IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Félix Fernández
2024-01-23

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- The methodologies were used to analyze data:
  - Data collection using web scrapping and SpaceX API.
  - EDA (Exploratory Data Analysis) with Python and SQL.
  - Machine Learning Models for predictions.
- Summary of all results
  - We collected valuable Data from Public Resources.
  - With EDA we identify valuable information from the Data and which features uses to use for the prediction models.
  - Identify the best model and features for the best use to the Machine Model Prediction.

Introduction

- With this project for the Space Y enterprise the idea is could be a serious competitor to Space X in the industry.

- The problem that we solved with this project is what strategies we can use to optimize the cost from launches, predicting the success pf the launch and where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The Data was collected from two sources:
    - Using the Space X API (https://api.spacexdata.com/v4/launches/past)
    - With Web Scrapping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Here we enrich the Data collected by creating a landing outcome label after analyzing the Data.

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL
  - We were able to answer various questions and find valuable aspects for our objectives.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Using these tools we were able to find the places and visualize where the successful and unsuccessful launches were carried out.
- Perform predictive analysis using classification models
  - We used a very well structured strategy that concluded with obtaining the best parameters for each model and knowing which model has the greatest precision.

Data Collection

- Data sets was collected using Space X API (https://api.spacexdata.com/v4/launches/past.) and using Web Scrapping techniques and extracting the data from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

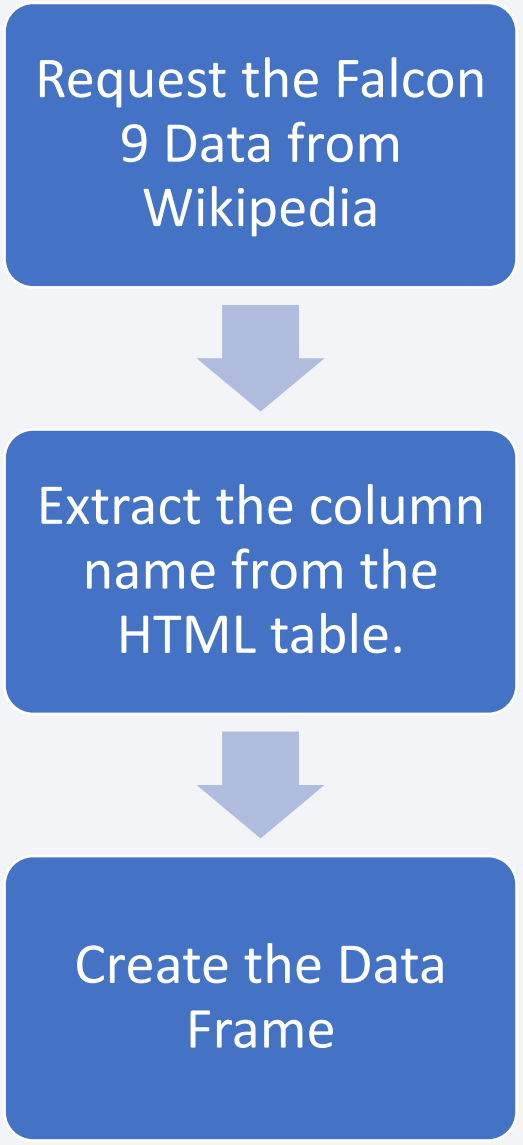
- SpaceX offer a public API that we where able to extract the Data and use it.

- Source code: https://github.com/frfv04/DScourseprojects/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Data Collection - Scraping

- The Data used from SpaceX falcon 9 launches is avaible in Wikipedia.

- Source Code: https://github.com/frfv04/DScourseprojects/blob/main/jupyter-labs-webscraping.ipynb.



Data Wrangling

- Start seeing the percentage of Null values.
- Realize some calculus to answer some questions about the Data.
- Convert the ‘Class’ data from categorical to numerical.

- Code source: https://github.com/frfv04/DScourseprojects/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb.

EDA with Data Visualization

- We use scatter and bar plots to visualize the relationship between the features.
- Code source: https://github.com/frfv04/DScourseprojects/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- We performed the following SQL queries:
  - Names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date when the first succesful landing outcome in ground pad was acheived
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - The total number of successful and failure mission outcomes
  - the names of the booster_versions which have carried the maximum payload mass.
  - The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - The count 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.
- Code source: https://github.com/frfv04/DScourseprojects/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers are useful to be able to point out the places where the launches were made.
- Circles serve to highlight an area using specific coordinates.
- Marker clusters are useful for indicating sets of events in an area
- Lines are useful to indicate the distance between two coordinates

- Code source: https://github.com/frfv04/DScourseprojects/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb.

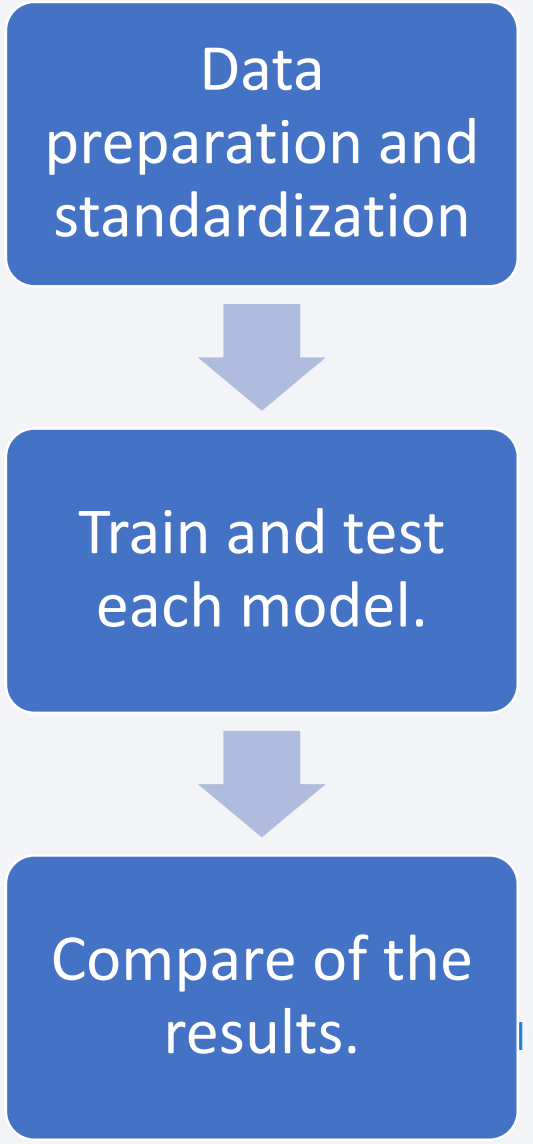
Build a Dashboard with Plotly Dash

- The dashboard had the following graphs to perform the analysis:
  - Porcentaje de lanzamientos por sitio.
  - Payload range.
- The dashboard was very useful for quick analysis of specific characteristics.

