IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Masoud Amirrezai Haradasht
15.09.2023

Outline

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

Executive Summary

❖ **Summary of methodologies**

- Data Collection – SpaceX API and Scraping
- Data Wrangling
- EDA with Data Visualization
- EDA with SQL
- Building an Interactive Map with Folium
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

❖ **Summary of all results**

- As the flight numbers increases, the first stage is more likely to land successfully.

- Having heavy payloads has a positive effect on the success rate of the lunches for Polar, LEO and ISS Orbits.

- The KSC LC-39A lunch sites has the highest success rate, while the CCAFS SLC- 40 has the lowest.

-"FT" Boster has the highest success rate among different Booster versions.

- From 2013 the success rate has started to increase continuesly except for the year 2014-2015. From 2017 to 2018 there is an updown trend, however it has started to improve from 2018 to almost 90%. From 2019 to 2020 the success rate is slightly decreased.

Introduction

❖ **Project background:**

- In this project, our goal is to forecast the successful landing of the Falcon 9 first stage. SpaceX showcases Falcon 9 rocket launches on its website, priced at 62 million dollars each, a significantly lower cost compared to other providers whose rockets can go upward of 165 million dollars. This substantial cost difference is primarily due to SpaceX's innovative approach of reusing the first stage of the rocket. Hence, by accurately predicting the first stage's landing, we can ascertain the overall launch cost. This crucial information becomes invaluable when a competing company seeks to bid against SpaceX for a rocket launch contract study both successful and unsuccessful launch examples.

❖ **Problems find answers for?**

- What are the features that has the most influence on landing of the first stage, like the amount of payload, the location of the launch site, the booster version, or the orbit that the missile get launched to.
- Is the first stage going to land successfully ?

Section 1

Methodology

Methodology

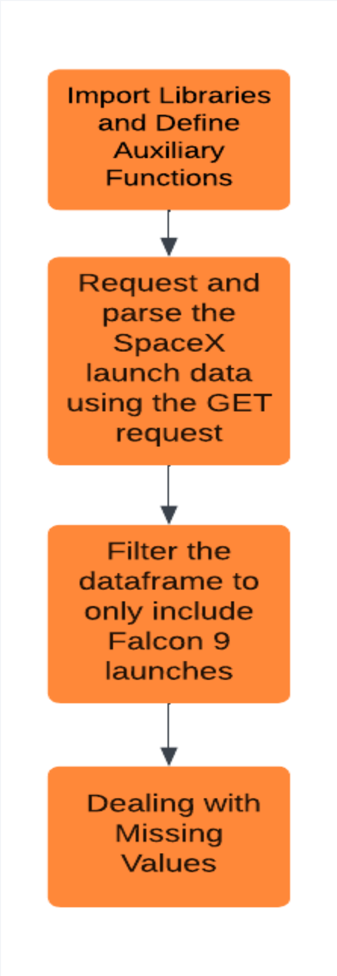
Executive Summary

- Data collection methodology
- Performing data wrangling
- Exploratory data analysis (EDA)
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection – SpaceX API

- The data collection with SpaceX REST API is presented

In the flowchart:



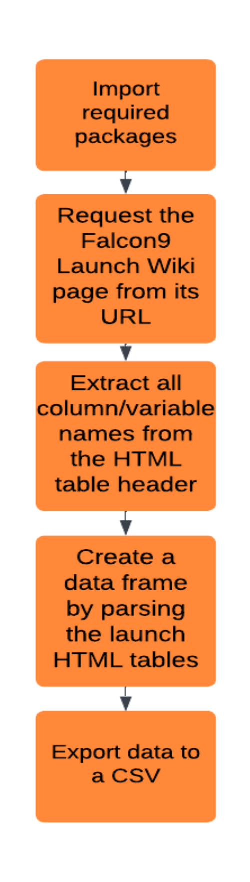
- GitHub URL:

https://github.com/Masoud6990/Courcera_Capstone.git

(Please see Data Collection_Webscraping.ipynb file in the link above)

Data Collection - Scraping

- The web scraping process is presented in the flowchart



- GitHub URL:

https://github.com/Masoud6990/Courcera_Capstone.git

(Please see Data Collection_Webscraping.ipynb file in the link above)

