IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wassim IZERGUINE
16/01/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

In this project, we will use machine learning classifier to determine if the first stage of Falcon 9 will land successfully. To do, we will split your data into training data and test data to find the best Hyperparameter for SVM, Classification Trees decision, Logistic Regression and K-NN to find the method that performs best using test data.

The best classifier model found is the Trees decision.

Introduction

Project background and context :

The goal of this project is to predict if the Falcon 9 first stage will land successfully, to answer this question, we will go through different steps : data collection, data wrangling, building a dashboard, machine learning algorithms and the final step is to detect which algorithms delivers the best performance.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Using API and web scrapping

- Perform data wrangling

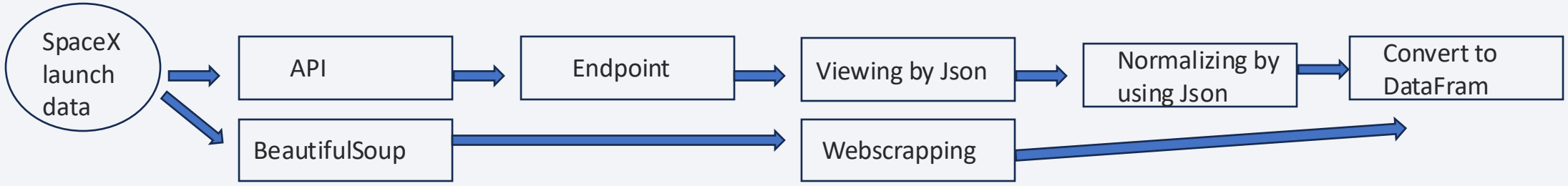
  Understand datas and create a landing outcome label

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

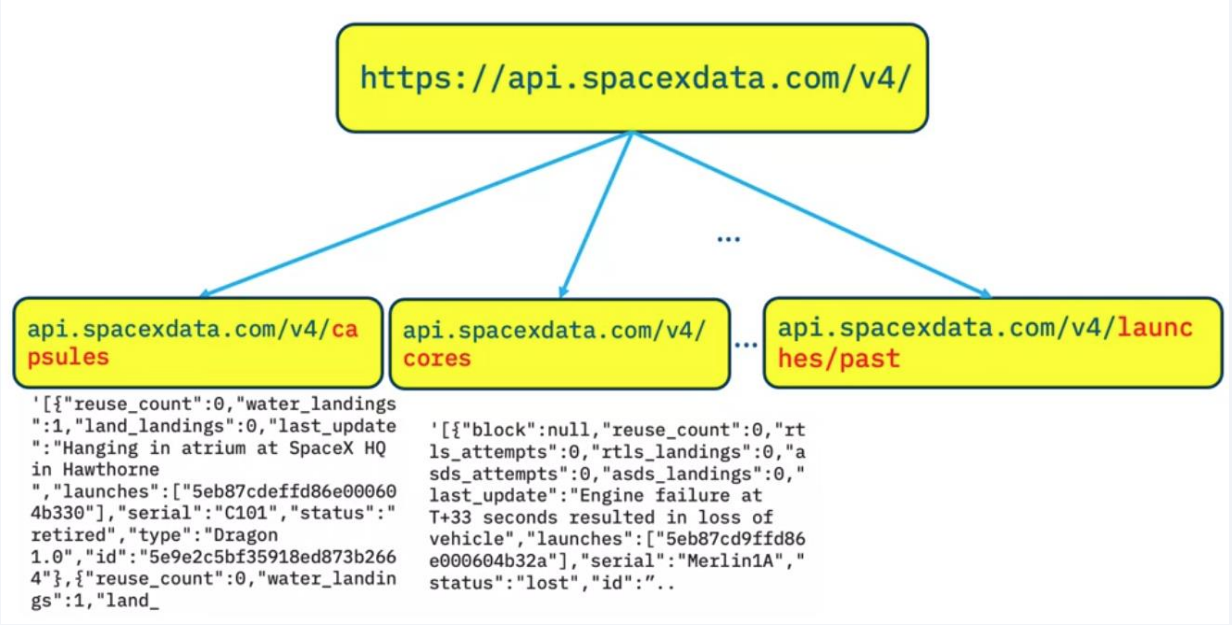
SpaceX launch data can be retrieved from an API using a GET request with the requests library in Python. The data obtained from the API is typically in JSON format. Once retrieved, the JSON file should be normalized to facilitate easier data handling. One common approach is to convert this normalized data into a pandas DataFrame.

Finally, when dealing with the DataFrame, it's important to handle null values appropriately. One method is to replace these null values with the mean value of the respective column.



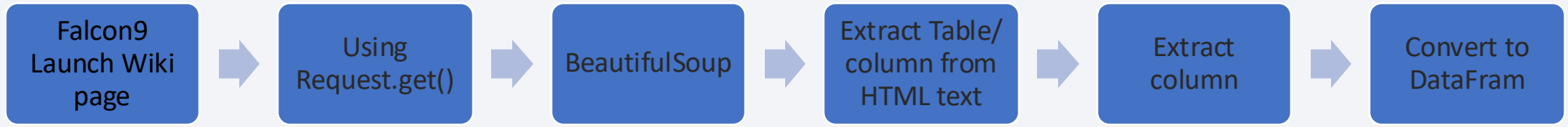
Data Collection – SpaceX API

- The URL of the SpaceX API calls notebook (https://github.com/LearnerDSS/Project_Capstone-/blob/main/Module1%20_n1_how%20to%20collect%20data%20API%20an%20organized%20in%20a%20table_(1).ipynb).



