NATIONAL CENTRE FOR
SCIENTIFIC RESEARCH "DEMOKRITOS"

Machine Learning

Stefopoulos Andreas

London – Heathrow Departure Delays

(*threshold 20 min*)

Supervisor : Theodoros Giannakopoulos

Phases of Machine Learning

The steps applied to the project are the following:

- **Phase 0:** Data Gathering – Initial Cleansing
- **Phase 1:** Data Loading and Preparation – Final Cleansing
- **Phase 2:** Data Exploration – Statistical Analysis
- **Phase 3:** Models – Evaluation
- **Phase 4:** Final Validation with current data
- **Phase 5:** Results Communication and Decision Making

Phase 0 (Data Gathering – Initial Cleansing)**:**

- Flightradar24 scrapper development
- Testing
- Data cleansing
- Data re-formatting
- Make .exe file of scrapper
- Run on daily basis (set as task) and append new data to a .csv file

First 10 rows sample of the .csv file generated

	FlightNo	ActualTimeDep	Status	AC Model	Lat Arr	Lon Arr	Dest Country	ScheduledTimeArr	ActualTimeArr	ActualTimeDep	Date	ScheduledTimeDep
0	LY316	Departed 14:48	departed	B789	32.011379	34.886662	IL	19:05:00	18:54:57	14:48:40	10-12-20	14:20:00
1	BA227	Departed 14:41	departed	B789	33.636719	-84.428001	US	0:05:00	23:16:44	14:41:48	10-12-20	14:20:00
2	EI165	Departed 14:40	departed	A320	53.421379	-6.27	IE	15:45:00	15:31:00	14:40:45	10-12-20	14:15:00
3	BA1448	Departed 14:47	departed	A320	55.950001	-3.3725	GB	15:35:00	15:40:47	14:47:53	10-12-20	14:10:00
4	BA818	Departed 14:26	departed	A320	55.616959	12.645637	DK	16:00:00	15:48:44	14:26:14	10-12-20	14:10:00
5	EW7461	Departed 14:21	departed	A320	53.630379	9.988228	DE	15:45:00	15:23:44	14:21:02	10-12-20	14:05:00
6	BA209	Departed 14:37	departed	B789	25.793249	-80.290497	US	0:10:00	23:50:39	14:37:54	10-12-20	14:05:00
7	BA35	Departed 14:11	departed	B788	12.99441	80.180511	IN	0:05:00	23:37:10	14:11:42	10-12-20	14:00:00
8	SV112	Departed 14:24	departed	B77W	21.67956	39.156528	SA	19:50:00	19:53:45	14:24:07	10-12-20	14:00:00

Phase 1 (Data Loading and Preparation):

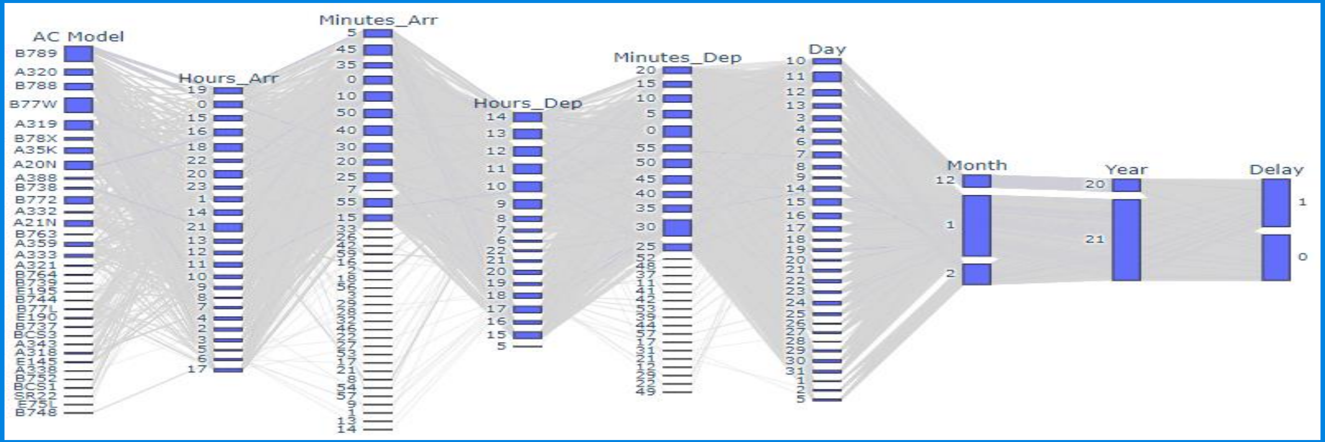
- Data Loading from GitHub
- Data Cleansing
  - Drop first column (index)
  - Drop duplicates
  - Keep only flights with status "departed"
  - Feature engineering by keeping first 2 letters of FlightNo as the name of the airline
  - Check for missing values
  - Fix issue with year change and different format of data from scrapper
- Dependent Variable creation (threshold 20min of delay)
  - Calculate the difference between schedule and actual time departure and if minutes>20 then we have delay (delay=1, class 1) else no delay (delay = 0, class 0)

