Project Report Template

1 INTRODUCTION

1.1 Overview

the present world, the major components of any transportation system include passenger airline, cargo airline, and air traffic control system. With the passage of time, nations around the world have tried to evolve numerous techniques of improving the airline transportation system. This has brought drastic change in the airline operations. Flight delays occasionally cause inconvenience to the modern passengers [1]. Every year approximately 20% of airline flights are canceled or delayed, costing passengers more than 20 billion dollars in money and their time.

1.2 Purpos

Flight delay is a significant problem that negatively impacts industry and costs billions of dollars each year. Most existing studies investigated this issue using various methods based on applying machine learning methods to predict the flight delay. However, due to the highly dynamic environments of the aviation industry, relying only on single route of airport may not be sufficient and applicable to forecast the future of flights. The purpose of this project is **to analyze a broader scope of factors which may potentially influence the flight delay it compares several machine learning-based models in designed generalized flight delay prediction tasks. In this project we have used flight delay dataset from US Department of Transportation (DOT) to predict flight delays. We have used supervised learning algorithms to predict flight departure delay** and then model evaluation is done to get best model and our model can identify which features were more important when predicting flight delays.

Many factors can come into play when a traveler chooses a flight. At Amadeus, we know that price is one of the main factors, but we also wondered what other criteria might be relevant?

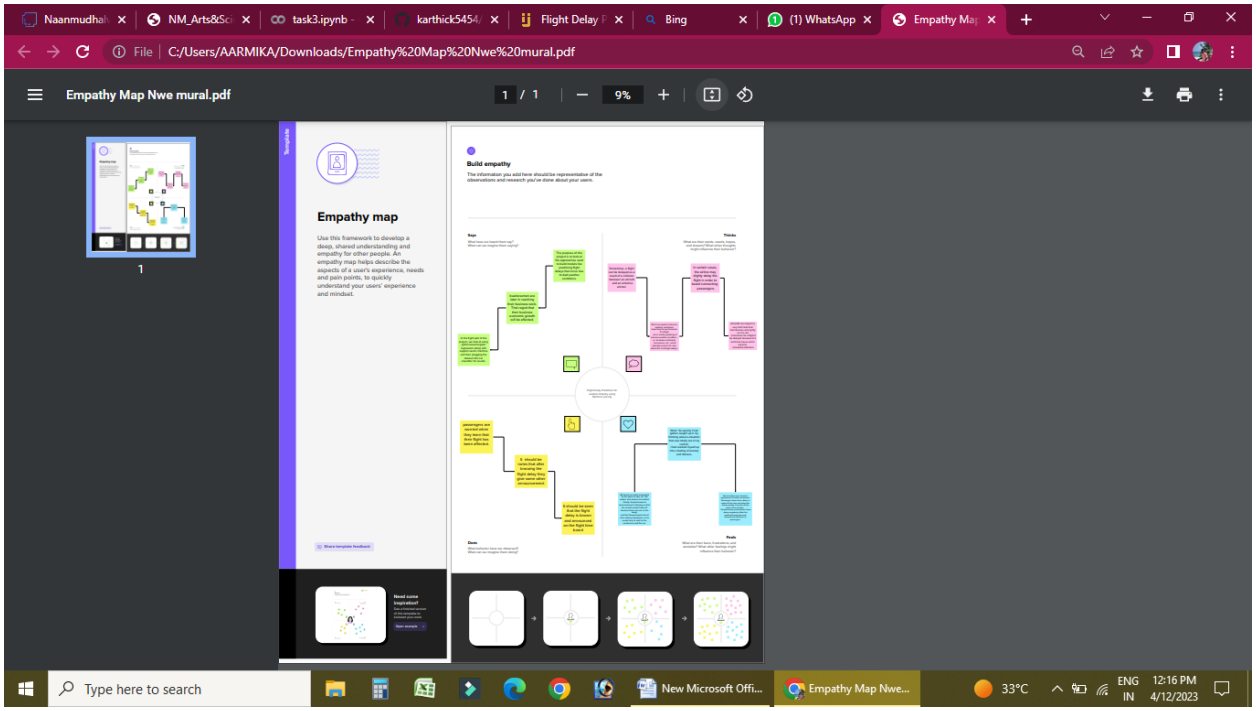
Flight delays and cancellations are two of the most common travel headaches, but what if you could know in advance the probability of a flight being delayed? Would you still book it? What if it was a connecting flight?

Knowing the probability of flight delay or cancellation is a crucial tool for travelers, so we set about creating a **model to predict long-term flight delays**. Rather than looking at disruptions caused by punctual factors like weather, we wanted to see which flights and itineraries had the highest probability of delays or cancellations over time.

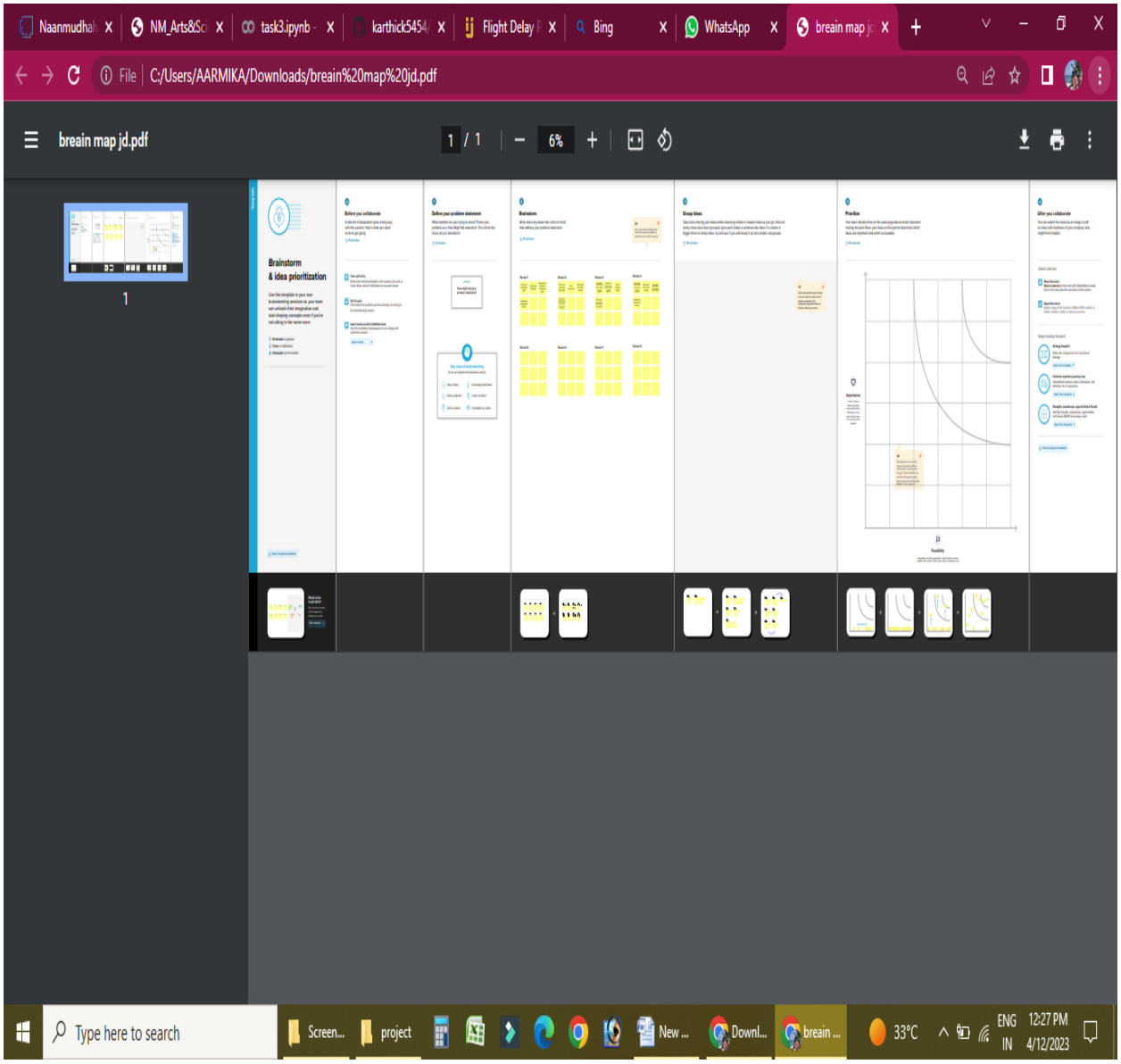
In this article, we'll explain **how we created our flight delay predictor** using machine learning and how you can integrate these insights into your app using the [Flight Delay Prediction API]