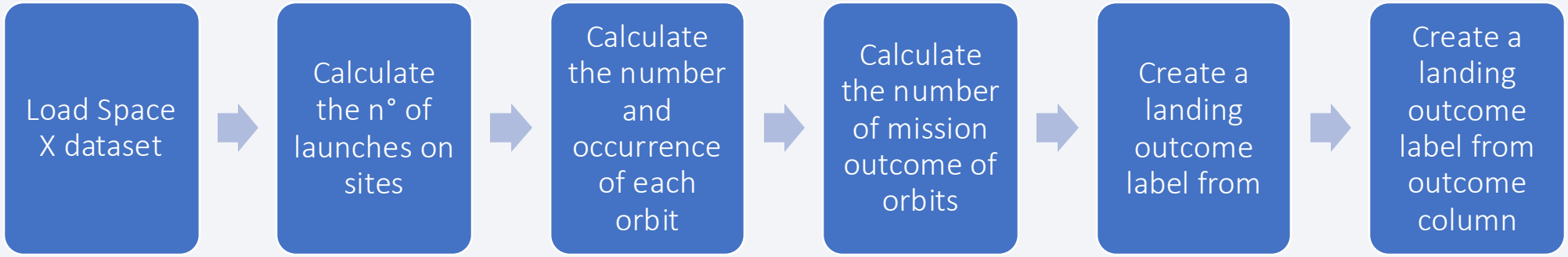
Data Collection - Scraping

- the GitHub URL of the web scraping notebook, (https://labs.cognitiveclass.ai/v2/tools/jupyterlab?ulid=ulid-d8f550da53a199b98550928754ae733955db18ec).



Data Wrangling

- The objectif of data wrangling is more than preparing and cleaning the data, it involves also tasks like handling missing data, transforming variables, and ensuring consistency.



GitHub URL of data wrangling notebooks :
https://github.com/LearnerDSS/Project_Capstone-/blob/main/Module1%20_n3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To find the relationship between variables such as : Payload Mass and Launch Site or Success rate and Orbit type we had to use charts. The charts that we have used in this part of the project are : Scatter plot, histogram and line chart.

- the GitHub URL of EDA with data visualization notebook : https://github.com/LearnerDSS/Project_Capstone-/blob/main/Module2%20_n1_edadataviz.ipynb

EDA with SQL

The SQL queries used are :

- Connecting SQLite database named my_data.db using %sql : %sql sqlite:///my_data1.db
- Convert DataFrame to SQL table : df.to_sql
- Remove Blank row from table : %sql create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
- Select the name of unique launch sites : %sql SELECT Unique(LAUNCH_SITE) from SPACEXTBL ;
- Select boosters which have have payload mass greater than 4000 and less than 6000 : %sql SELECT Booster_Version FROM SPACEXTBL where PAYLOAD_MASS__KG_> 4000 AND PAYLOAD_MASS__KG_<6000 ;

Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose:
https://github.com/LearnerDSS/Project_Capstone-/blob/main/Module2%20_n2_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Adding circle, line, etc, to folium map help analyzing launch site locations and find factors that influence the success rate.
- the GitHub URL of the interactive map with Folium is : https://github.com/LearnerDSS/Project_Capstone-/blob/main/Module3%20_n1_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

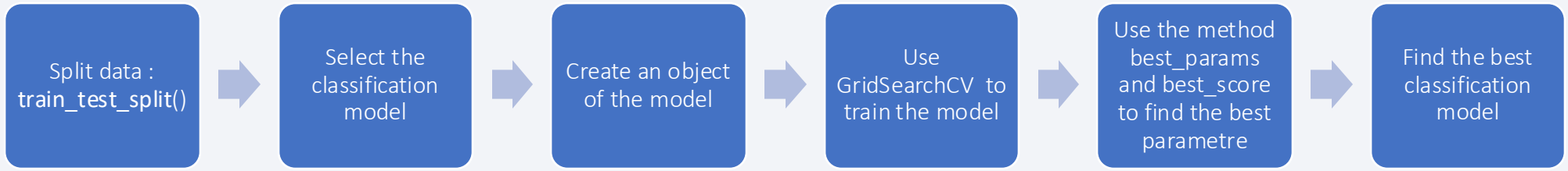
We have used in the Dashboard the Pie chart and Scatter point chart.

- Pie chart : to display the success for each launch site
- Scatter point chart : to display the correlation between payload mass, class and booster version class categorie

The GitHub URL of Plotly Dash lab : https://github.com/LearnerDSS/Project_Capstone-/blob/main/Module3%20_n2_spacex_dash_app.py

Predictive Analysis (Classification)

The method involves splitting data into training and testing sets, selecting a model, find its best parameters, and calculate its accuracy. Once all the classification model were calculated, the best one is chosen based on the highest accuracy achieved.



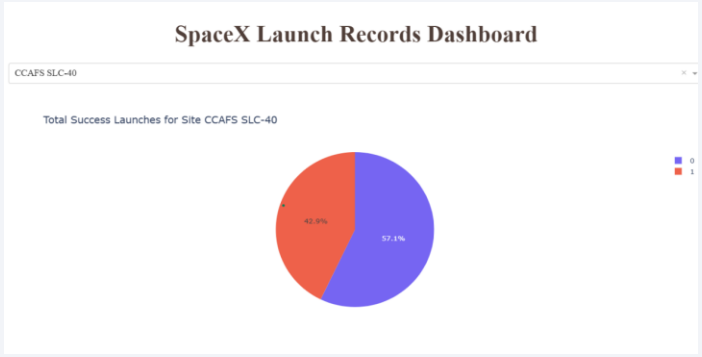
The GitHub URL of predictive analysis lab : https://github.com/LearnerDSS/Project_Capstone-/blob/main/Module4%20_n1_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

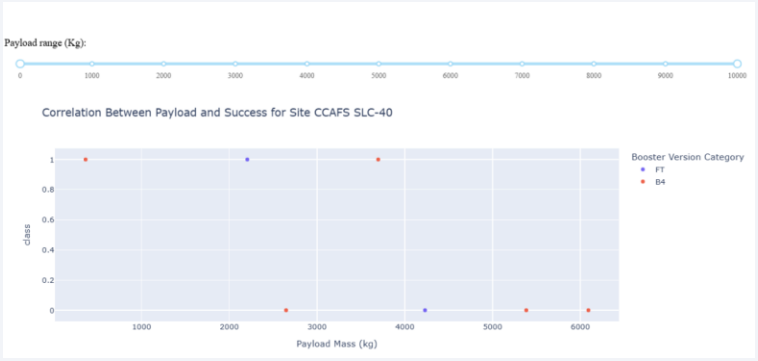
Results

Exploratory data analysis results

- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10 000).
- In the LEO orbit, success seems to be related to the number of flight. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.
- we can observe that the success rate since 2013 kept increasing till 2020.

