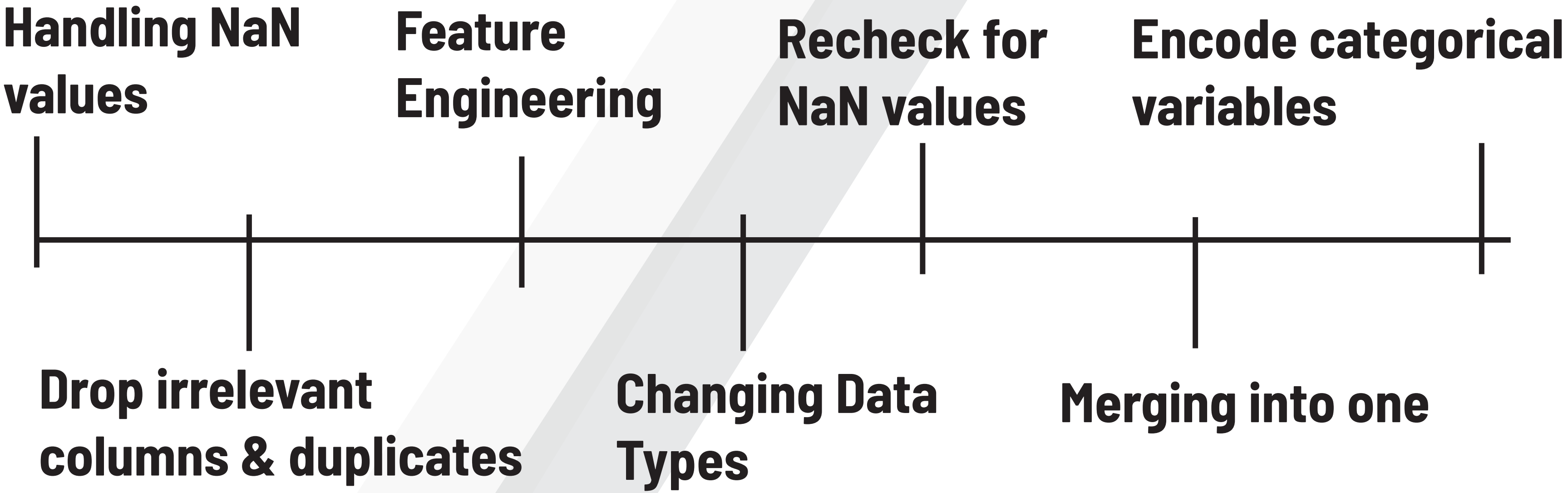
Classification dataset with detailed airline, weather, airport and employment information. Optional cancellation and delay reasons for multiclass applications.

Content

This is a classification dataset with detailed airline, weather, airport and employment information. If using the