Predictive Analysis (Classification)

- The model used were:
  - Logistic Regression.
  - Support Vector Machine.
  - Decision Tree.
  - K nearest neighbors.

- Code source: https://github.com/frfv04/DScourseprojects/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

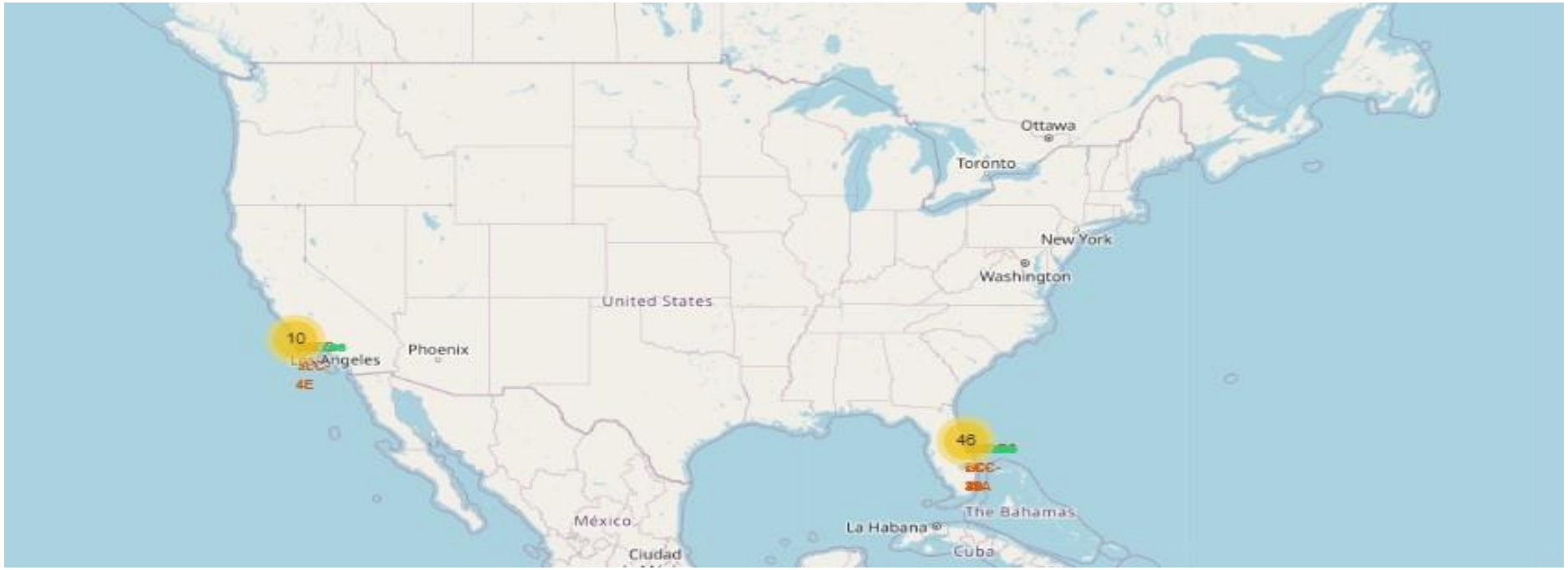


Results

- Exploratory data analysis results
  - SpaceX have 4 unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS) is 48213Kg.
  - The average payload mass carried by booster version F9 v1.1 is 2928.4 Kg
  - The date when the first succesful landing outcome in ground pad was achieved was the 2015-12-22.
  - The total number of successful and failure mission outcomes is 100.