First rows sample of the generated dataset

	FlightNo	AC Model	Lat Arr	Lon Arr	Dest Country	Hours_Arr	Minutes_Arr	Hours_Dep	Minutes_Dep	Day	Month	Year	Airline	Delay
0	LY316	B789	32.011379	34.886662	IL	19	5	14	20	10	12	20	LY	1
1	BA227	B789	33.636719	-84.428001	US	0	5	14	20	10	12	20	BA	1
2	EI165	A320	53.421379	-6.270000	IE	15	45	14	15	10	12	20	EI	1
3	BA1448	A320	55.950001	-3.372500	GB	15	35	14	10	10	12	20	BA	1
4	BA818	A320	55.616959	12.645637	DK	16	0	14	10	10	12	20	BA	0

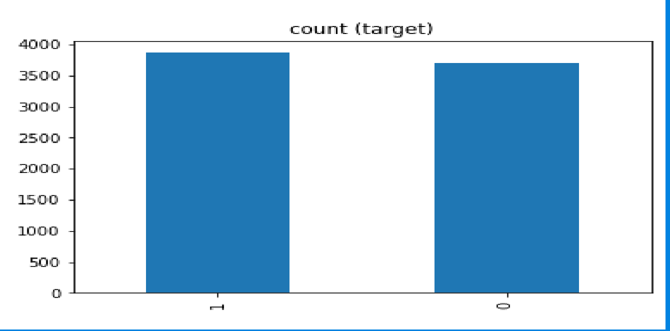
Phase 2 (Data Exploration – Statistical Analysis):

- **Statistical Analysis:**
  - Value allocation



  - Checking balance of dataset

```
1     51.05117
0     48.94883
Name: Delay, dtype: float64
class 0: 3702
class 1: 3861
```



Phase 2 (Data Exploration – Statistical Analysis):

- Calculating mean values of every column in every class (numerical only)

	0	1
Lat Arr	37.926836	36.164094
Lon Arr	0.365289	4.849247
Hours_Arr	13.399514	12.746957
Minutes_Arr	28.406537	28.462574
Hours_Dep	13.360886	12.913235
Minutes_Dep	26.282820	27.277389
Day	13.938682	14.577053
Month	2.030524	3.217301
Year	20.927337	20.816887
Delay	0.000000	1.000000

- Counting the frequency each categorical value appears in every class for "Airline", "AC Model" and "FlightNo" *

Delay	0	1
Airline		
0B	20	31
A3	27	20
AA	325	77
AC	50	45
AF	70	84
...	...	...
UK	3	13
UL	16	4
VS	184	188
WB	8	2
WY	24	6

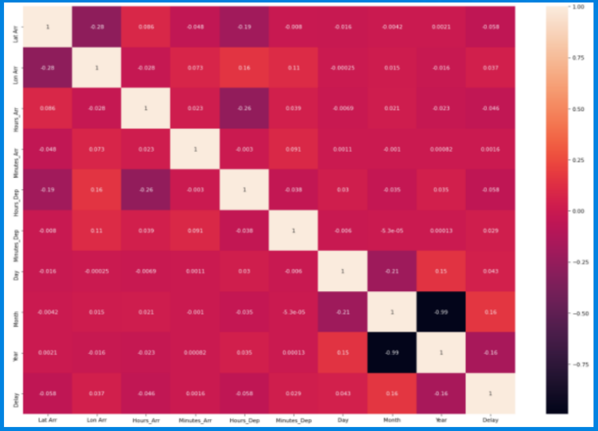
Delay	0	1
AC Model		
A20N	307	336
A21N	200	217
A318	54	21
A319	339	383
A320	245	219
A321	35	33
A332	22	45
A333	110	100
A338	0	3
A343	7	6
A359	176	108