2.problem definition & design thinking

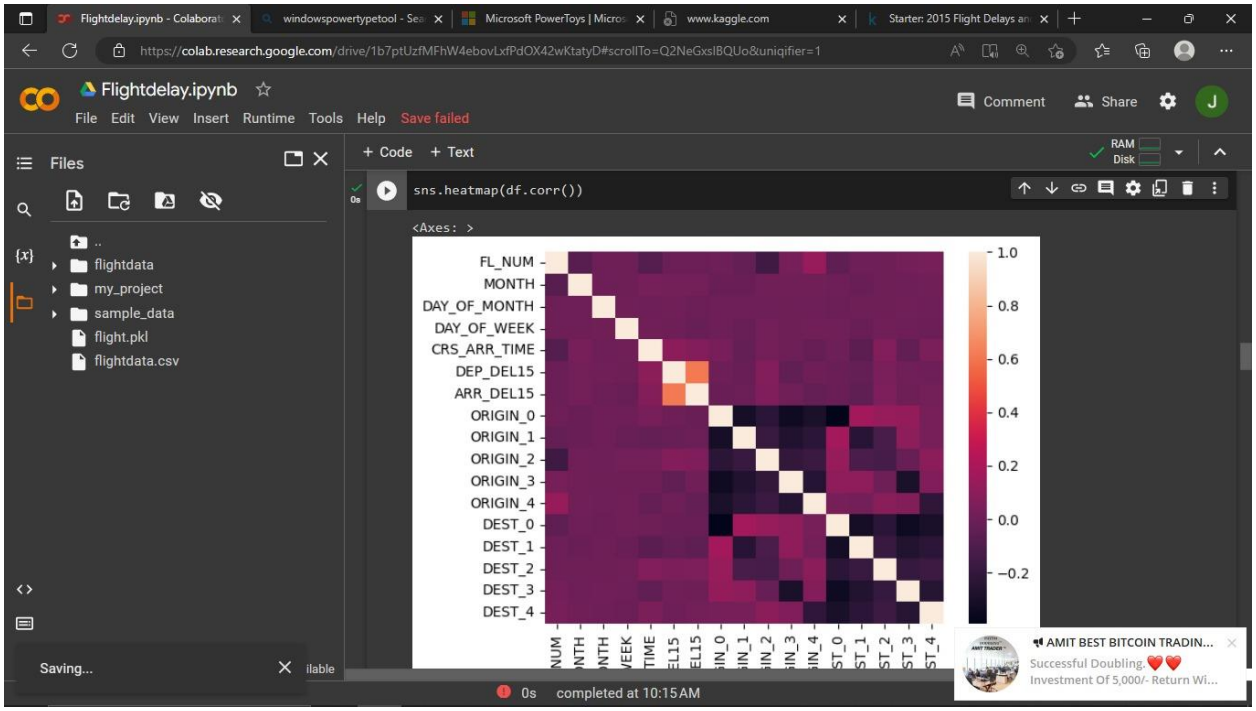
2.1 EMPATHY MAP

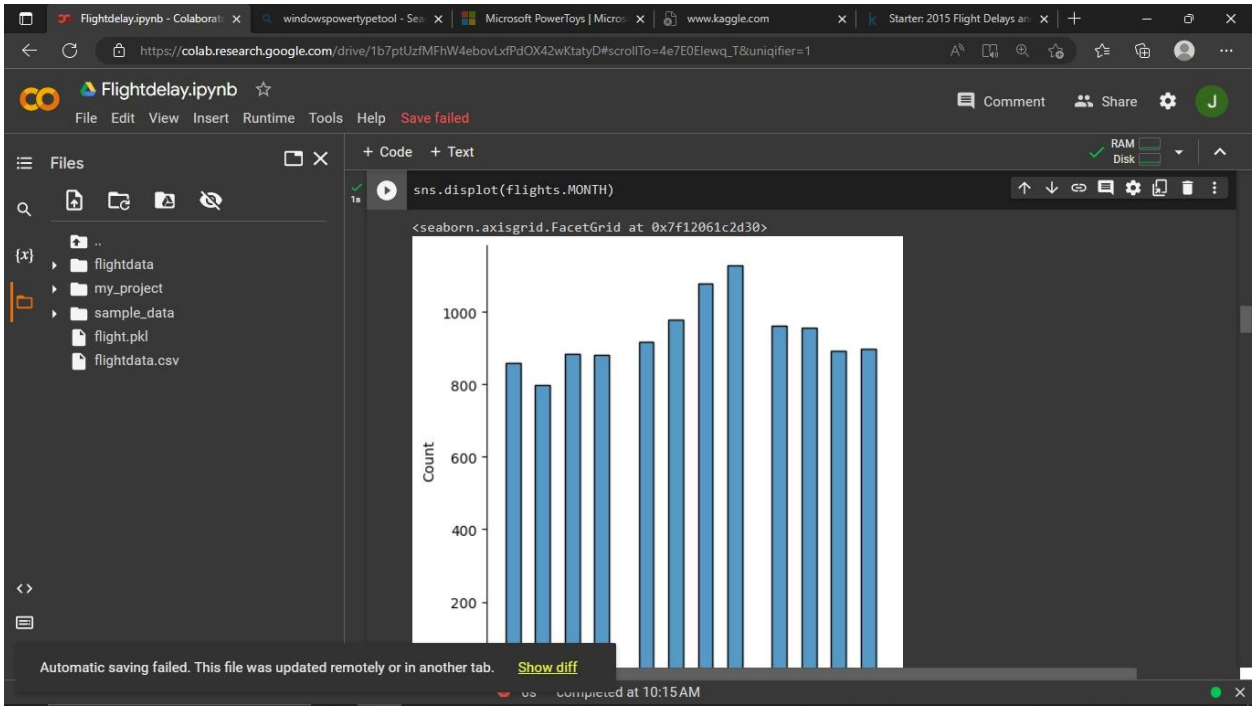


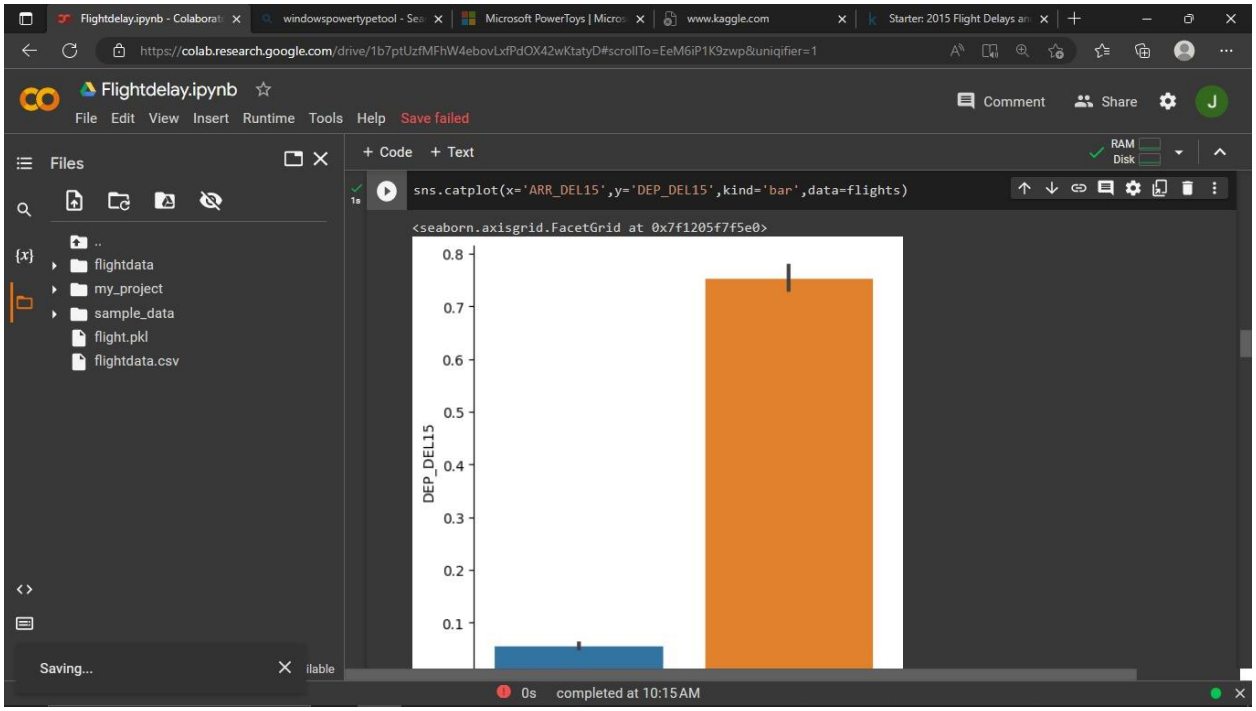
2.2 Ideation & brainstorming map

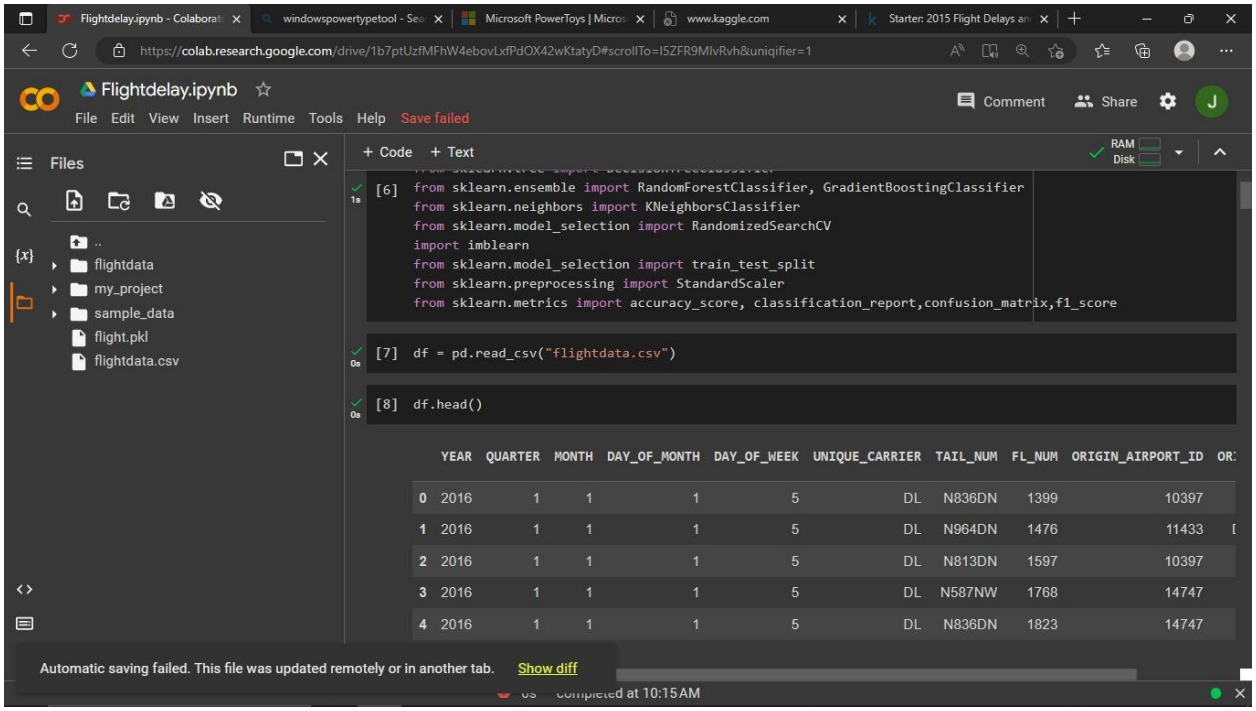