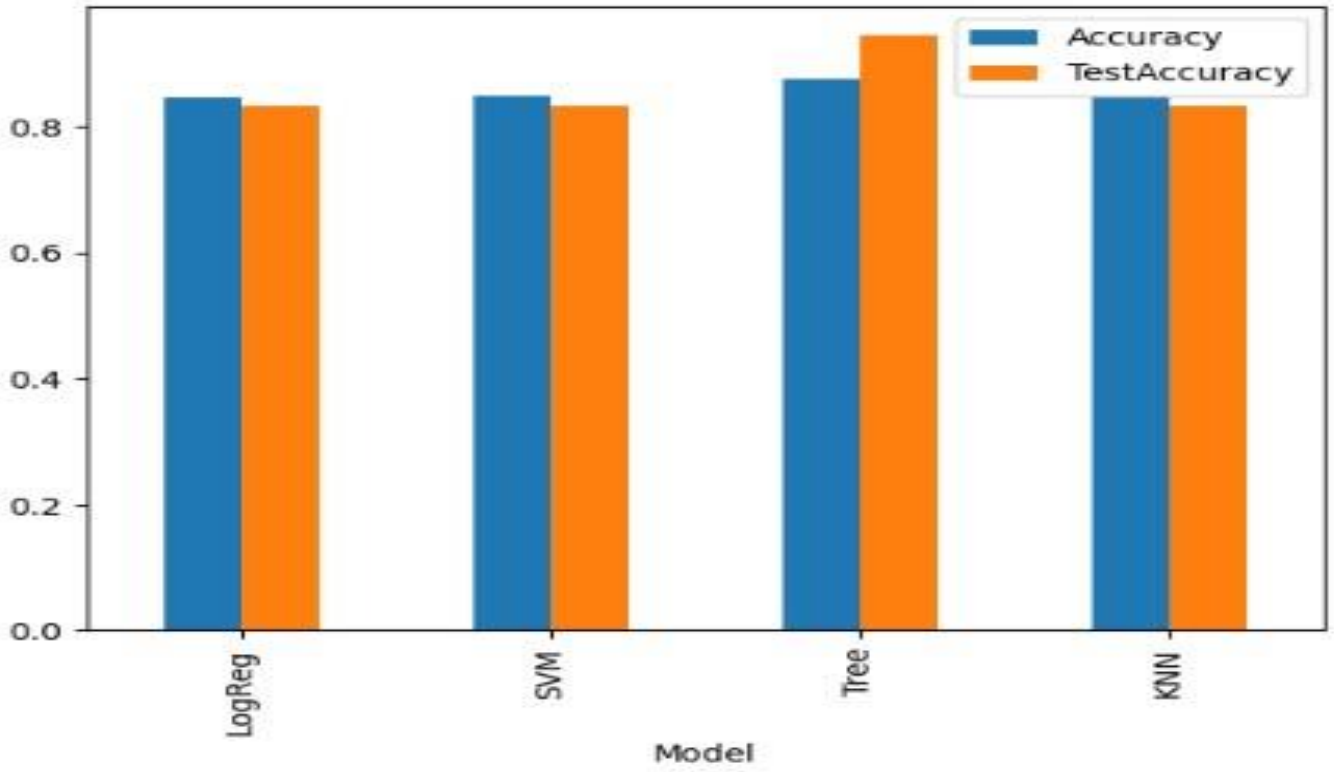
Results

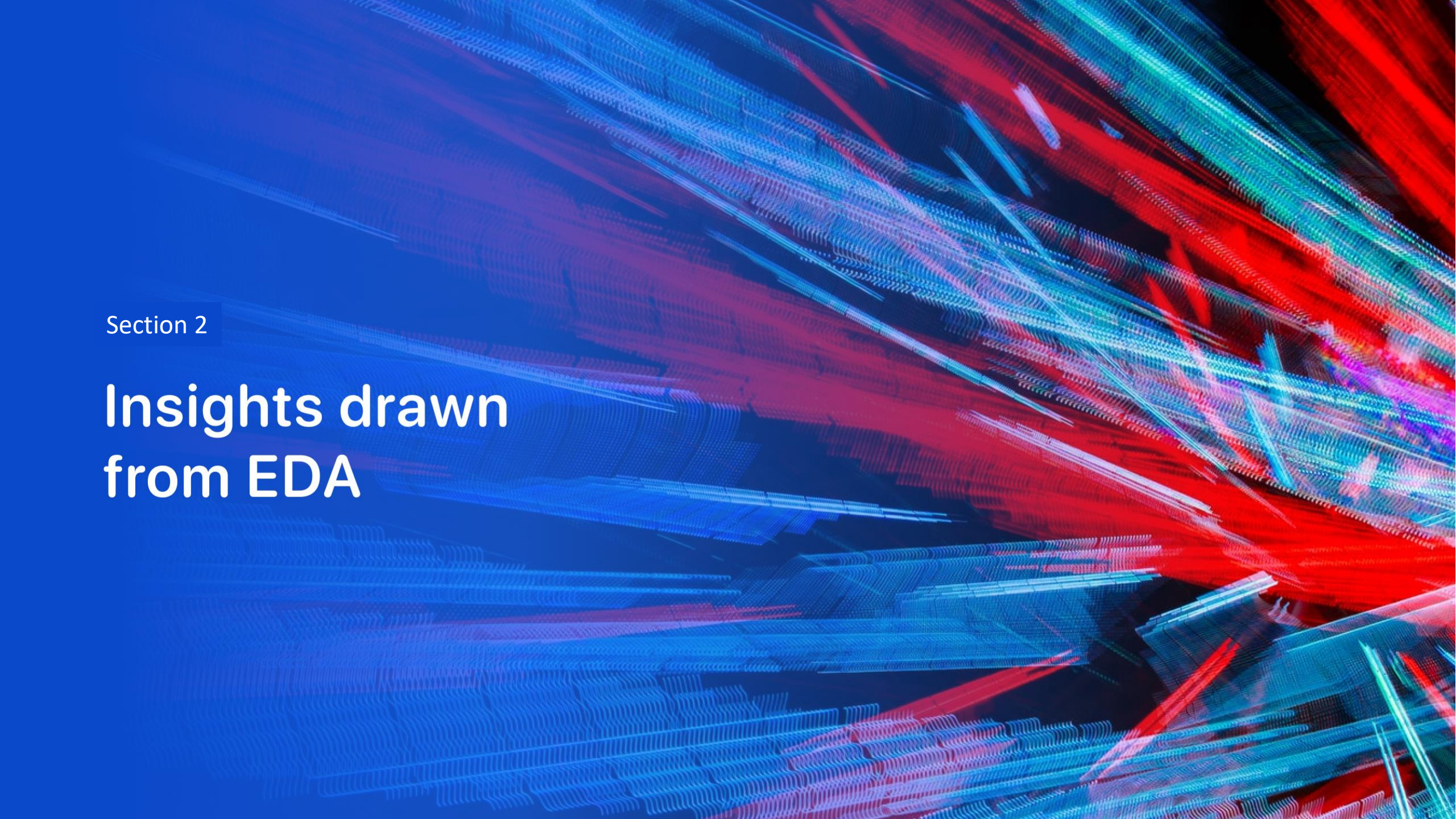
- With interactive analytics was possible to see that the launches were made near to the sea:



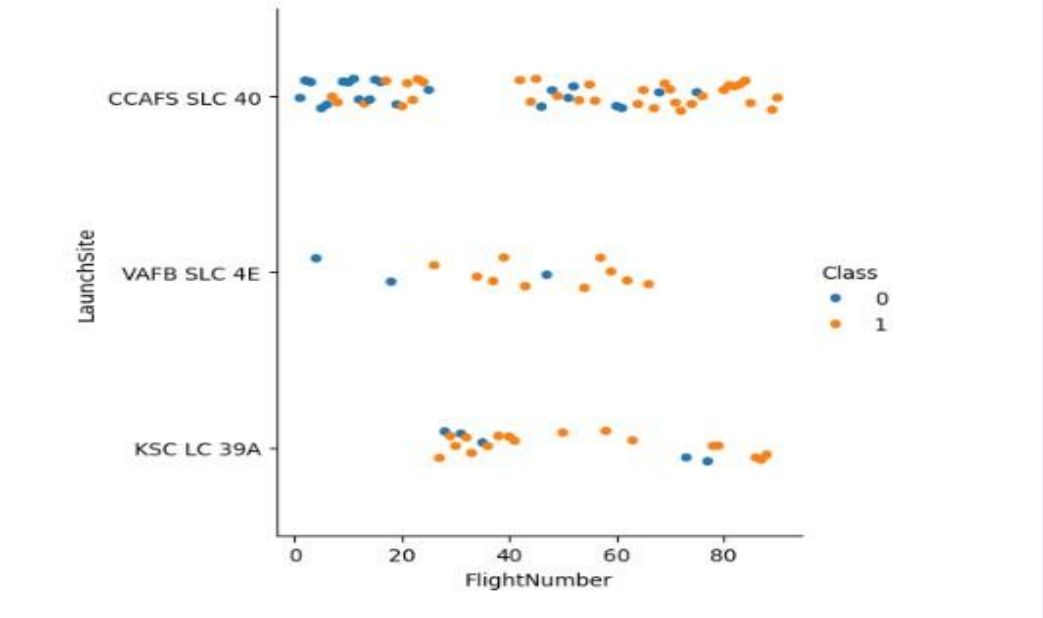
Results

- The predictive analysis results showed that the best model which we can use is the decision tree.