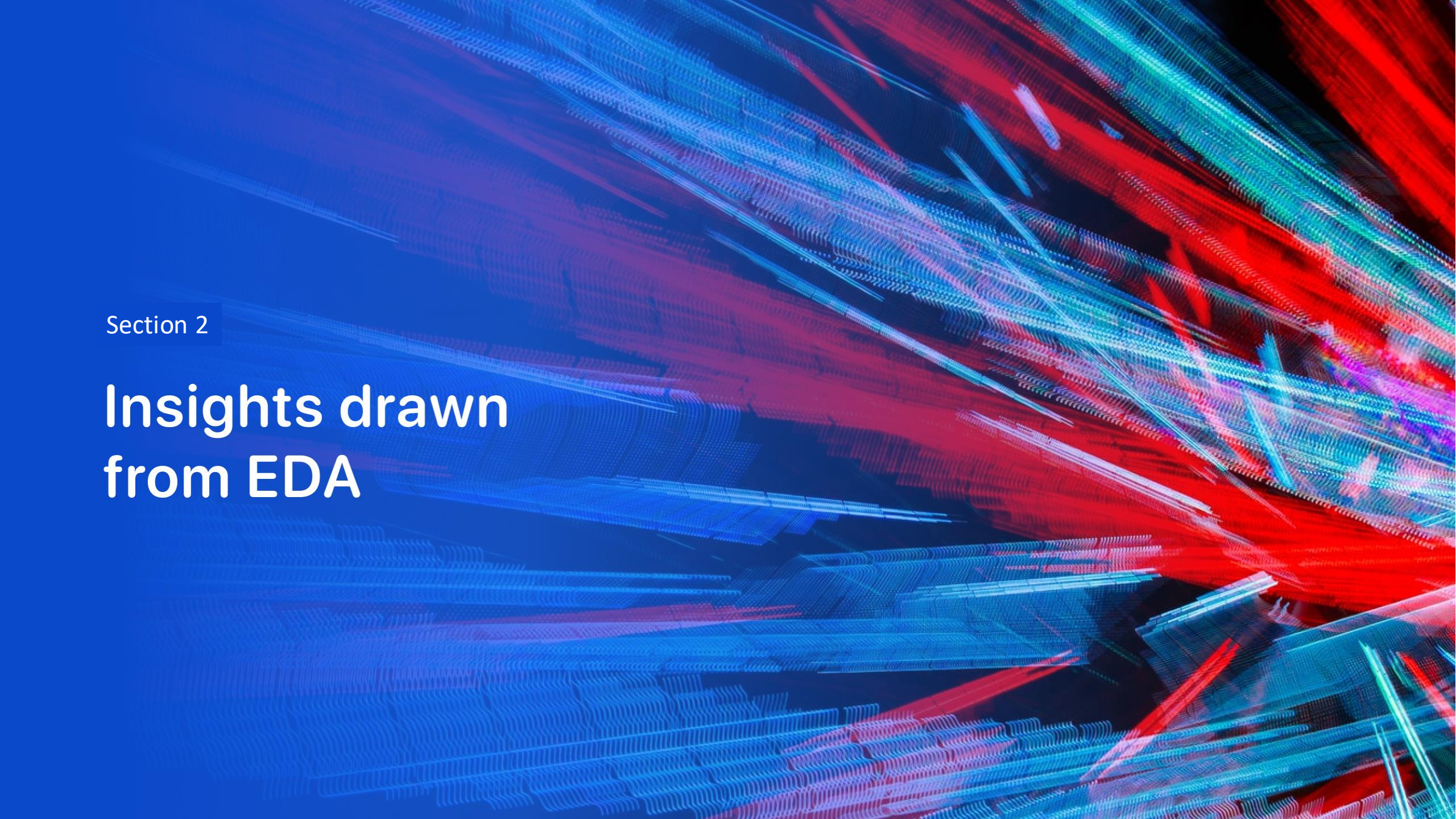
Interactive analytics demo in screenshots





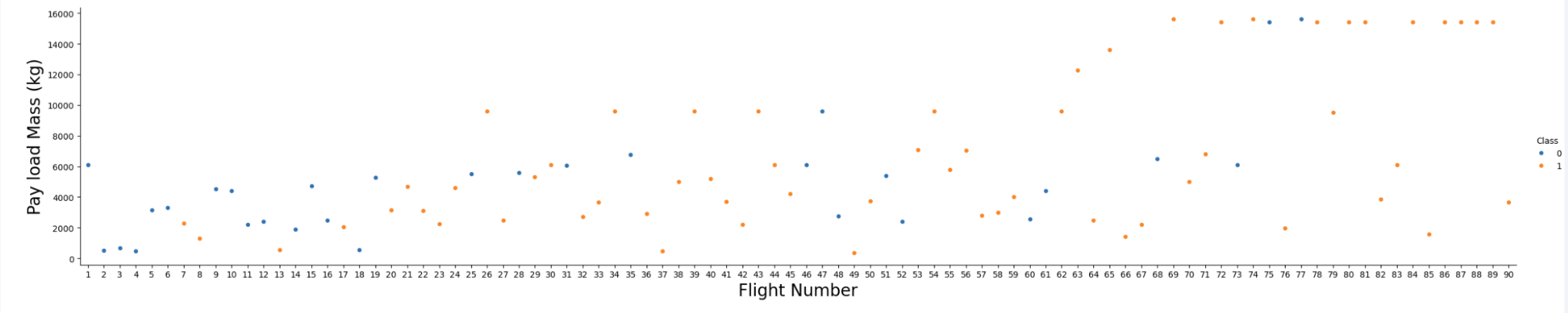
Predictive analysis results

Based on the accuracy, the decision tree is the best model.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