Delay	0	1
FlightNo		
0B1332	13	15
0B1432	7	16
A3601	11	8
A3603	16	12
AA105	48	4
...	..	..
VS525	2	1
VS623	6	2
WB701	8	2
WY104	23	5
WY2104	1	1

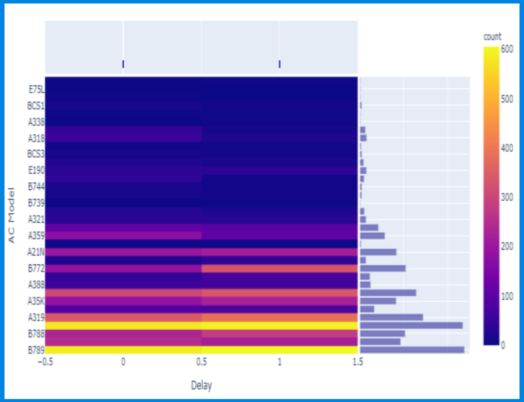
** images represent samples of the data*

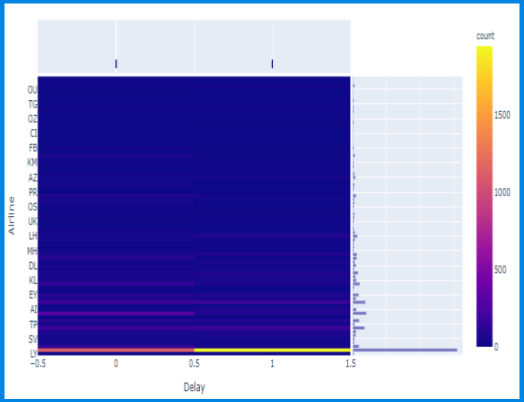
Phase 2 (Data Exploration – Statistical Analysis):

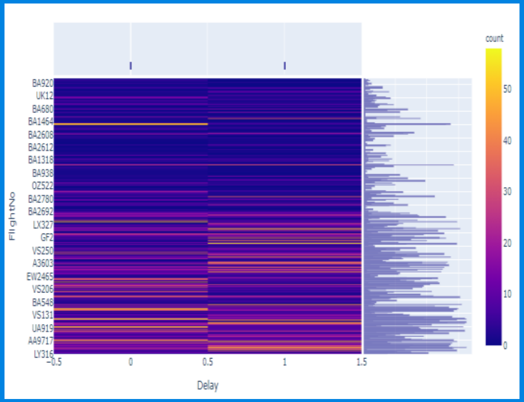
- Calculating correlations (numerical only). Low correlated features with the dependent variable because of their origin.



- Counting the frequency each categorical value appears in every class for “Airline”, “AC Model” and “FlightNo” (with plotly) *



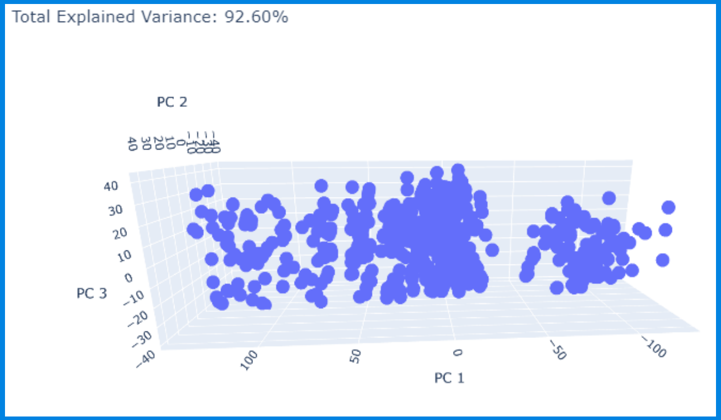




*** images represent samples of the data***

Phase 2 (Data Exploration – Statistical Analysis):

- Performing PCA for visualization purposes (X_train)