Section 2
Insights drawn from EDA

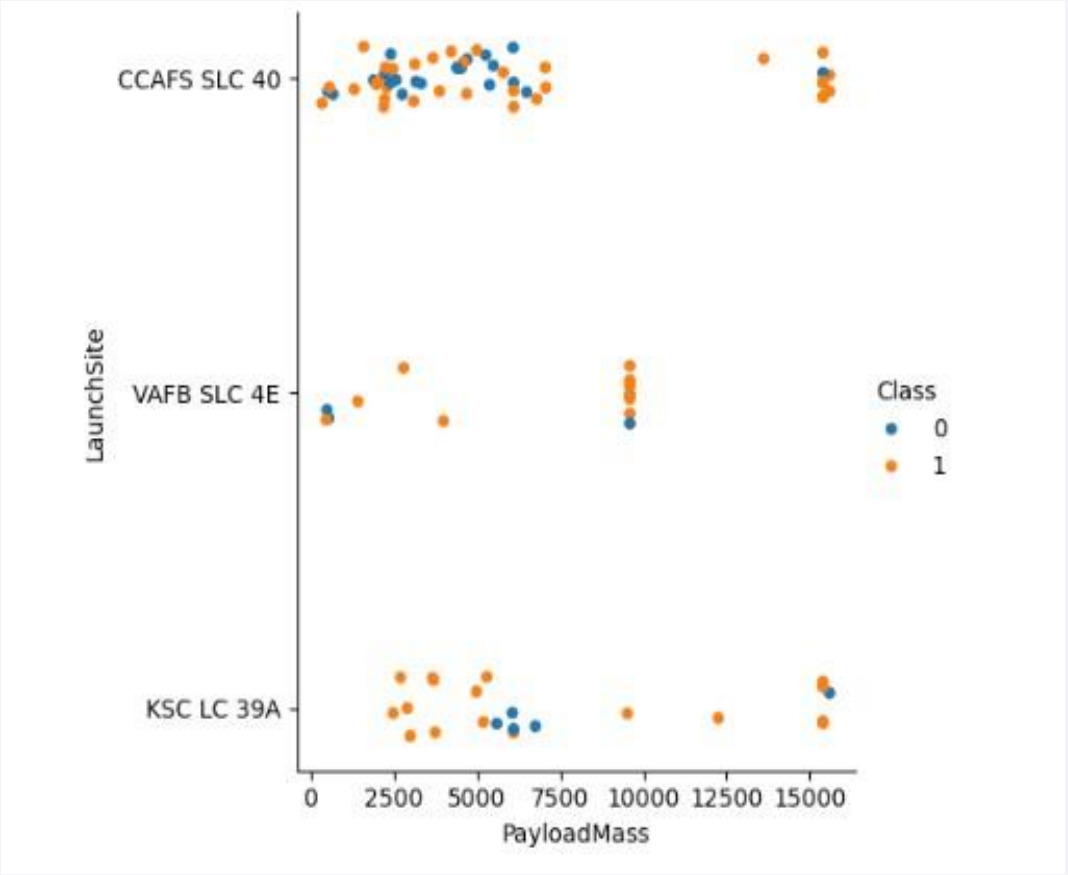
Flight Number vs. Launch Site



- The majority of launches were made on VAFB SLC 4E was successful.
- The majority of the launches were made on CCAFS SLC 40.
- The successful rate increase over the time.

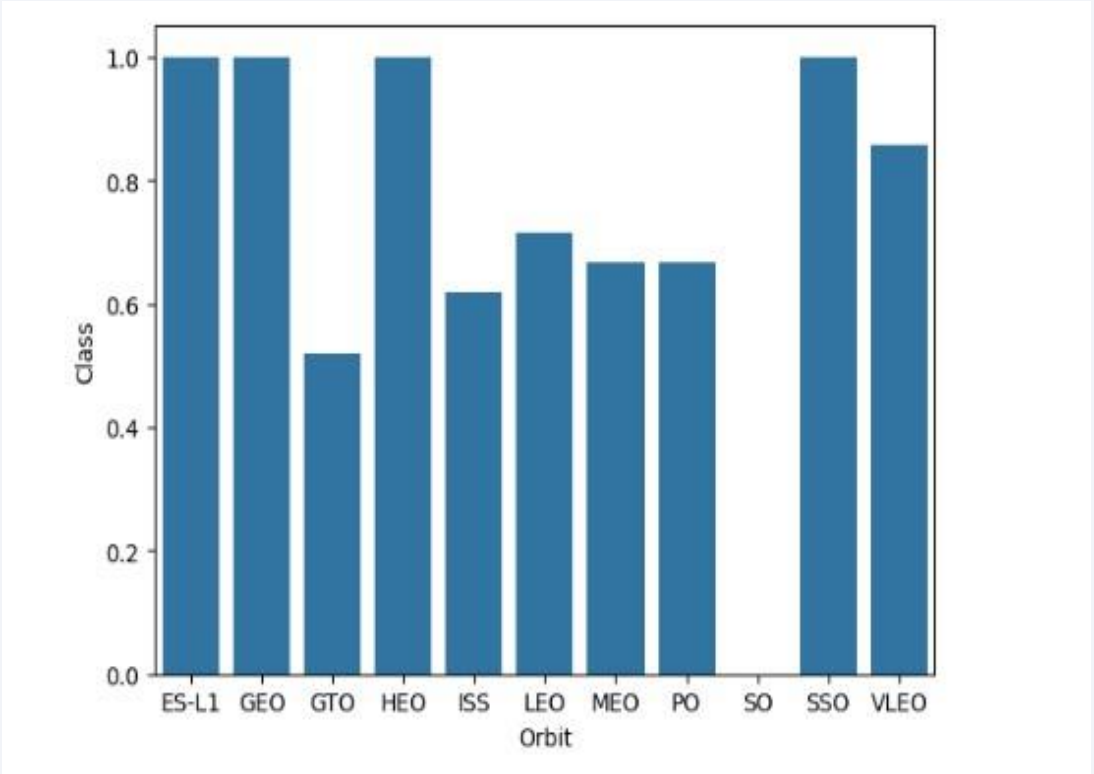
Payload vs. Launch Site

- Most launches over 12000Kg were a success and only possible in CCAFS SLC 40 and KSC LC 39A
- The 9000 KG launchers at VAFB SLC 4E were mostly successful