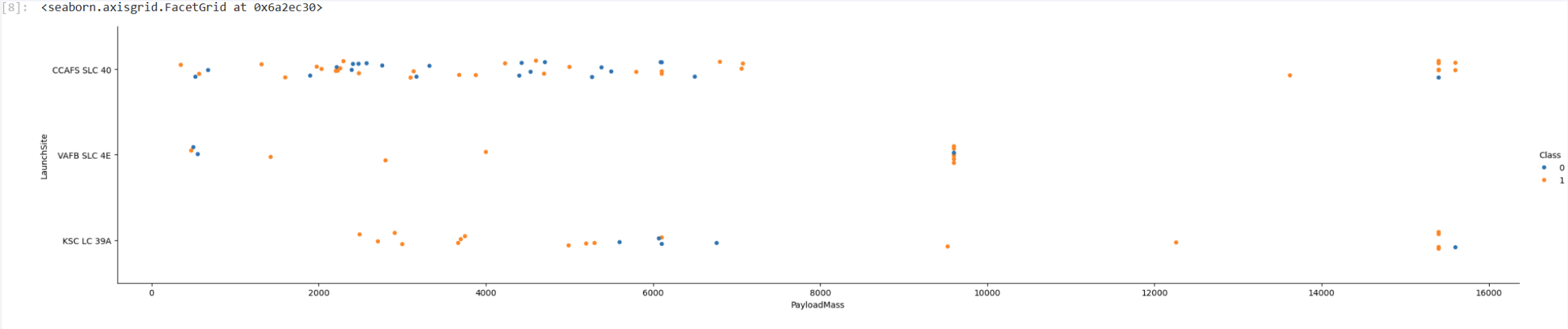
Scatter plot of Flight Number vs. Launch Site :



The lunch of payload mass of 16 000 kg started at the flight n° 69, the majority of this lunches was successful.

Payload vs. Launch Site

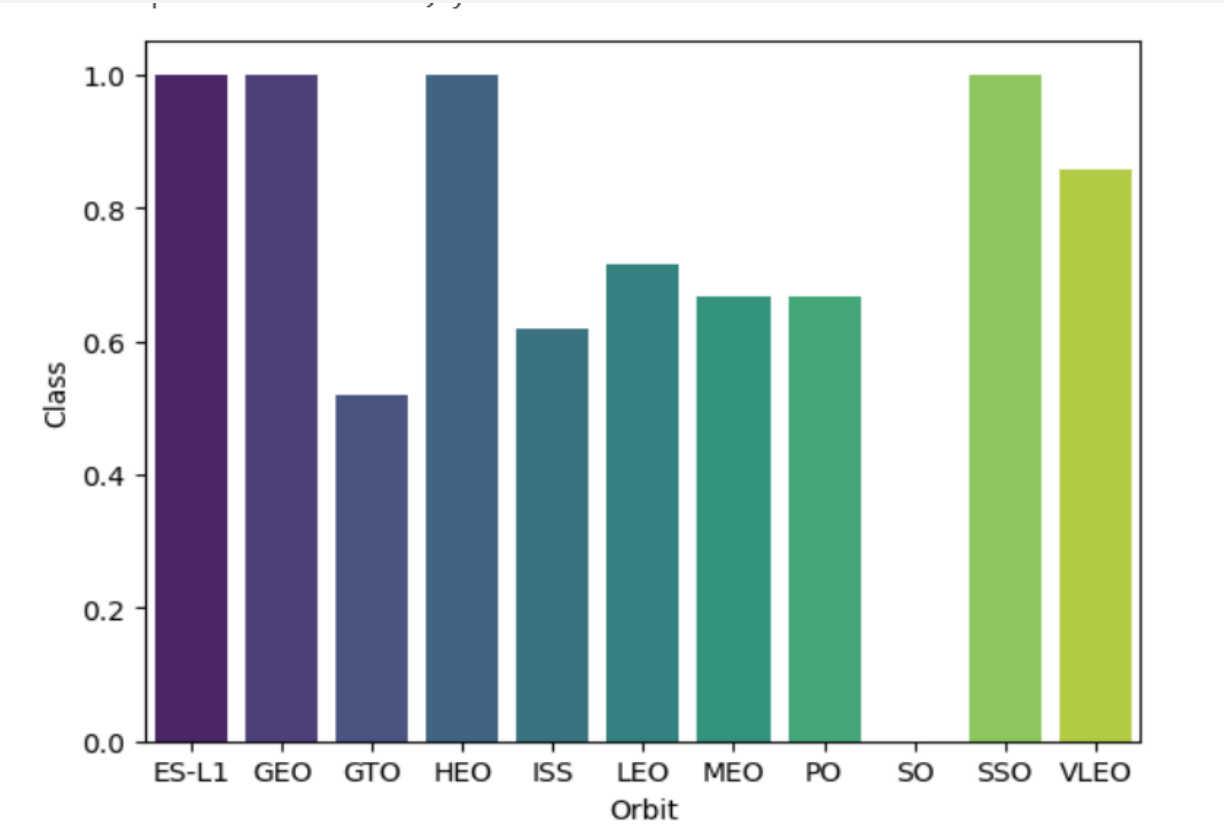
- A scatter plot of Payload vs. Launch Site :



- We observe that the majority of launches was in CCAFS SLC 40.
- The KSC LC 39A presents a lot of success launches compared to the other site.