- Final description of the dataset before Standard Scaling

	Lat Arr	Lon Arr	Hours_Arr	Minutes_Arr	Hours_Dep	Minutes_Dep	Day	Month	Year	Delay
count	7563.000000	7563.000000	7563.000000	7563.000000	7563.000000	7563.000000	7563.000000	7563.000000	7563.000000	7563.000000
mean	37.026936	2.654402	13.066376	28.435145	13.132355	26.790559	14.264578	2.636388	20.870951	0.510512
std	15.235396	60.996570	7.047490	17.631753	3.843894	17.262808	7.422718	3.627091	0.335277	0.499923
min	-34.822201	-123.183998	0.000000	0.000000	5.000000	0.000000	1.000000	1.000000	20.000000	0.000000
25%	25.793249	-16.572399	8.000000	10.000000	10.000000	10.000000	9.000000	1.000000	21.000000	0.000000
50%	40.639751	6.108950	15.000000	30.000000	13.000000	30.000000	13.000000	1.000000	21.000000	1.000000
75%	49.012516	46.698769	19.000000	45.000000	16.000000	40.000000	19.000000	2.000000	21.000000	1.000000
max	63.985001	139.779602	23.000000	59.000000	22.000000	57.000000	31.000000	12.000000	21.000000	1.000000

Phase 3 (Models - Evaluation):

- Before scaling, we call the scrapper to gather data for the flights departed the last 24 hours, and we align the two sets in order to be able to predict in the end of the script
- Scaling data with StandardScaler (subtraction of mean value and division with standard deviation

Models Tested and Evaluated:

- SVM
- Decision Tree
- Random Forest
- kNN
- XGBoost
- Logistic Regression
- Simple ANN Classifier

Evaluation Metrics

Actions:

- Set hyperparameters of models (after testing)
- model.fit(X_train) data
- Predict X_test and print confusion matrix, classification report, score of train and test (accuracy)
- Perform 5-Fold cross validation and save the mean f1, precision and recall
- Predict probabilities and save recall, precision and auc and plot them in the end

Phase 3 (Models - Evaluation):

Results of all methods used for every model:

- SVM

```
[[709 395]
 [371 794]]
0.6624063464081092
              precision    recall  f1-score   support

           0       0.66      0.64      0.65      1104
           1       0.67      0.68      0.67      1165

    accuracy                           0.66      2269
   macro avg       0.66      0.66      0.66      2269
weighted avg       0.66      0.66      0.66      2269

Training set score: 0.721
Test set score: 0.662
```

```
Mean F1 Score = 64.33% - SD F1 Score = 3.48%
Mean Recall Score = 66.30% - SD Recall = 9.45%
Mean Precision Score = 63.79% - SD Precision = 3.91%
```

```
SVM kernel: f1=0.675 auc=0.752
```

- Decision Tree

```
[[906 198]
 [203 962]]
0.8232701630674306
              precision    recall  f1-score   support

           0       0.82      0.82      0.82      1104
           1       0.83      0.83      0.83      1165

    accuracy                           0.82      2269
   macro avg       0.82      0.82      0.82      2269
weighted avg       0.82      0.82      0.82      2269

Training set score: 1.000
Test set score: 0.823
```

```
Mean F1 Score = 61.51% - SD F1 Score = 5.26%
Mean Recall Score = 59.57% - SD Recall = 8.59%
Mean Precision Score = 64.02% - SD Precision = 2.63%
```

```
Decision Tree: f1=0.828 auc=0.872
```

Phase 3 (Models - Evaluation):

- Random Forest

```
[[911 193]
 [247 918]]
0.8060819744380785
              precision    recall  f1-score   support

           0       0.79      0.83      0.81      1104
           1       0.83      0.79      0.81      1165

    accuracy                           0.81      2269
   macro avg       0.81      0.81      0.81      2269
weighted avg       0.81      0.81      0.81      2269

Training set score: 0.984
Test set score: 0.806
```

```
Mean F1 Score = 60.86% - SD F1 Score = 7.11%
Mean Recall Score = 56.35% - SD Recall = 12.88%
Mean Precision Score = 68.56% - SD Precision = 2.60%
```

```
Random Forest: f1=0.807 auc=0.875
```

- kNN

```
[[754 350]
 [315 850]]
0.7069193477302776
              precision    recall  f1-score   support

           0       0.71      0.68      0.69      1104
           1       0.71      0.73      0.72      1165

    accuracy                           0.71      2269
   macro avg       0.71      0.71      0.71      2269
weighted avg       0.71      0.71      0.71      2269

Training set score: 0.874
Test set score: 0.707
```

```
Mean F1 Score = 59.54% - SD F1 Score = 5.34%
Mean Recall Score = 58.17% - SD Recall = 11.23%
Mean Precision Score = 62.86% - SD Precision = 2.86%
```

```
KNN: f1=0.719 auc=0.811
```