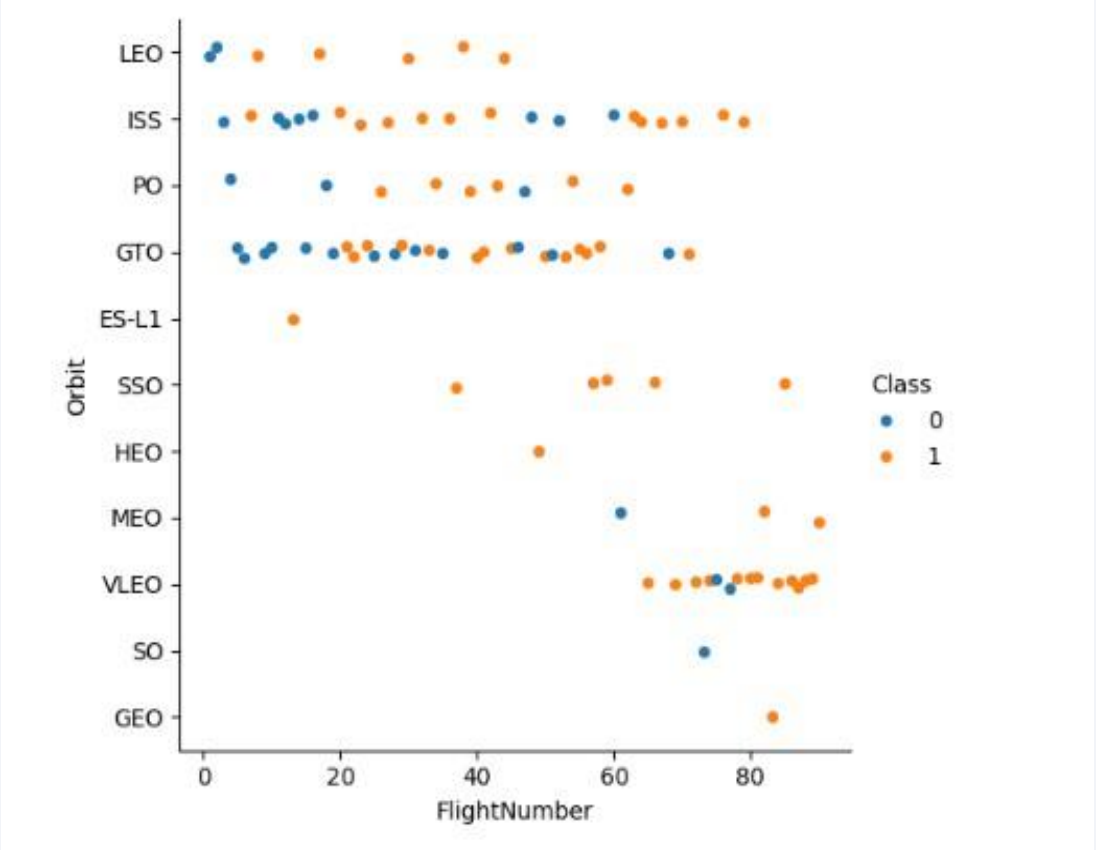
Success Rate vs. Orbit Type

- The Orbit with more success in average were:
  - ES-L1
  - GEO
  - HEO
  - SSD



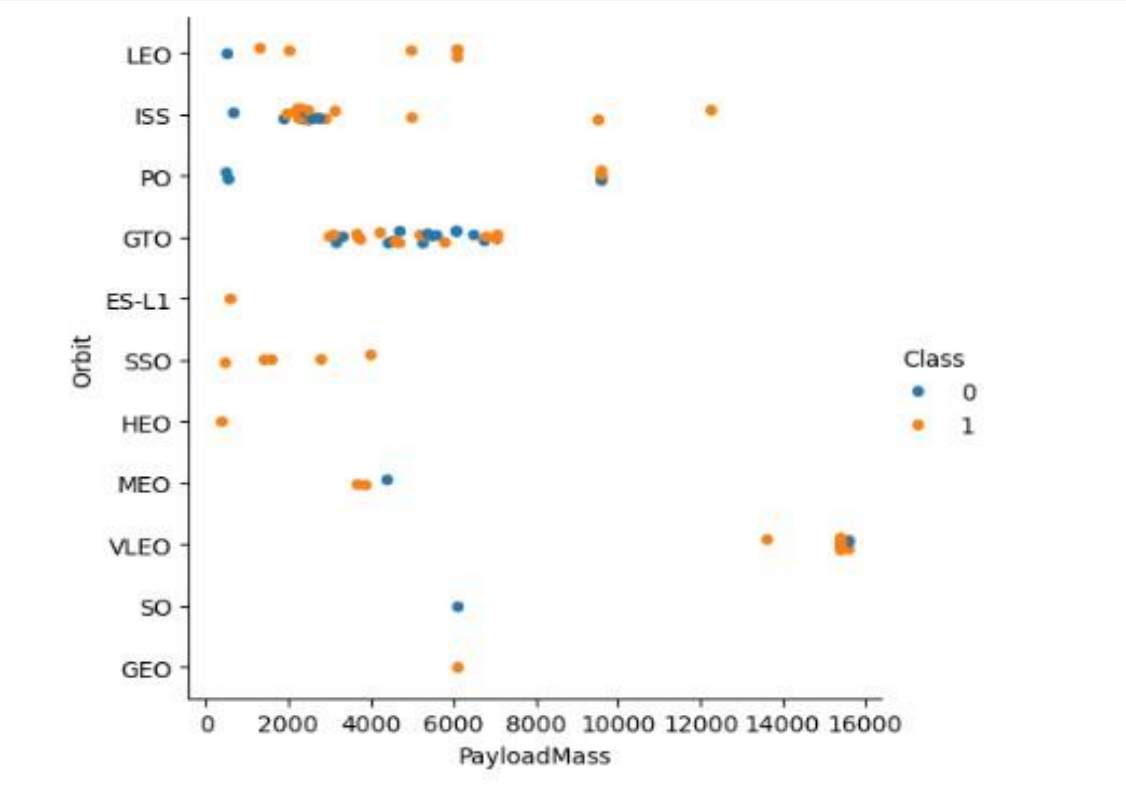
Flight Number vs. Orbit Type

- VLEO that start recently shown a great success rate.
- The success increase over the time.