Airline

Data Preparation



Handling NaN values

Drop irrelevant columns & duplicates

Feature Engineering

Changing Data Types

Recheck for NaN values

Merging into one

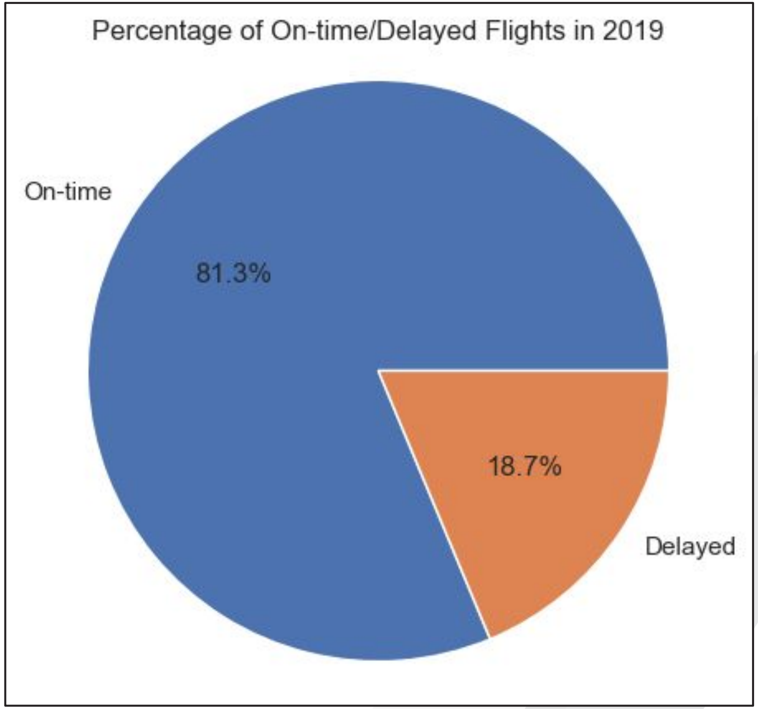
Encode categorical variables

Airline

EDA

02 Exploratory Data Analysis

EXPLORATORY DATA ANALYSIS

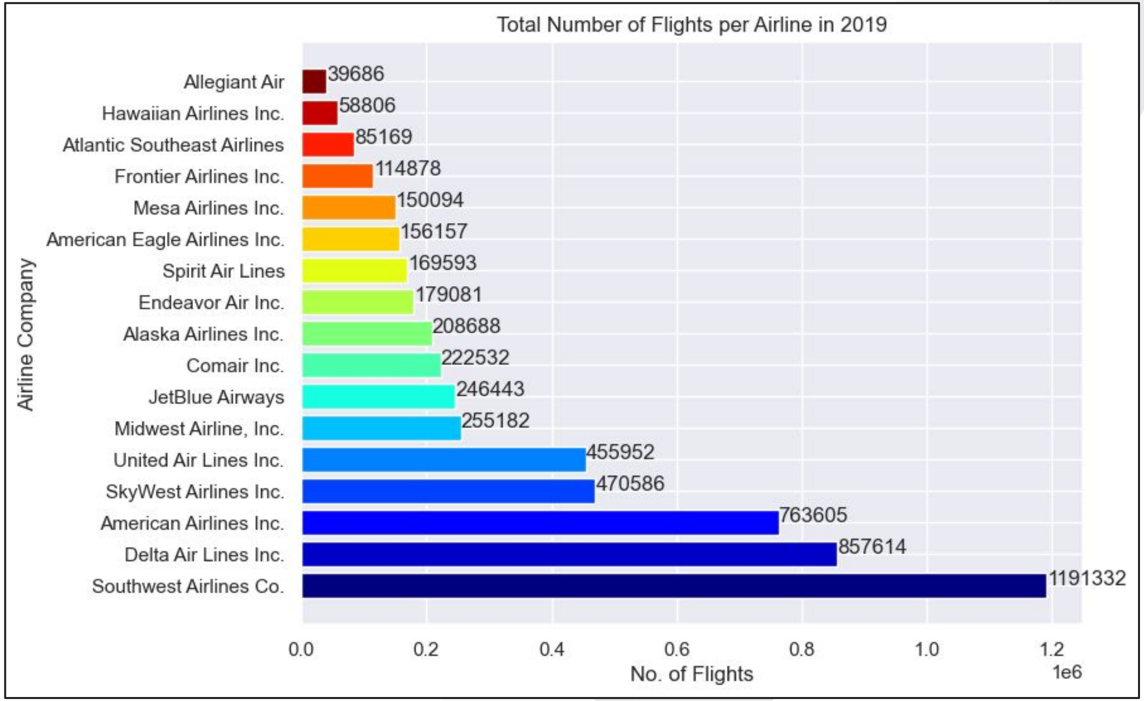


On-time vs Delayed

| ON-TIME: 4.5 million flights