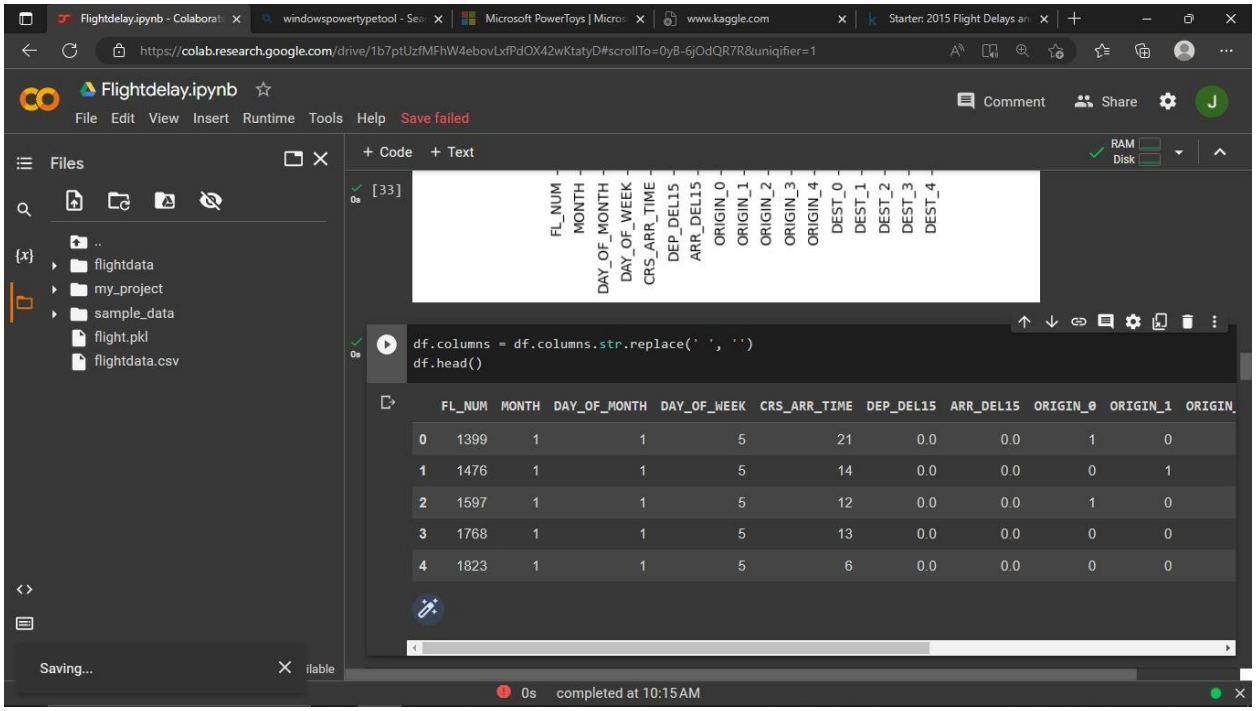
3.RESULT

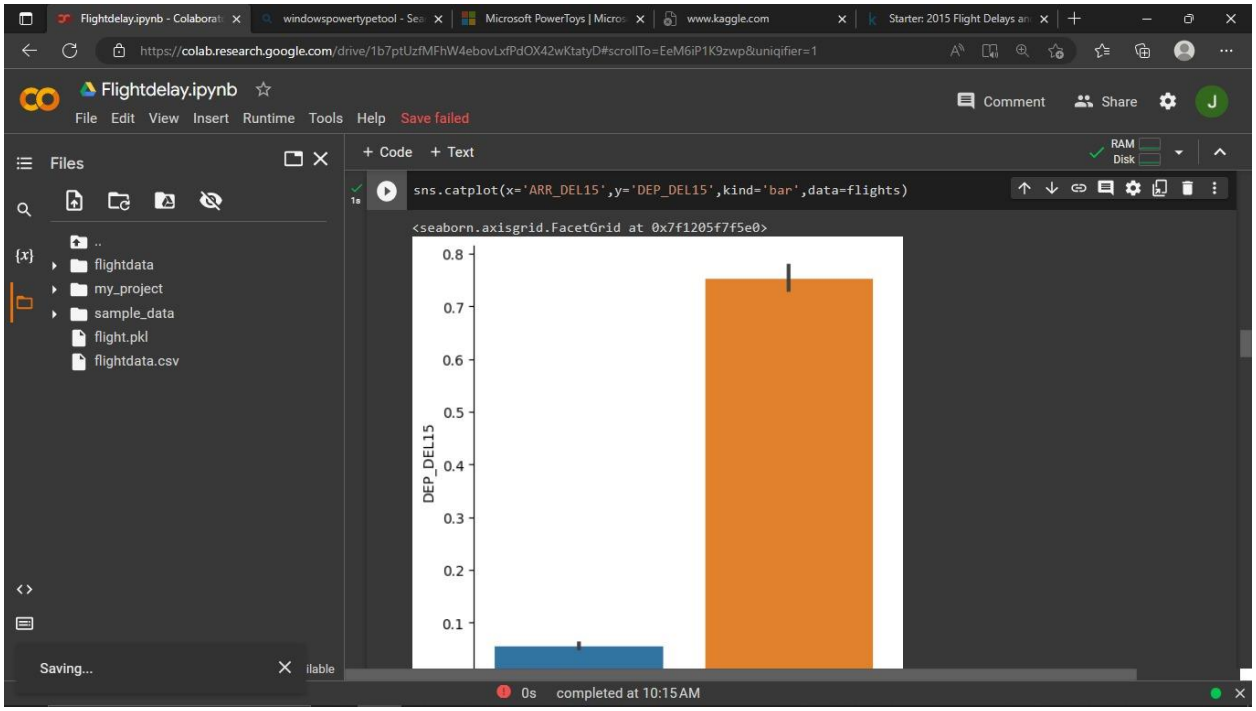

Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
+ Code + Text
RAM
Disk
Files
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
sns.heatmap(df.corr())
<Axes: >
FL_NUM
MONTH
DAY_OF_MONTH
DAY_OF_WEEK
CRS_ARR_TIME
DEP_DEL15
ARR_DEL15
ORIGIN_0
ORIGIN_1
ORIGIN_2
ORIGIN_3
ORIGIN_4
DEST_0
DEST_1
DEST_2
DEST_3
DEST_4
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Saving...
0s completed at 10:15AM