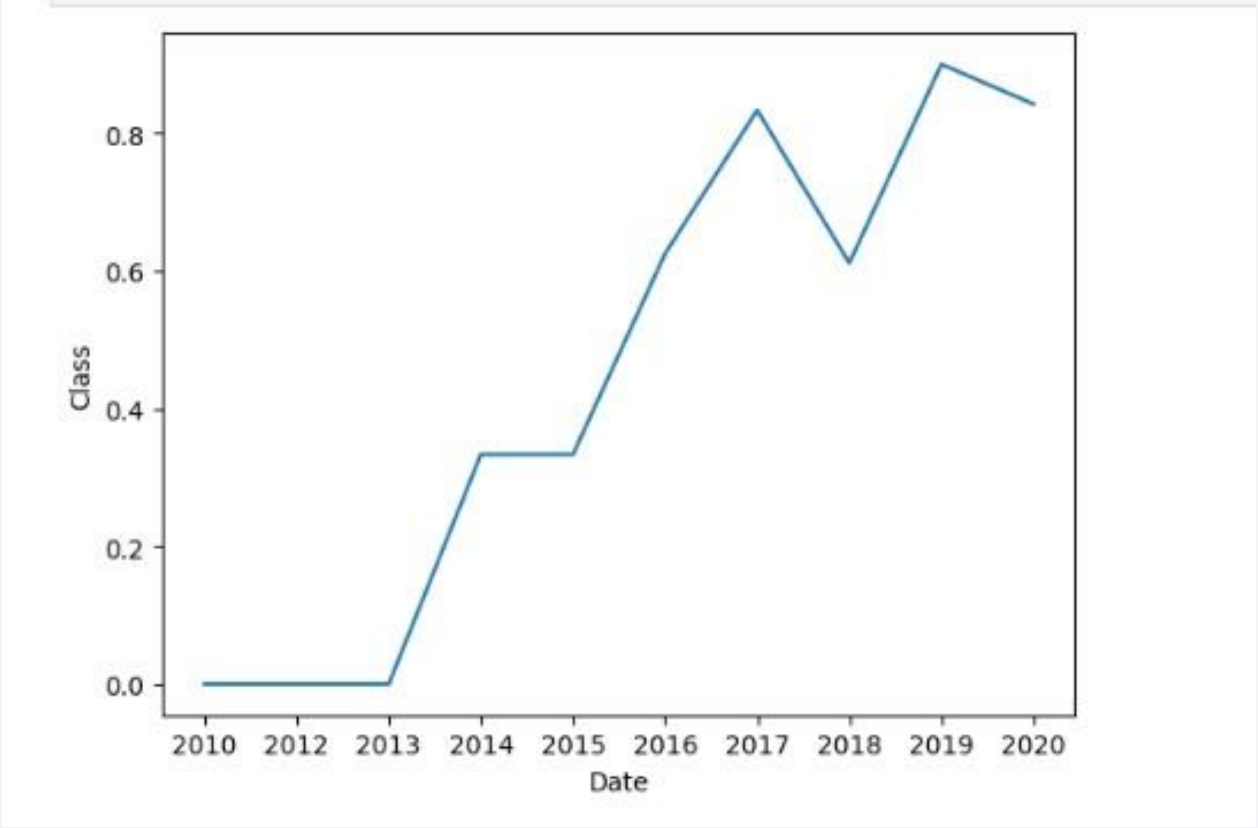
Payload vs. Orbit Type

- Apparently there aren't a relationship between the success rate with the Orbit and Payload Mass.



Launch Success Yearly Trend

- The success rate increase over the time from 2013.
- In 2018 and 2020 decrease a little bit but nothing worried.



All Launch Site Names

- Selecting unique ‘Launch_Site’ found that there are four launch sites.

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

Launch Site Names Begin with 'CCA'

- The first launch that begin with ‘CCA’ start in 2010-06-04.
- Found this was possible selecting all the records where the Launch_Site was ‘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

Total Payload Mass

- Found this was possible summing Payload mass where the customer was the NASA (CRS)

sum(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

- Found this was possible selecting average payload mass carried where the booster version was F9 v1.1.

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Found this was possible selecting de the min Date where the Landing_Outcome was 'Success (ground pad)'.

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Found this was possible selecting distinct Booster_Version where Landing_Outcome was 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

- Found this was possible selecting Mission_Outcome, count(Mission_Outcome) and then grouping by Mission_Outcome.

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

Boosters Carried Maximum Payload

- Found this was possible selecting the Booster_Version and then filtering the PAYLOAD_MASS__KG_, this was possible using a subquery where select the max PAYLOAD_MASS__KG_

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

- There are a lot of Landing_Outcome that ‘No attempt’.

- This was possible selecting Landing_Outcome, COUNT(*) AS outcome_count, WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY outcome_count DESC;

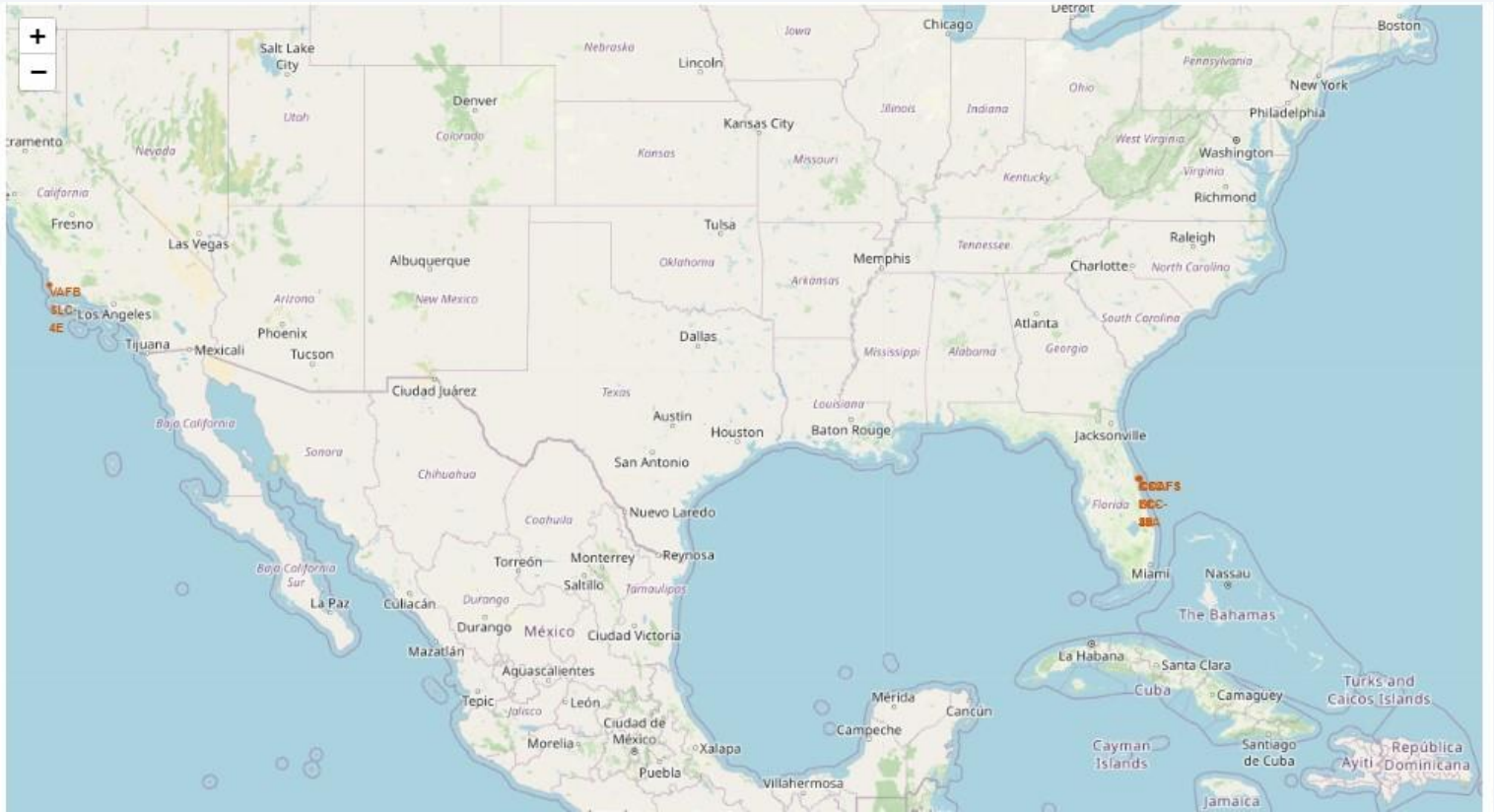
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