| DELAYED: 1 million flights

EXPLORATORY DATA ANALYSIS



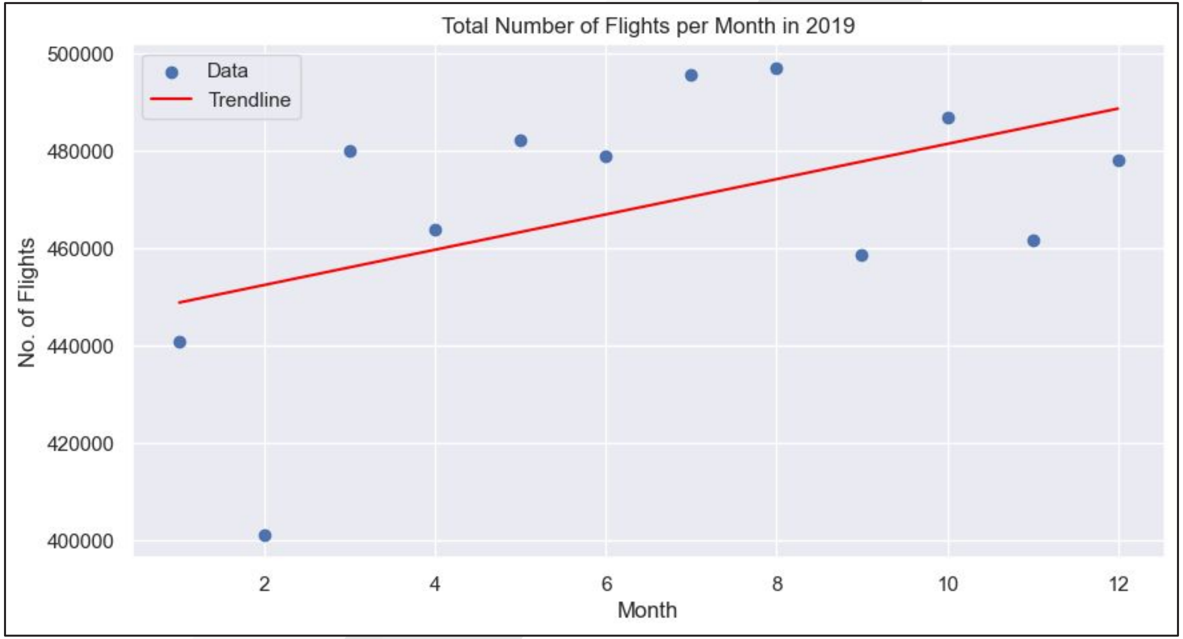
Airlines w/ Most Flights

1. Southwest Airline Co.
   - 1.1M / 21.2%

2. Delta Air Lines Inc.
   - 850k / 15.2%

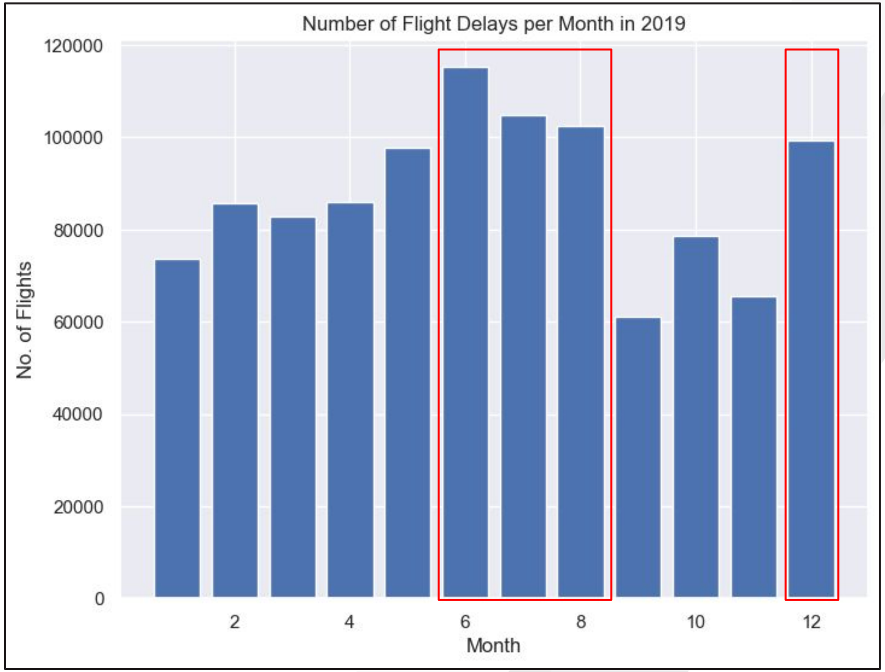
3. American Airlines Inc.
   - 760k / 13.6%

EXPLORATORY DATA ANALYSIS



Increasing trend in the number of flights per month

Which month has the most flight delays?