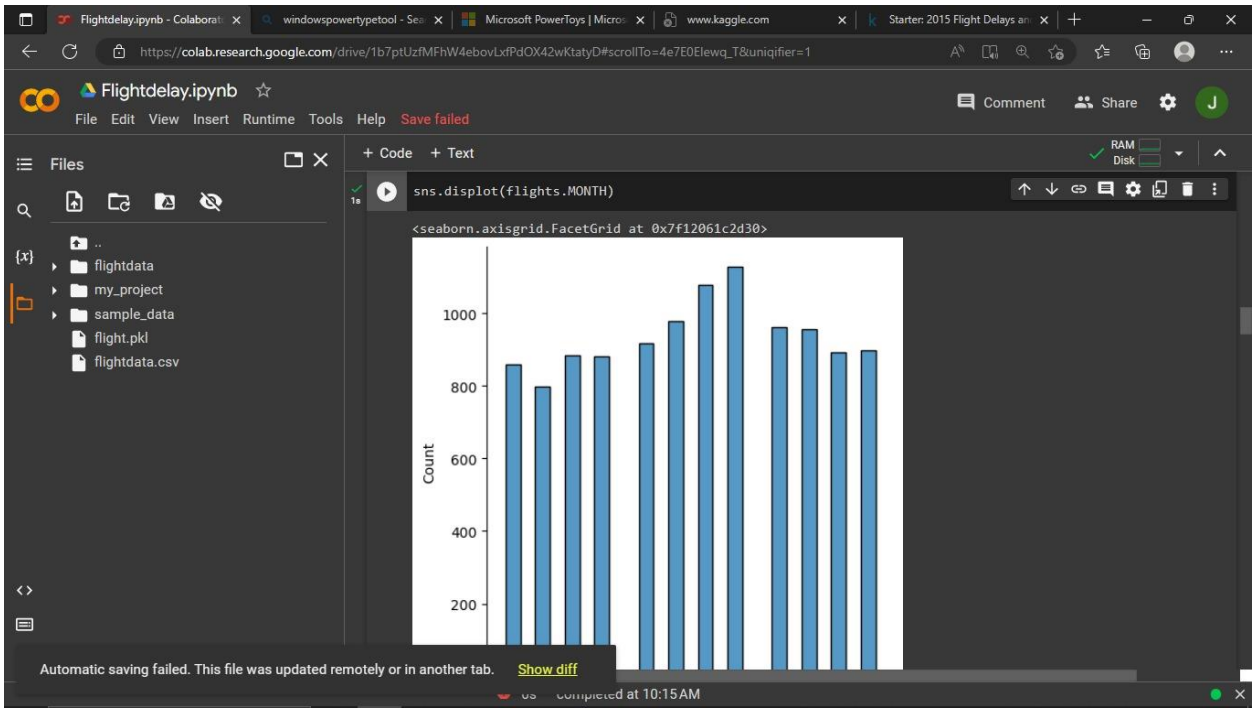

Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
+ Code + Text
RAM
Disk
Files
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
sns.displot(flights.MONTH)
<seaborn.axisgrid.FacetGrid at 0x7f12061c2d30>
Count
1000
800
600
400
200
Automatic saving failed. This file was updated remotely or in another tab.
Show diff

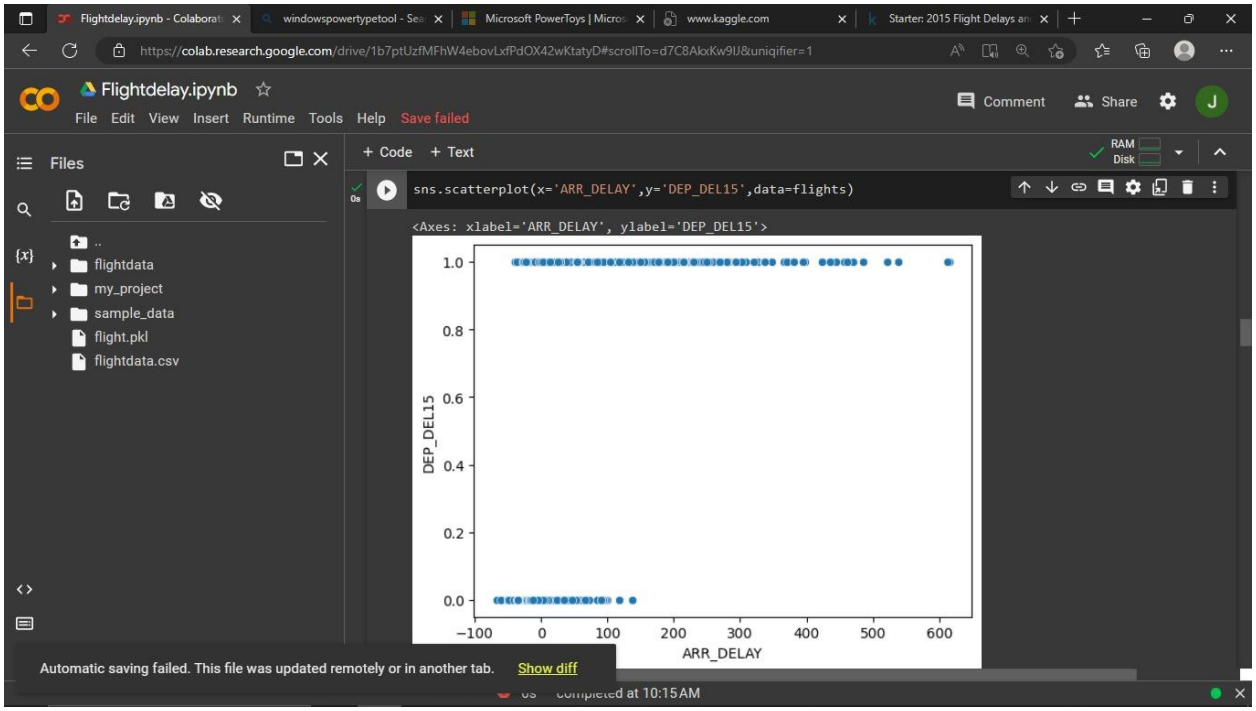

Flightdelay.ipynb - Colaborat
windowspowertypetool - Sea
Microsoft PowerToys | Micros
www.kaggle.com
Starter: 2015 Flight Delays an
https://colab.research.google.com/drive/1b7ptUzfMFhW4ebovLxfPdOX42wKtatyD#scrollTo=EeM6iP1K9zwp&uniqifier=1
Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
J
Files
+ Code + Text
RAM
Disk
sns.catplot(x='ARR_DEL15',y='DEP_DEL15',kind='bar',data=flights)
<seaborn.axisgrid.FacetGrid at 0x7f1205f7f5e0>
DEP_DEL15
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
..
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
Saving...
ilable
0s completed at 10:15 AM