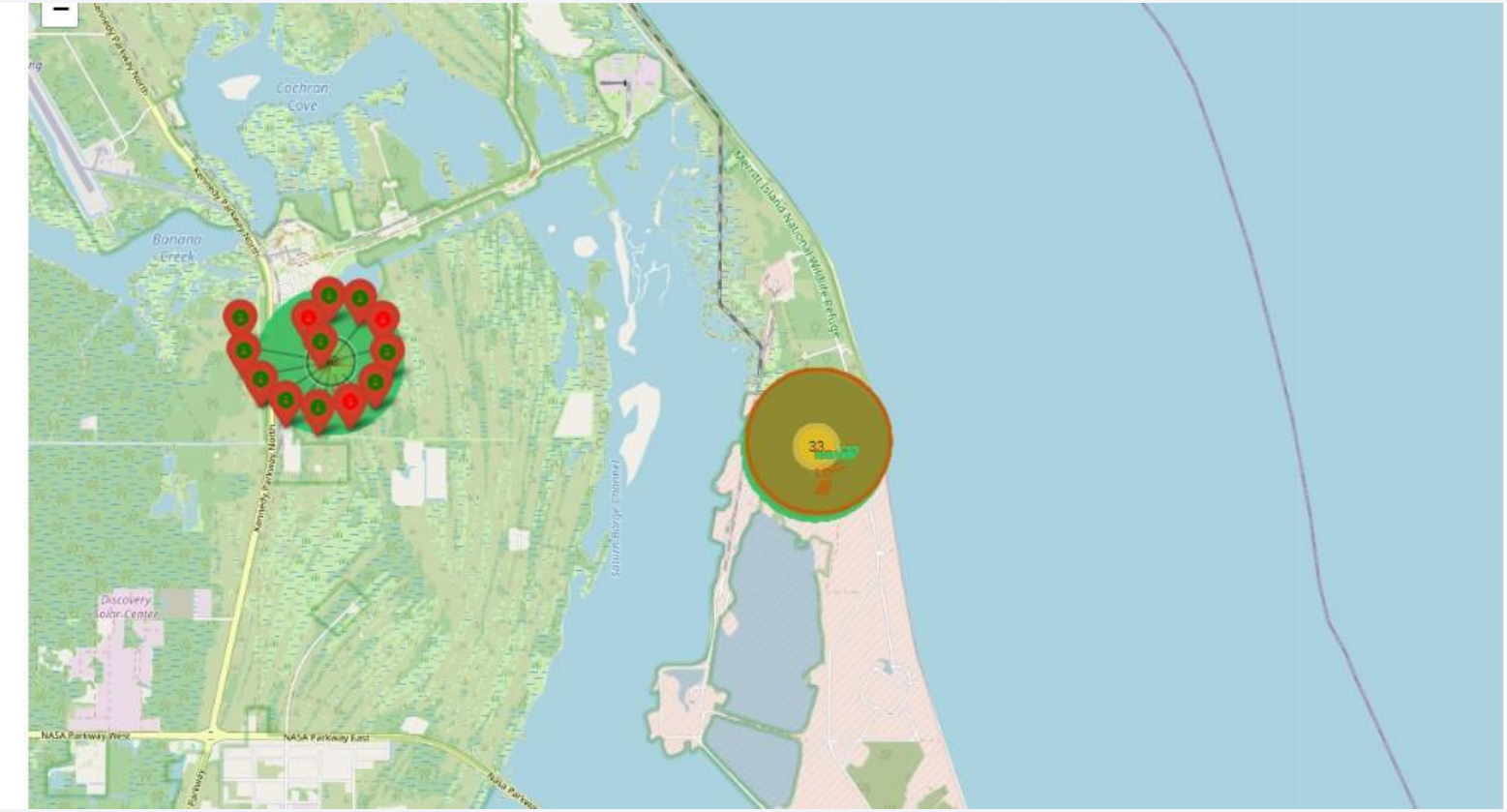
Map with all launch sites

- The launch site are so near to the sea.