1. June, July & August (Summer)

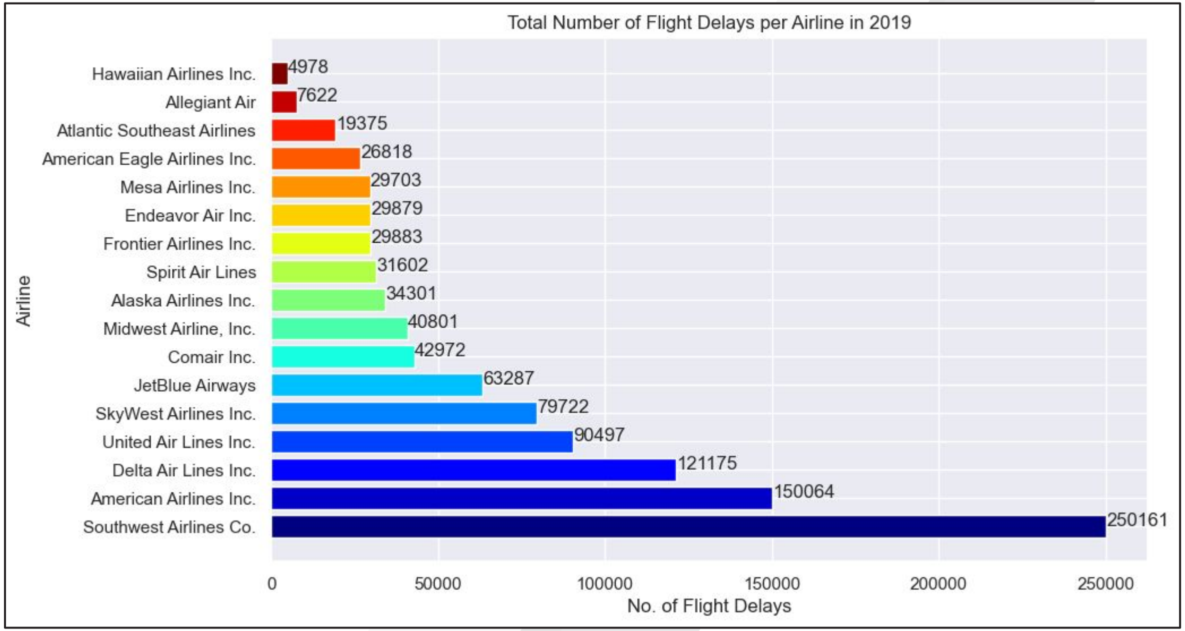
2. December (Winter)

Possible reasons:

- Adverse weather conditions
- Vacation period
- High number of concurrent flights

EXPLORATORY DATA ANALYSIS

Which airline has the most flight delays?



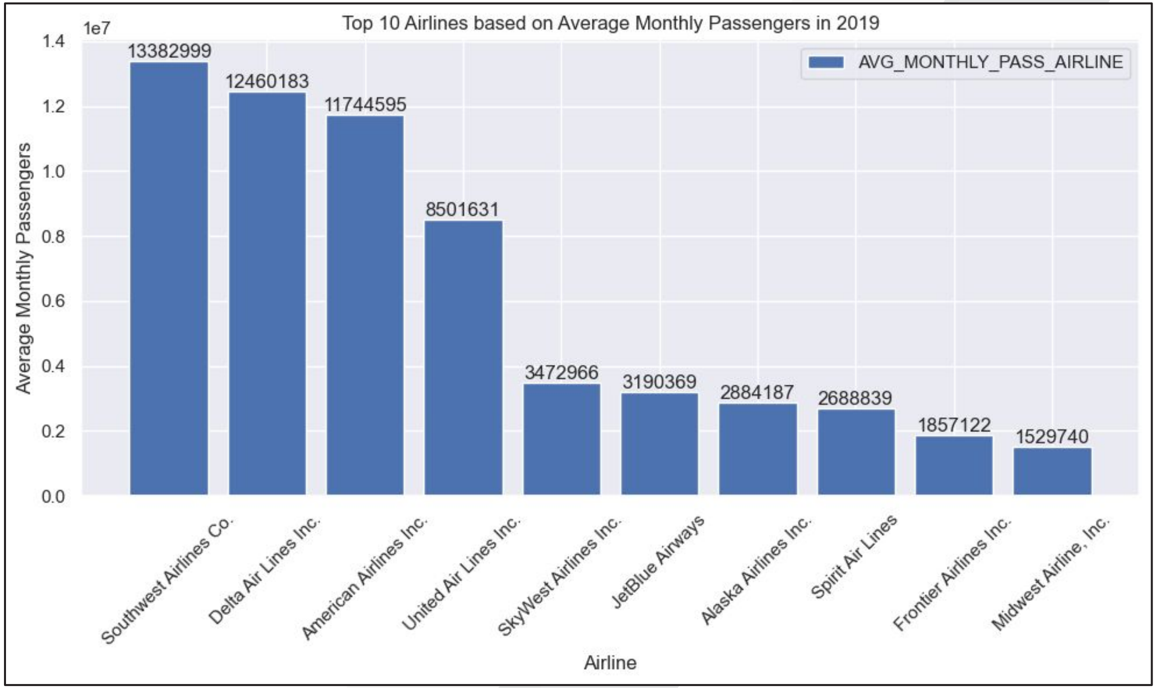
Most number of flight delays:

Southwest Airlines Co.