Phase 3 (Models - Evaluation):

- XGBoost

```
[[747 357]
 [325 840]]
0.6994270603790216
              precision    recall  f1-score   support

           0       0.70      0.68      0.69      1104
           1       0.70      0.72      0.71      1165

    accuracy                           0.70      2269
   macro avg       0.70      0.70      0.70      2269
weighted avg       0.70      0.70      0.70      2269

Training set score: 0.722
Test set score: 0.699
```

```
Mean F1 Score = 60.10% - SD F1 Score = 9.56%
Mean Recall Score = 57.02% - SD Recall = 15.90%
Mean Precision Score = 67.99% - SD Precision = 6.86%
```

```
XGB: f1=0.711 auc=0.787
```

- Logistic Regression

```
[[733 371]
 [347 818]]
0.6835610401057735
              precision    recall  f1-score   support

           0       0.68      0.66      0.67      1104
           1       0.69      0.70      0.69      1165

    accuracy                           0.68      2269
   macro avg       0.68      0.68      0.68      2269
weighted avg       0.68      0.68      0.68      2269

Training set score: 0.736
Test set score: 0.684
```

```
Mean F1 Score = 64.94% - SD F1 Score = 6.67%
Mean Recall Score = 65.46% - SD Recall = 12.61%
Mean Precision Score = 66.13% - SD Precision = 3.29%
```

```
Logistic Regression: f1=0.695 auc=0.787
```

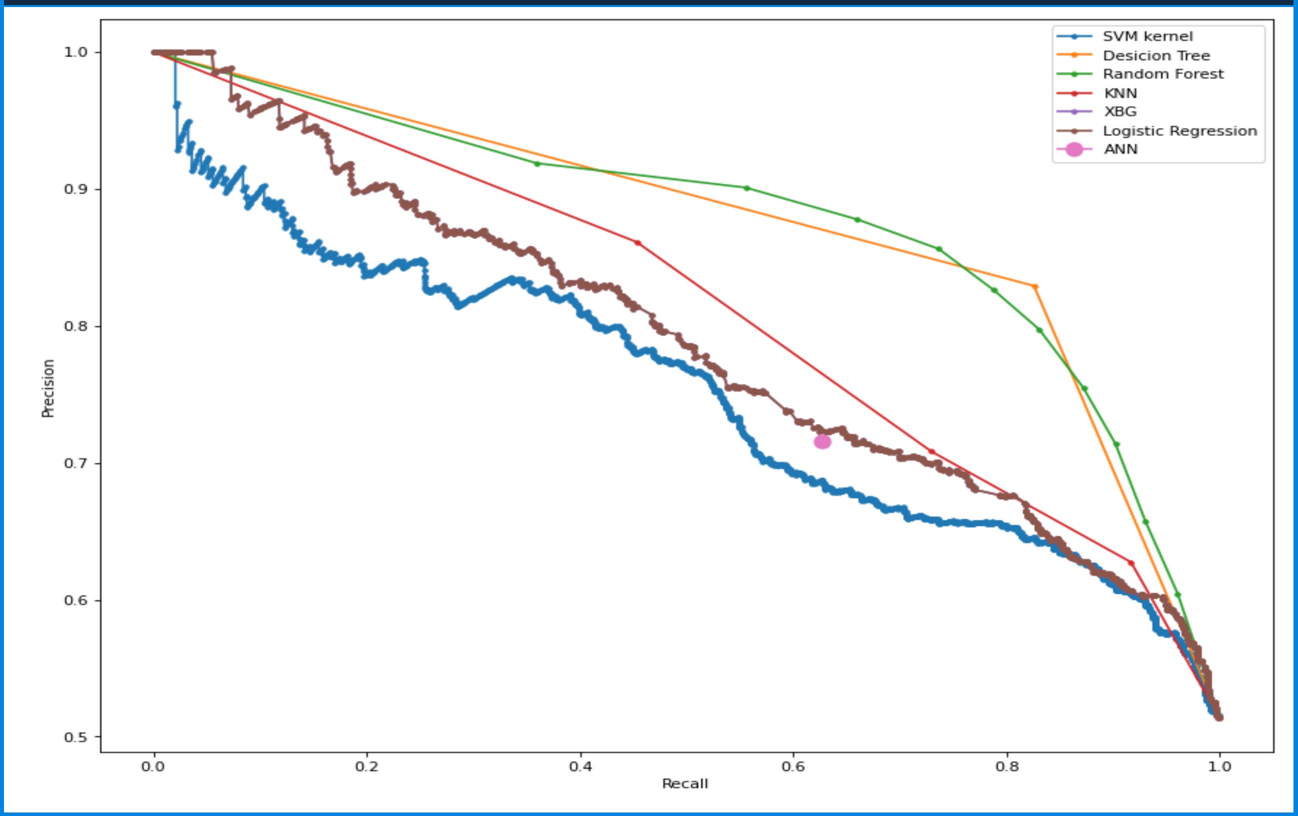
- ANN

```
[[814 290]
 [435 730]]
Accuracy: 0.680476
Precision: 0.715686
Recall: 0.626609
F1 score: 0.668192
```