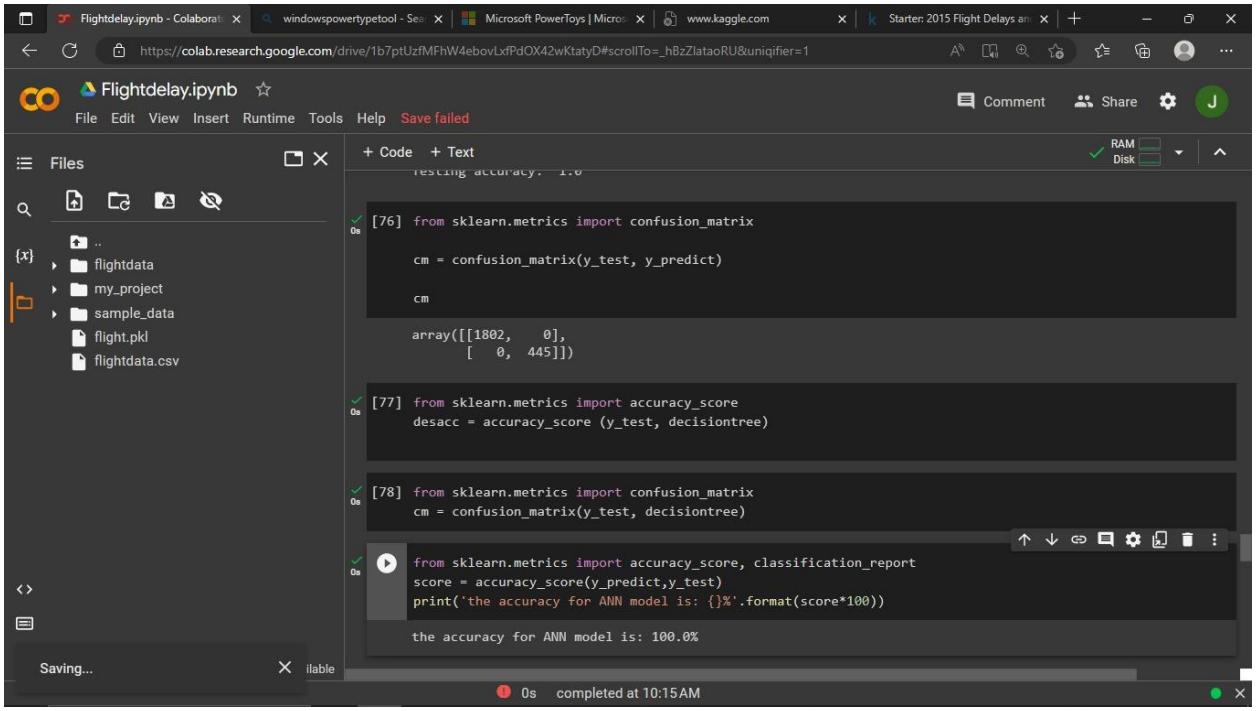

Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
Files
+ Code + Text
RAM
Disk
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
[6] from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import RandomizedSearchCV
import imblearn
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report,confusion_matrix,f1_score
[7] df = pd.read_csv("flightdata.csv")
[8] df.head()
YEAR QUARTER MONTH DAY_OF_MONTH DAY_OF_WEEK UNIQUE_CARRIER TAIL_NUM FL_NUM ORIGIN_AIRPORT_ID
0 2016 1 1 1 5 DL N836DN 1399 10397
1 2016 1 1 1 5 DL N964DN 1476 11433
2 2016 1 1 1 5 DL N813DN 1597 10397
3 2016 1 1 1 5 DL N587NW 1768 14747
4 2016 1 1 1 5 DL N836DN 1823 14747
Automatic saving failed. This file was updated remotely or in another tab.
Show diff


Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
Files
+ Code + Text
RAM
Disk
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
[33]
FL_NUM
MONTH
DAY_OF_MONTH
DAY_OF_WEEK
CRS_ARR_TIME
DEP_DEL15
ARR_DEL15
ORIGIN_0
ORIGIN_1
ORIGIN_2
ORIGIN_3
ORIGIN_4
DEST_0
DEST_1
DEST_2
DEST_3
DEST_4
df.columns = df.columns.str.replace(' ', '')
df.head()
FL_NUM MONTH DAY_OF_MONTH DAY_OF_WEEK CRS_ARR_TIME DEP_DEL15 ARR_DEL15 ORIGIN_0 ORIGIN_1 ORIGIN_
0 1399 1 1 5 21 0.0 0.0 1 0
1 1476 1 1 5 14 0.0 0.0 0 1
2 1597 1 1 5 12 0.0 0.0 1 0
3 1768 1 1 5 13 0.0 0.0 0 0
4 1823 1 1 5 6 0.0 0.0 0 0
Saving...
0s completed at 10:15 AM


Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
Files
+ Code + Text
RAM
Disk
sns.catplot(x='ARR_DEL15',y='DEP_DEL15',kind='bar',data=flights)
<seaborn.axisgrid.FacetGrid at 0x7f1205f7f5e0>
DEP_DEL15
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
Saving...
0s completed at 10:15AM


Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
Files
+ Code + Text
RAM
Disk
sns.displot(flights.MONTH)
<seaborn.axisgrid.FacetGrid at 0x7f12061c2d30>
Count
1000
800
600
400
200
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
Automatic saving failed. This file was updated remotely or in another tab.
Show diff


Flightdelay.ipynb - Colaborat
windowspowertypetool - Sea
Microsoft PowerToys | Micros
www.kaggle.com
Starter: 2015 Flight Delays an
https://colab.research.google.com/drive/1b7ptUzfMFhW4ebovLxfPdOX42wKtatyD#scrollTo=d7C8AkxKw9IJ&uniqifier=1
Flightdelay.ipynb
File Edit View Insert Runtime Tools Help Save failed
Comment
Share
Files
+ Code + Text
RAM
Disk
flightdata
my_project
sample_data
flight.pkl
flightdata.csv
sns.scatterplot(x='ARR_DELAY',y='DEP_DEL15',data=flights)
<Axes: xlabel='ARR_DELAY', ylabel='DEP_DEL15'>
DEP_DEL15
ARR_DELAY
Automatic saving failed. This file was updated remotely or in another tab.
Show diff
completed at 10:15 AM



4.Advantages & disadvantages

4.1 Advantages

1. **Better data analytics** not only gives us better insights into flightdelays, it also helps our policy makers make better-informed choices about which strategies are most effective

2. .in addition to that this will be beneficial also for the travelers to understand aircraft departure and arrival dates in various cases, whether they are normal or there are delays if there are natural disasters or fluctuating weather conditions.