Success Rate vs. Orbit Type

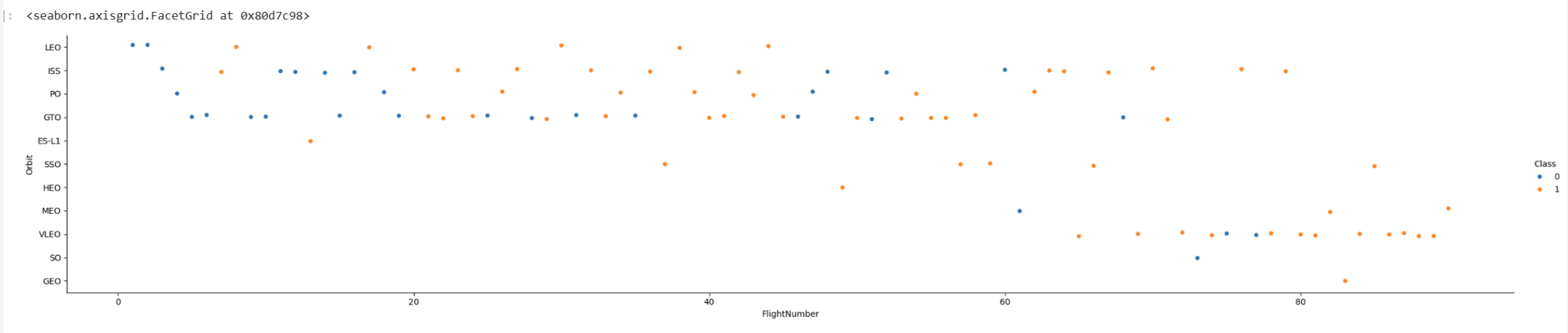
Bar chart for the success rate of each orbit type :



The Orbits that have the most success rate are : ES-L1 GEO SSO.

Flight Number vs. Orbit Type

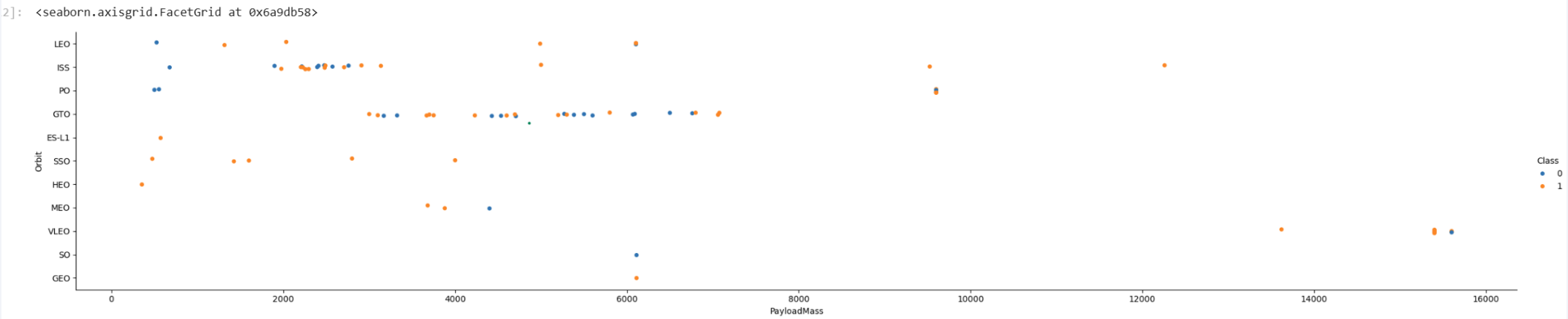
Scatter point of Flight number vs. Orbit type :



- we observe that after the flight number n° 50, the orbit MEO,VLEO,SO and GEO start to be used.

Payload vs. Orbit Type

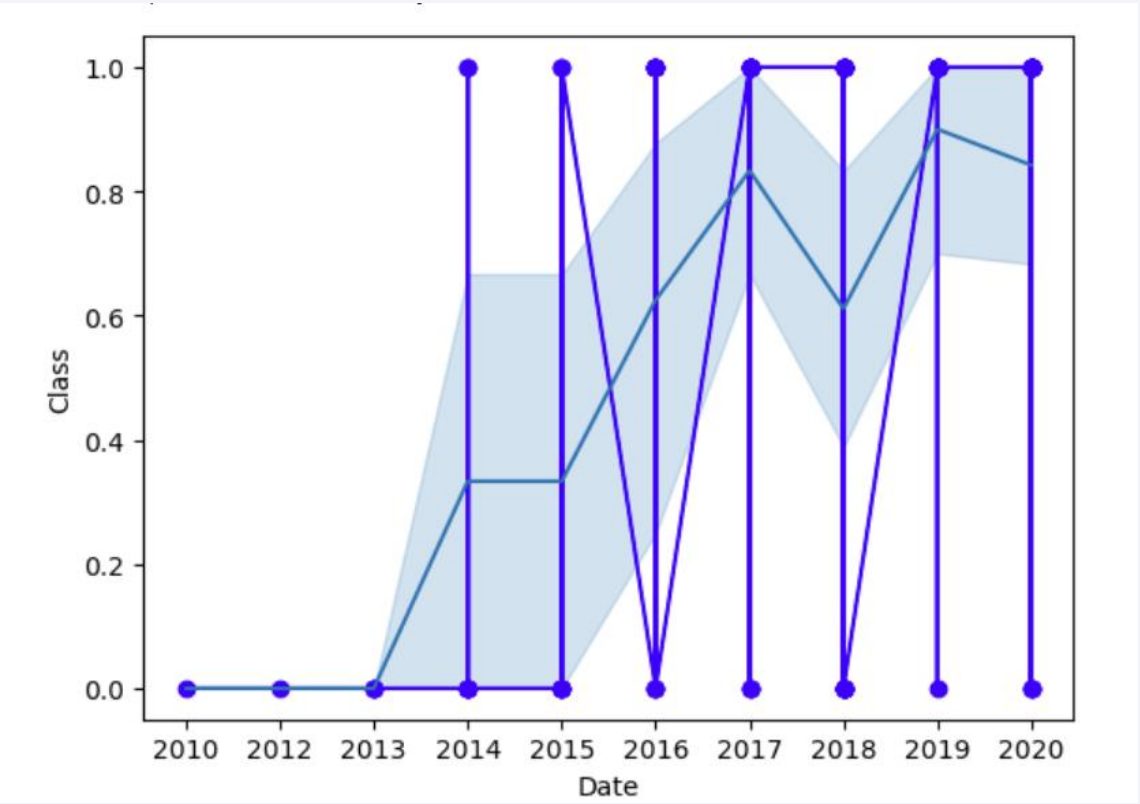
Scatter point of payload vs. orbit type :



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- The orbit SSO presents a 100 % successful landing rate.