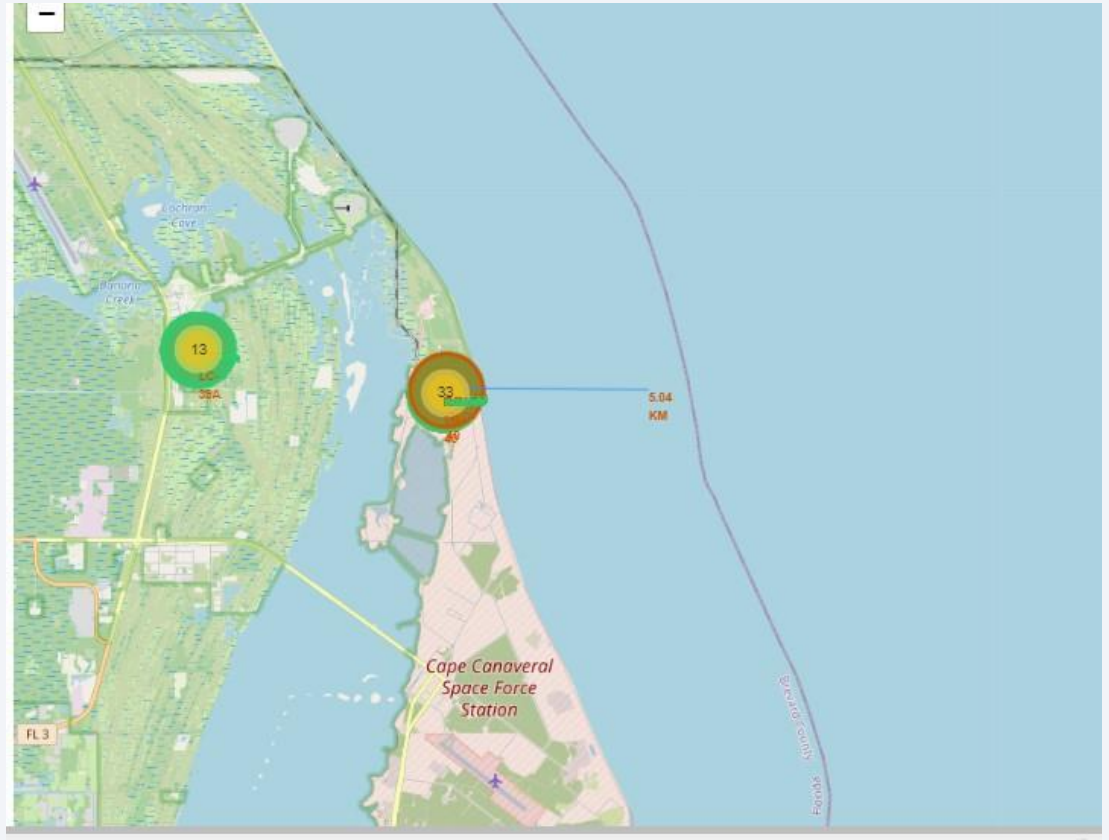
Launch Outcomes

- The area where there a lot of launch outcomes, the green marker means success and the red means fail.



Line to the coast

- In the map we can see a line and the distance between one launch site and the a point.

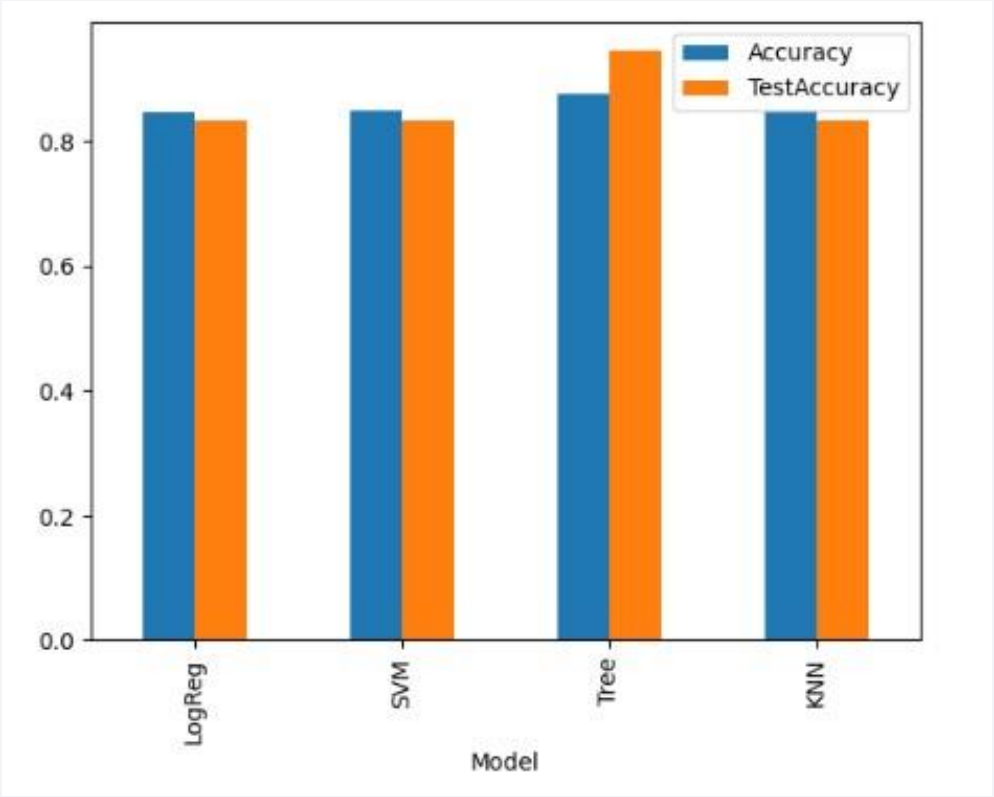


Section 4

Predictive Analysis (Classification)

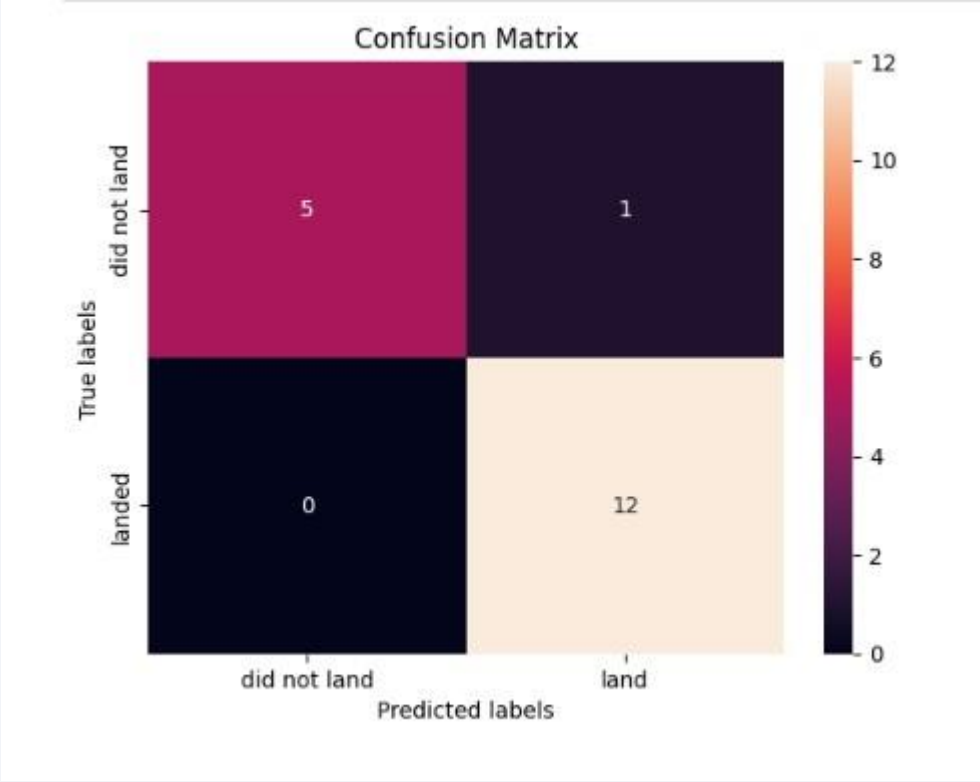
Classification Accuracy

- The Classification model with most accuracy and test accuracy was de Decision Tree model.



Confusion Matrix

- The Confusion Matrix shown as 5 of 6 attempts acert that did not land, and 12 of 12 attempts acert that did land.



Conclusions

- The success rate increase over the time.
- The launch sites are near to the coast.
- Apparently there aren’t a relationship between the success rate with the Orbit and Payload Mass.
- The Decision Tree is the best model for our purpose.

Thank you!