3. using python Programming Language for building Machine learning and Deep Learning Models and Tableau for holistic view of aviation industry using storytelling dashboards

4. Even if it is at the airport, you can enjoy your holiday destination for a little longer. Enjoy the local food, write and send a holiday card from the airport or sit back and relax and enjoy the vibrant life at the airport.

5. Reality teaches us that once you are home, you often do not take time to sort out your holiday photos... let alone make a photo album. Are you delayed? Then sort them out to create that photo album back home that you are so looking forward to!

6. Time loss is an annoying side effect... According to Regulation 261/2004 you are entitled to compensation for time loss in the event of an arrival delay of three hours or more. Depending on the flight distance, you are entitled to €250, €400 or €600 per person.

4.2 Disadvantages

1. One of the biggest disadvantages of flying is the cost. It can be very expensive to purchase a plane ticket, especially if you're flying internationally. Sure, budget airlines might offer some cheap flights, but they often come with their own set of problems (more on that below

2. Even for hardened travelers, flying can be a hassle, especially if you're dealing with delays, cancellations, or lost baggage. It can be frustrating to deal with the logistics of air travel, and it's not always a smooth or easy process.

3. Flying can be inconvenient, especially if you have to travel on short notice. It can be difficult to find a flight that fits your schedule, and you may have to deal with unforeseen delays or cancellations. You also need to make sure your plane tickets, passport, and other documents are in order. And did we mention the long lines at security?

4. One of the biggest headaches of flying is dealing with missed connections. If your flight is delayed or canceled, it can throw off your whole travel schedule. When buying tickets, you should leave plenty of time between connecting flights

5. Jet lag is a real problem for many people who fly frequently. It's a disruption of your body's natural sleep cycle, and it can be difficult to adjust to a new time zone. If your air travels involve a long journey, you're likely to experience some jet lag.

5 Applications

Air help App

If you're not up to speed on flyer's rights, AirHelp can help you figure out if you're eligible to receive compensation for a delay or cancellation. Unfortunately, you're out of luck if your flight was canceled or delayed because of a weather-related event. But if your plans are disrupted because of a technical or staffing issue, you might be owed something.

Start by punching in your flight details on the AirHelp website or mobile app. If you're eligible, the company will help you file a claim. AirHelp's mobile app will even scan your boarding pass and automatically upload all the necessary information to make the process as painless as possible.

Follow me on LinkedIn. Check out my website. Send me a secure tip.

6. Conclusion

Flight delays are an important subject in the literature due to their economic and environmental impacts. They may increase costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system. In this context, researchers created flight delay models for delay prediction over the last years, and this work contributes with an analysis of these models from a Data Science perspective. We developed a taxonomy scheme and classified models in respect of detailed components.

Mainly, the taxonomy includes domain and Data Science branches. The former branch categorizes the problem (flight delay prediction) and the scope. The last branch groups methods and data handling. It was observed that the flight delay prediction is classified into two main categories, such as delay propagation and root delay and cancellation. Besides, the scope determines one of the three specific extents: airline, airport, en-route airspace or an ensemble of them. Additionally, considering Data Science branch, we aimed at the datum, by categorizing data sources, dimensions that can be used in the models, and data management techniques to preprocess data and improve prediction models efficiency. We also studied and divided the main methods into five categories: statistical analysis, probabilistic models, network representation, operations research, and machine learning. Those categories have been grouped as their use on specific forecast models for flight delays. Besides the taxonomic scheme, we also presented a timeline with all articles to spot trends and relationships involving the main elements in the taxonomy. In the light of the domain-problem classification, this timeline showed a dominance of delay propagation and root delay over cancellation analysis. Researchers used to focus on statistical analysis and operational research approaches in the past. However, as the data volume grows, we noticed the use of machine learning and data management is increasing significantly. This clearly characterizes a Data Science trend. Researchers from airlines, airports, and academia will require a combination of skills of both domain specialists and data scientists to enable knowledge discovery from flight Big Data.

7. Future scope

Further supportive study is required to correlate all the problem, scope and method for getting most accurate result. Although weather conditions are the major reasons for flight delay, other unprecedented events such as major calamities , natural or man-made can cause major delay in flight.

8.APPENDIX

```
pwd

import pandas as pd

from google.colab import files

upload = files.upload()

df = pd.read_csv("flightdata.csv")

dataset.info()

dataset dataset.drop('Unnamed: 25', axises) dataset.isnull().sun()

Brilter the dataset to eliminate colums that aren't relevant to a predictive model. dataset
dataset[["FL_NUM", "MONTH", "DAY OF MONTH", "DAY OF MEEE", "ORIGIN", "DEST",
"CRS_AIR_TIME","DEP DELIS", "ARR DELSS"}] dataset.isnull()-xum()

dataset["DCP_3"]()

dataset dataset.fillnat("ARK ELIS 3))


dataset dataset. Filina(('DEP_DELLS's 0)) dataset, iloc[377:185]

import math


for index, row in dataset.iterrows():


dataset.loc[index, 'CRS ARR TIME') math.floor(row['CRS ARR_TIME"]/100) ARR_TIME']


dataset.head()

from sklearn.preprocessing import LabelEncoder


le LabelEncoder()
```

```
dataset['DEST'] le.fit_transform(dataset['DEST'])


dataset['ORIGIN"'] le.fit transform(dataset['ORIGIN'])

dataset [ ORIGIN].unique()

Red

array([0, 1, 4, 3, 2])


dataset = pd.get_dummies (dataset, columns=['ORIGIN, 'DEST']) dataset.head()


x = dataset.iloc[:, 0:8].values


y dataset.iloc[:, 8:9].values

from sklearn.preprocessing import OneHotEncoder


oh = OneHotEncoder()


z=oh.fit_transform(x[:,4:5]).toarray()


t-oh.fit_transform(x[:,5:6]).toarray()


#x=np.delete(x, [4,7], axis=1)

x=np.delete(x, [4,5],axis=1)

from sklearn.tree import DecisionTreeClassifier classifier DecisionTreeClassifier (random_state = 0)
classifier.fit(x_train,y_train
```

DecisionTreeClassifier (random state-0)

decisiontree classifier.predict(x_test)

4]

decisiontree

5]

array([1., 0., 0., ..., 0., 0., 1.])

from sklearn.metrics import accuracy_score desacc = accuracy_score (y_test, decisiontree)

from sklearn.ensemble import RandomForestClassifier rfe RandomForestClassifier(n_estimators-10, criterion-'entropy

rfc.fit(x_train,y_train)