Launch Success Yearly Trend

Line chart of yearly average success rate :



We observed that success rate began in 2013 and continued until 2020.

All Launch Site Names

The unique launch sites are :

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

The 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

All the landing outcomes have failed for this launch sites

Total Payload Mass

The total payload carried by boosters from NASA is 45 596 Kg

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is :

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad is :

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are :

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes is :

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

We observe that the number of failed missions is 2 which is so small compared to successful missions.

Boosters Carried Maximum Payload

List of the names of the booster which have carried the maximum payload mass :

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Ranking of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order :

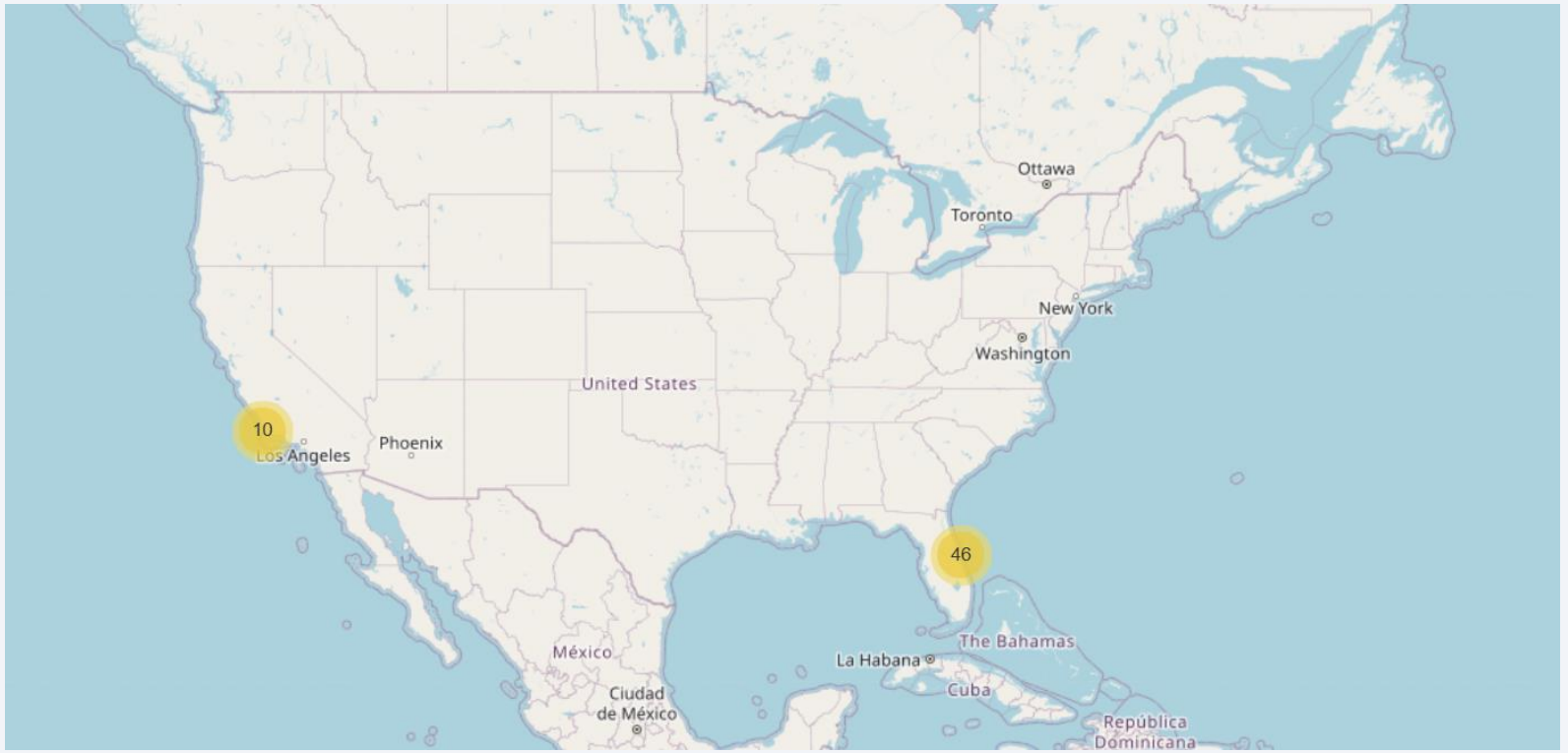
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