Possible reasons:

- Budget airline
- Popular

Most Popular Airlines



1. Southwest Airlines Co.
2. Delta Air Lines Inc.
3. American Airlines Inc.

- Major American airlines

Airline

03 Models

1st

Decision Tree

2nd

Random Forest

3rd

AdaBoost

Airline

4th

| XGBoost

5th

| Naive Bayes

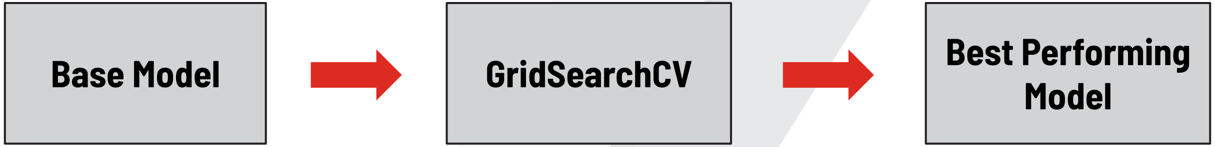
Comparison Summary

Metrics used for models

Accuracy	% of correct classifications
Precision	% of true positive/(true pos + false pos)
Recall	% of true positive/(true pos + false neg)
F1 Score	Mean of precision and recall

Models

Decision Tree Classifier



Default parameters

criterion: gini
max_depth: 4

Hyperparameter Tuning using GridSearchCV:

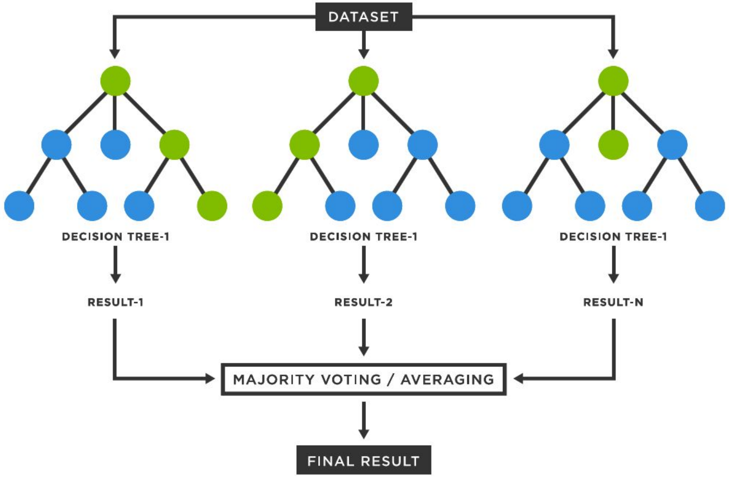
- To obtain the best parameters to generate the model with the highest accuracy

Accuracy	81.3%
Precision	67.0%
Recall	0%
F1-Score	0%

Random Forest Classifier

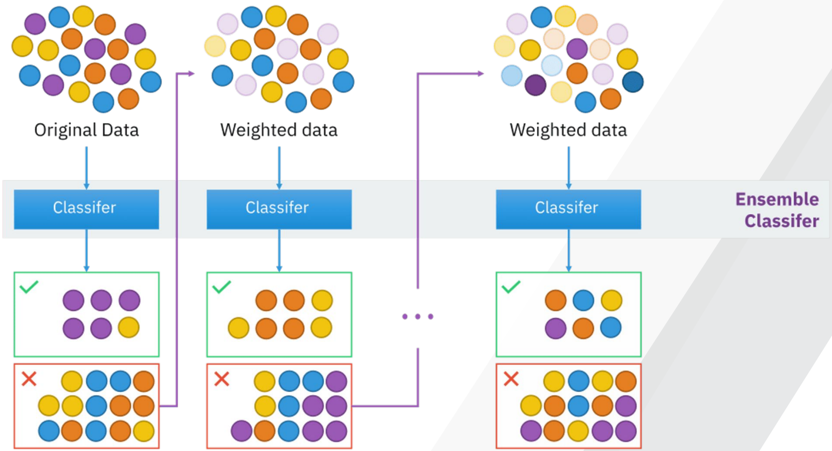
Base Model → K-folds Cross Validation

- Combination of decision trees
- Prevent overfitting
- Result = avg no. of models, evaluated k times