<ipython-input-125-b87bb2ba9825>:1: DataConversionlarning: A column-vector y ravel().

rfc.fit(x_train,y_train)

RandomForestClassifier (criterion="entropy', n_estimators-10)

y predict rfc.predict(x_test)

Impor ng the Keras libraries and packages

```
impo tensorflow

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense

+

# Creating ANN skleton view

classification Sequential()

classification.add(Dense (30, activation='relu'))

classification.add(Dense (128, activation= 'relu'))

classification.add (Dense (64, activation='relu'))

classification.add(Dense (32, activation='relu'))

classification. add (Dense (1, activation= sigmoid'))

classification.compile(optimizer='am', loss='binary_crossentropy, metrics=['accuracy'])

# Training the model
```

```
classification.fit(x_train,y_train, batch_size=4, validation_split-0.2,epochs=100)

y_pred classifier.predict([[129,99,1,0,0,1,0,1,1,1,0,1,1,1,1,1]])


print(y_pred) (y_pred)

y_pred = rfc.predict([[129,99,1,0,0,1,0,1,1,1,0,1,1,1,1,1]])


print(y_pred) (y_pred)

Day 1

dfs


= []


models = [


('RF', RandomForestClassifier()),


('DecisionTree', DecisionTreeClassifier()),


('ANN', MLPClassifier())


results = []


names = []


target_names = ['no delay', 'delay']
```

```
for name, model in models:

scoring = ['accuracy', 'precision_weighted', 'recall_weighted', 'f1_weighted', 'roc_auc'] kfold =
model_selection. KFold (n_splits=5, shuffle=True, random_state=90210) cv_results
model_selection.cross_validate(model, x_train, y_train, cv=kfold, scoring=scoring)

clf = model.fit(x_train, y_train) y_pred = clf.predict(x_test)

print(name)

print(classification_report (y_test, y_pred, target_names=target_names))

results.append(cv_results)

names.append(name)

this_df = pd.DataFrame(cv_results)

this_df['model'] = name

dfs.append(this_df)

final = pd.concat(dfs, ignore_index=True)

return final
```

```
mport numpy as np

import pickle

import matplotlib.pyplot as plt

%matplotlib inline

import seaborn as sns

import sklearn

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.model_selection import RandomizedSearchCV

import imblearn

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import accuracy_score, classification_report,confusion_matrix,f1_score

# Compiling the ANN model


classification.compile(optimizer adam', loss='binary_crossentropy', metrics=['accuracy'])


# Training the model


classification.fit(x_train,y_train, batch_size=4, validation_split-0.2, epochs-100)

Flight_Delay_Prediction_for_aviation_Industry_using_Machine_Lea...


#Importing the Keras libraries and packages
```

```
import tensorflow

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense

# Creating ANN skleton view

classification = Sequential()

classification.add(Dense (30, activation='relu'))

classification.add(Dense (128, activation='relu'))

classification.add(Dense (64, activation='relu'))

classification.add(Dense (32, activation='relu'))

classification.add(Dense (1,activation='sigmoid'))

print('Training accuracy: print('Testing accuracy: ,accuracy_score (y_train, y_predict_train))
accuracy_score (y_test,y_predict))

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_predict)

cm
```

```
from sklearn.metrics import accuracy_score

from sklearn.metrics import confusion_matrix


cm = confusion_matrix(y_test, y_predict)


cm

from sklearn.metrics import accuracy_score

from sklearn.metrics import confusion_matrix cm = confusion_matrix(y_test, decisiontree)

desacc = accuracy_score (y_test, decisiontree)

score = accuracy_score (y_pred,y_test)


from sklearn.metrics import accuracy_score, classification_report print('The accuracy for ANN model is: {}%'.format(score*100))

from sklearn.metrics import confusion_matrix cm = confusion_matrix(y_test, y_pred)


cm

from sklearn.metrics import accuracy_score, classification_report

from sklearn.metrics import confusion_matrix cm = confusion_matrix(y_test, y_pred)


cm

= RandomizedSearchCV(estimator=rf, param_distributions-parameters, cv=10,n_iter=4)

parameters = {


'n_estimators': [1,20,30,55,68,74,98,120,115],


'criterion': ['gini', 'entropy'], 'max_features: ["auto", "sqrt", "log2"],
```

'max_depth : [2,5,8,10], 'verbose' [1,2,3,4,6,8,9,10]

bt_params = RCV.best_params_

bt_score = RCV.best_score_

model = RandomForestClassifier (verbose= 10, n_estimators= 120, max_features= 'log2', max_depth= 10, criterion= 'entropy')

RFC=accuracy_score (y_test,y_predict_rf) RFC

y_predict_rf RCV.predict(x_test)

import pickle

pickle.dump (RCV, open('flight.pkl', 'wb'))

from flask import Flask, request, render_template

import numpy as np

import pandas as pd

import pickle

import os

model = pickle.load (open ('flight.pkl', 'rb'))

app Flask (

name

```
) #initializing the app

@app.route('/')

def home ():

return render_template("index.html")

@app.route('/prediction', methods = ['POST'])

def predict():

name = request.form['name']

month = request.form['month']

dayofmonth = request.form['dayofmonth']

dayofweek= request.form['dayofweek']

origin request.form['origin']

if (origin == "map"):

origini, origin2, origin3, origin4, orgin5 - 0,0,0,0,1
```

```
if (origin == "dtw"):

origini, origin2, origin3, origin4, orgin51,0,0,0,0

originl,origin2,

if (origin == "jfk"):

origini, origin2, origin3, origin4, orgin5=0,0,1,0,0

if (origin == "sea"):

origini, origin2, origin3, origin4, orgin5 0,1,0,0,0

if (origin == "alt"):

origini, origin2, origin3, origin4, orgin5=0,0,0,1,0

destination = request.form['destination']

if (destination "msp"):

N destinationl, destination2, destination3, destination4, destination5 0,0,0,0,1

if (destination "dtw"):
```

```
destinationl, destination2, destination3, destination4, destination5 = 1,0,0,0,0

if (destination "jfk"):

destinationl, destination2, destination3, destination4, destination5 = 0,0,1,0,0

if (destination == "sea"):

destinationl, destination2, destination3, destination4, destination5 = 0,1,0,0,0

if (destination == "alt"):

destinationl, destination2, destination3, destination4, destination5 = 0,0,0,1,0

dept = request.form['dept']

arrtime = request.form['arrtime']

actdept = request.form['actdept']

dept15-int (dept)-int (actdept)

total = [[name, month, dayofmonth, dayofweek, origin…

if
```

name

main = True)