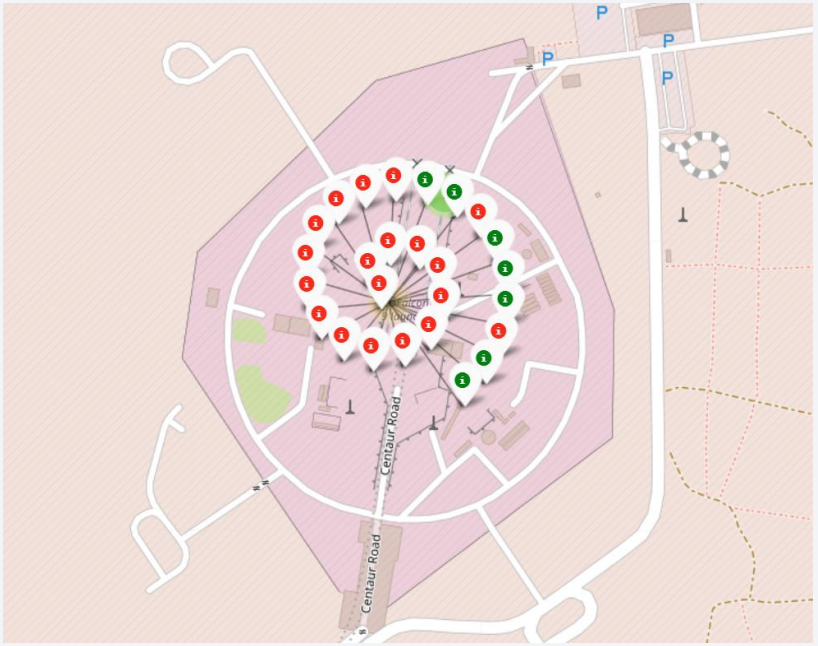
Launch sites on folium map

The generated folium map contained all the launch sites



Color labeled launch outcomes on the folium map

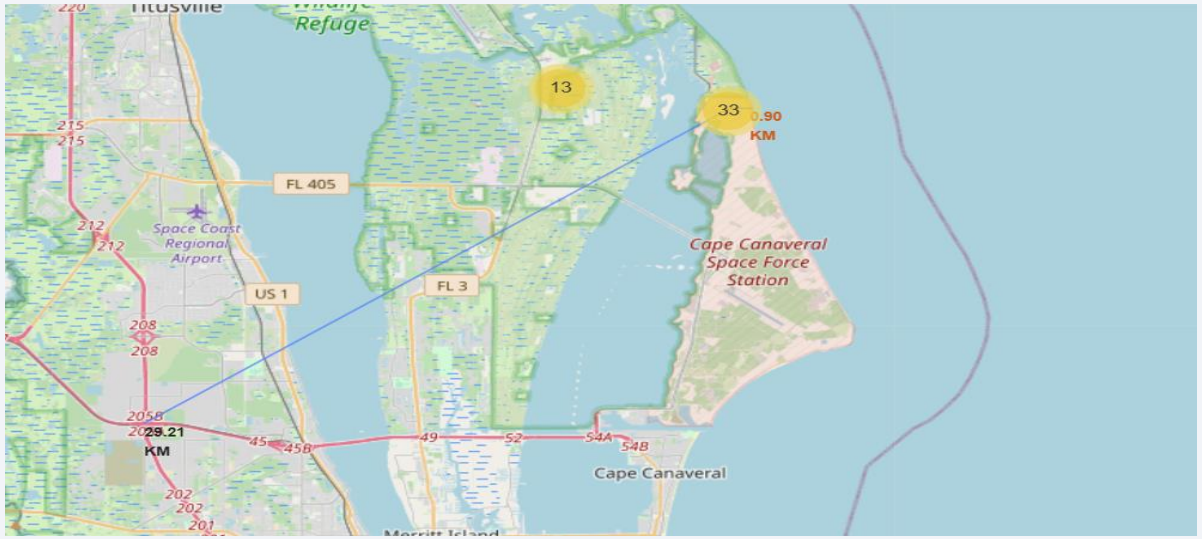
Screenshot shows the color-labeled launch outcomes on the map

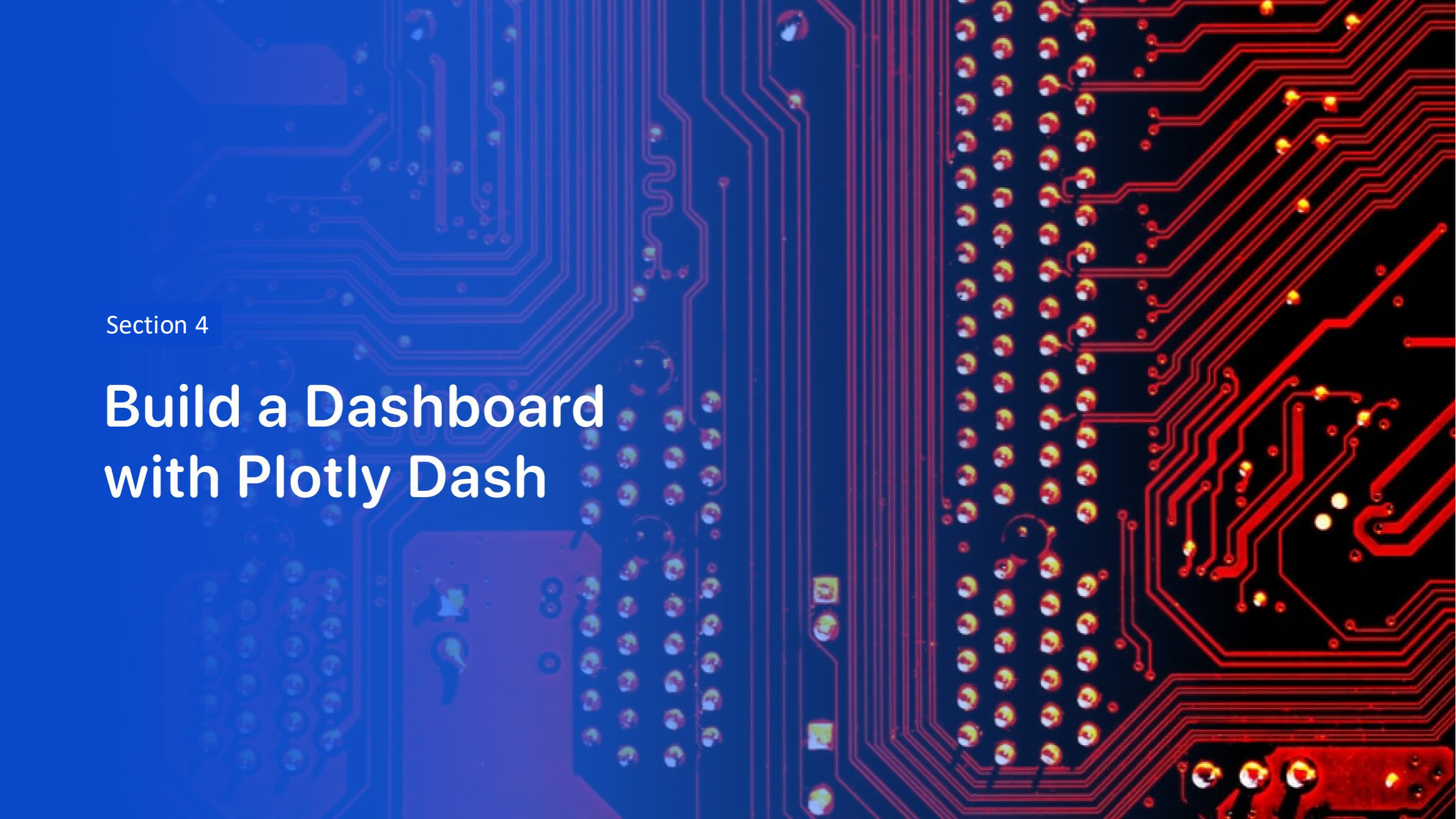


Is it very practical to have a color-labeled launch outcomes because it helps us to distinguish easily a successful or failed result