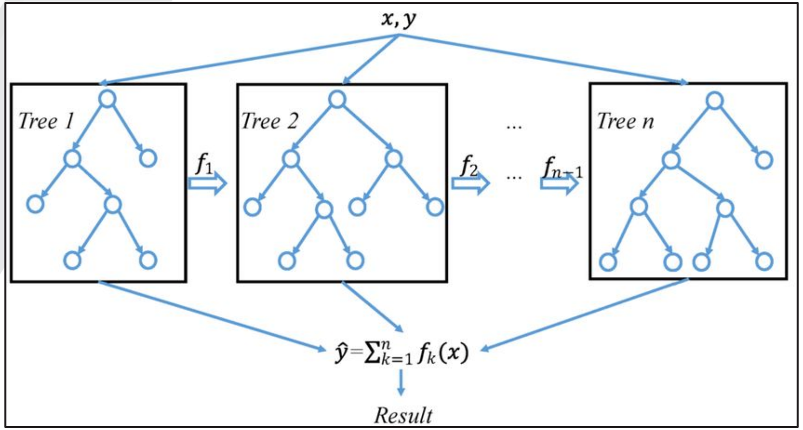


Models

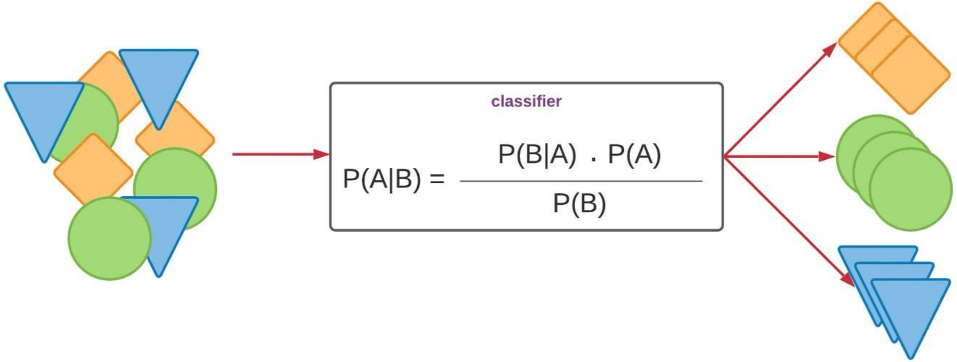
AdaBoost Classifier



XGBoost Classifier

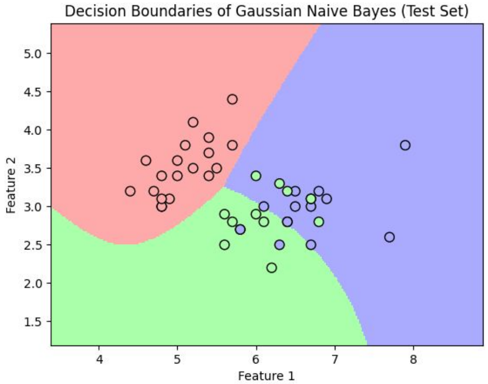


Naive Bayes Classifier



classifier

$$P(A|B) = \frac{P(B|A)\ .\ P(A)}{P(B)}$$

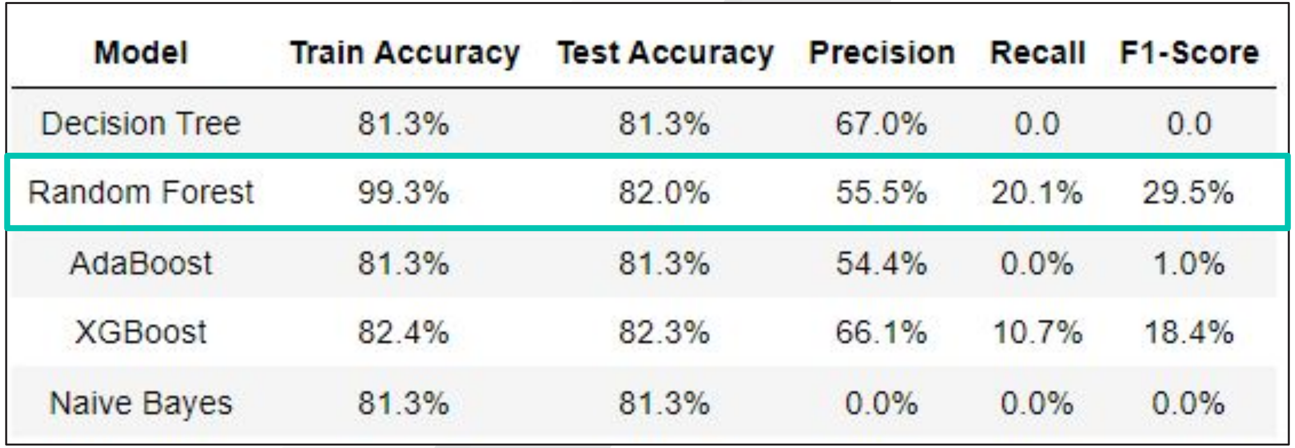


Decision Boundaries of Gaussian Naive Bayes (Test Set)

- Supervised learning
- Probabilistic approach + Gaussian distribution

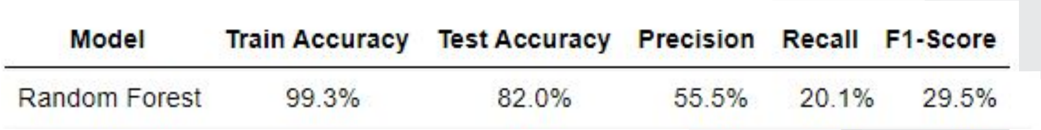
Models

Summary of Classification Models

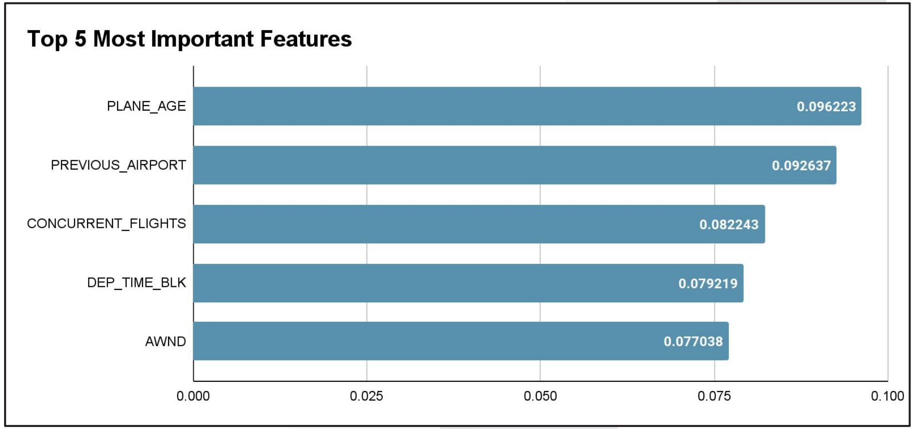


Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Decision Tree	81.3%	81.3%	67.0%	0.0	0.0
Random Forest	99.3%	82.0%	55.5%	20.1%	29.5%
AdaBoost	81.3%	81.3%	54.4%	0.0%	1.0%
XGBoost	82.4%	82.3%	66.1%	10.7%	18.4%
Naive Bayes	81.3%	81.3%	0.0%	0.0%	0.0%

Feature Importances on the Best Model



Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Random Forest	99.3%	82.0%	55.5%	20.1%	29.5%



1. Age of Plane
2. Previous Airport
3. Concurrent Flights
4. Departure Time Block
5. Max Wind Speed

Airline

Data Driven Insights & Recommendations

Travellers

Choose less popular seasons

Airlines

Prioritize newer aircraft models

Airport

Interact with each other

Thank You !

References:

- Slotnick, D. (2020, February 19). Why Your Airplane (Still) Might Be Delayed. The New York Times. Retrieved from https://www.nytimes.com/2020/02/19/business/air-travel-delays-airlines.html
- A general architecture of XGBoost. (n.d.). In ResearchGate. Retrieved from https://www.researchgate.net/figure/A-general-architecture-of-XGBoost_fig3_335483097
- Naive Bayes. (n.d.). In scikit-learn: Machine Learning in Python. Retrieved from https://scikit-learn.org/stable/modules/naive_bayes.html
- Dancerworld60. (n.d.). Demystifying Naïve Bayes—Simple Yet Powerful for Text Classification. Medium. Retrieved from https://medium.com/@dancerworld60/demystifying-na%C3%AFve-bayes-simple-yet-powerful-for-text-classification-ad92b14a5c7