Phase 4 (Final Validation with current data):

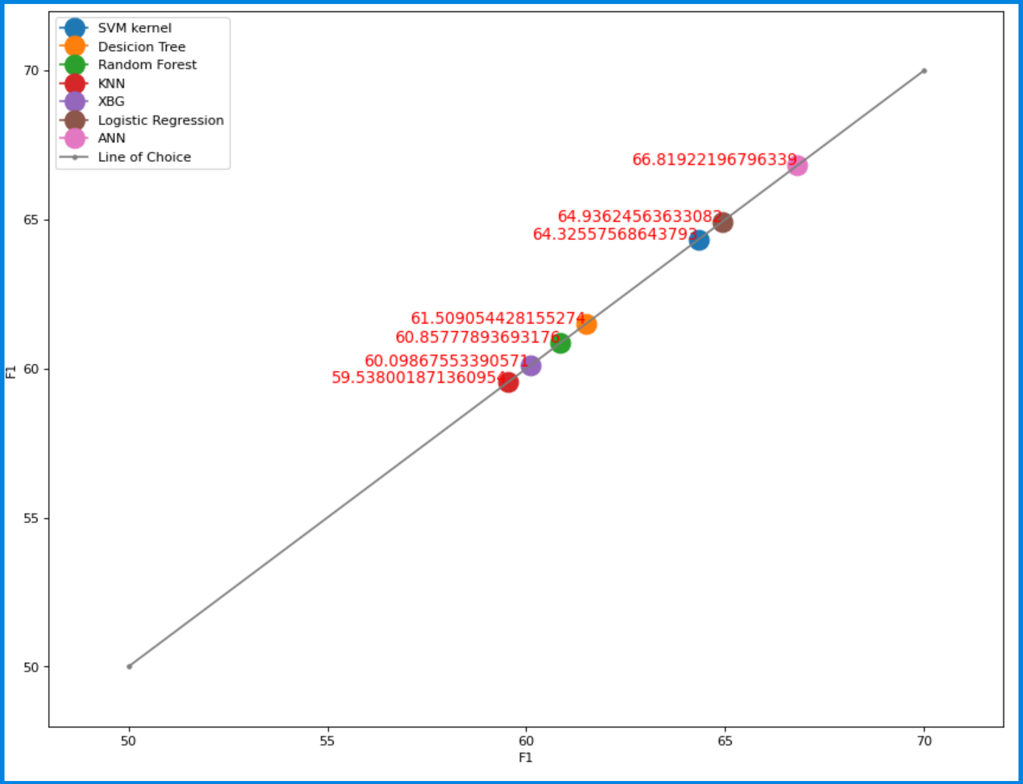
Before performing final validation, we have to choose a model. The criteria I used was the mean f1 score from the 5-Fold cross validation. Firtsly, i plotted the recall and precision for all models and then the mean f1 scores as shown below:

Recall and Precision plot



Phase 4 (Final Validation with current data):

Mean f1 scores:



In the last step, we can perform the validation and print the output of our model with data that happened just 24 hours ago. The model I chose to use is Random Forest due to lack of numerical features and nice performance on recall-precision outputs

Phase 5 (Results Communication & Decision Making)**:**

Finally, flight delays is a parameter that gets affected by many factors, and there is also a timeseries issue due to heavy traffic during specific seasons. As a result of the above, we can have a better prediction with more data in respect of the timeline we investigate.

Thank you

for your time