launch site and its proximities on folium map

A screenshot of a selected launch site and its proximities (railway, highway, coastline) with calculated distance.



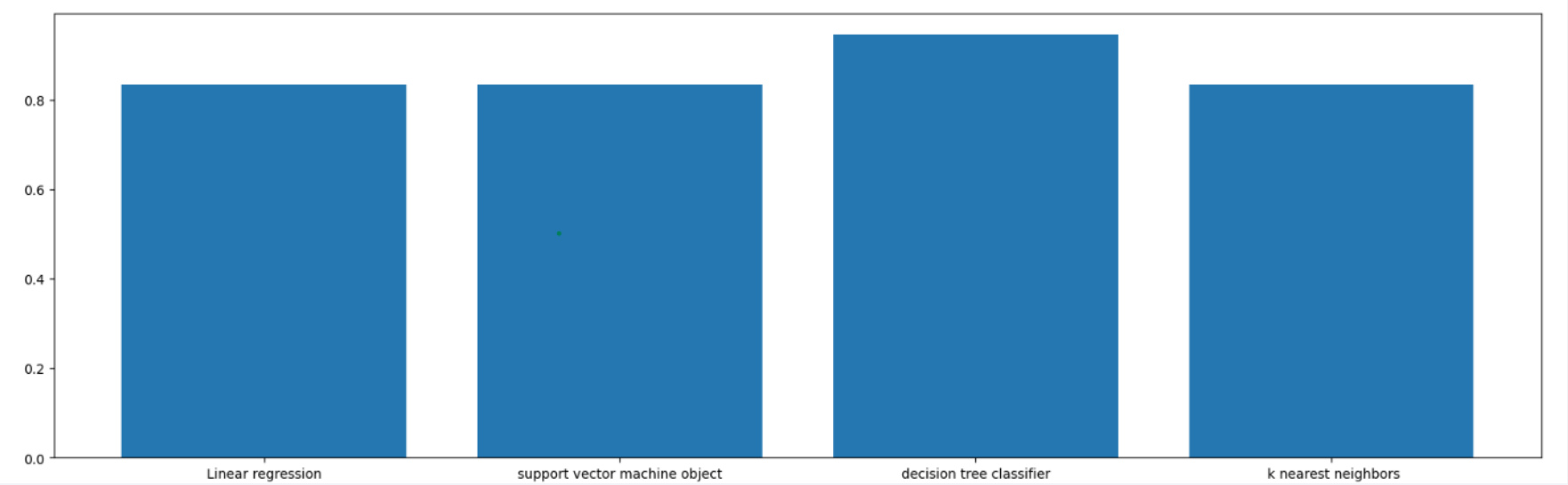

Section 4
Build a Dashboard with Plotly Dash

Section 5

Predictive Analysis (Classification)

Classification Accuracy

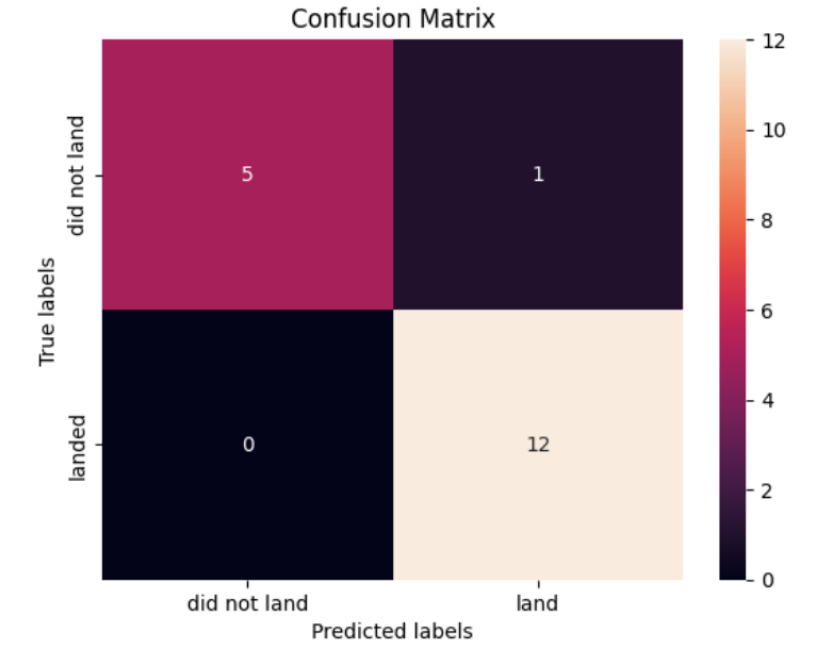
A bar chart of accuracy of classification models :



Based on the bar chart, we observe that the best module is the decision tree classifier

Confusion Matrix

The confusion matrix of the Tree classifier model



We can observe that this classifier model is the one that has the highest matching number between Y_test and Y_hat.

Conclusions

The principal steps for predictive analysis are :

- Create column for the class
- Standardize data
- Split into training and test data
- Training this models : SVM, Classification Tree decision, K-NN and logistic regression
- Based on accuracy we choose the best model
- The confusion matrix is a good way to determine the performance of the model

Thank you!