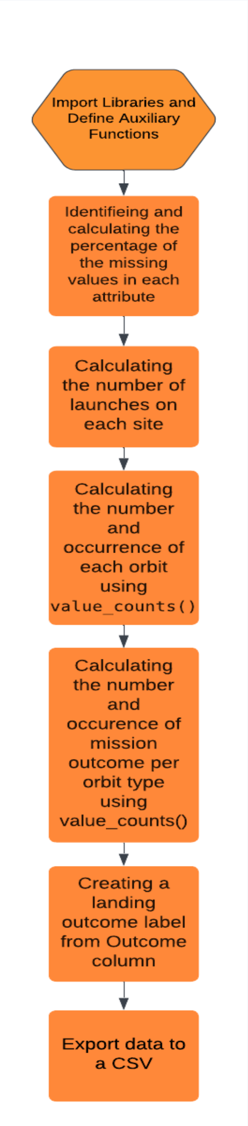
Data Wrangling

- The data wrangling process is carried out in order to perform exploratory Data Analysis and determine training Labels

- GitHub URL:

https://github.com/Masoud6990/Courcera_Capstone.git

(Please see Data Wrangling.ipynb file in the link above)



EDA with Data Visualization

- To obtain some preliminary insights about how each variable would affect the success rate, and thus selecting features that has the most influence the following figures have been visualized:

  - a scatter plot of flightnumber vs. payloadmass while overlaying the outcome of the launch

  - a scatter plot of flightnumber vs. lunchsite while overlaying the outcome of the launch

  - a scatter plot of lunchsite vs. payloadmass while overlaying the outcome of the launch

- To find which orbits have highest success rate

  - a bar chart of success rate vs. orbit type

- To reveal the relationship between Payload and Orbit type

  - a scatter point chart with x axis to be payload and y axis to be the orbit, and hue to be the class value

- To get the average launch success trend

  - a line chart with x axis to be Year and y axis to be average success rate

- GitHub URL: https://github.com/Masoud6990/Courcera_Capstone.git (Please see EDA with **Data Visualisation.ipynb** file in the link above)

EDA with SQL

Summarizing the SQL queries performed

- **Loading the SQL extension and establish a connection with the database :**

  - %load_ext sql

- **Displaying the names of the unique launch sites in the space mission:**

  - %sql select DISTINCT "Launch_Site" from SPACEXTABLE

- **Displaying 5 records where launch sites begin with the string 'CCA':**

  - %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5

- **Displaying the total payload mass carried by boosters launched by NASA (CRS):**

  - %sql select sum("PAYLOAD_MASS__KG_") as 'total payload_NASA(CRS)' from SPACEXTABLE where Customer='NASA (CRS)'

- **Displaying average payload mass carried by booster version F9 v1.1:**

  - %sql select avg("PAYLOAD_MASS__KG_") as 'average_payload_versio_F9_v1.1' from SPACEXTABLE where Booster_Version like'F9 v1.1%'

- **Listing the date when the first successful landing outcome in ground pad was achieved:**

  - %sql select min(Date) from SPACEXTABLE where Landing_Outcome ='Success (ground pad)'

EDA with SQL

Summarizing the SQL queries performed

- **Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:**

  - %sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome ='Success (drone ship)' and "PAYLOAD_MASS__KG_">4000 and "PAYLOAD_MASS__KG_"<6000

- **Listing the total number of successful and failure mission outcomes:**

  -%sql select SUM(CASE WHEN Landing_Outcome like '%Success%' THEN 1 ELSE 0 END) as 'No._successful_outcome',SUM(CASE WHEN Landing_Outcome like '%Failure%' THEN 1 ELSE 0 END) as 'No._Failure_outcome' from SPACEXTABLE

Listing the names of the booster_versions which have carried the maximum payload mass. Use a subquery:

  - %sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

- **Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015:**

  -%sql select substr(Date, 1, 4) as Year, substr(Date, 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome='Failure (drone ship)' and substr(Date, 1, 4)= '2015'

- **Ranking 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:**

  -%sql SELECT Date, Landing_Outcome, COUNT(*) as Count FROM SPACEXTABLE where Date >='2010-06-04' and Date <='2017-03-20' GROUP BY Landing_Outcome Order by Count Desc

GitHub URL:

- https://github.com/Masoud6990/Courcera_Capstone.git (Please see EDA with SQL.ipynb file in the link above)

Build an Interactive Map with Folium

Added map objects to the folium map and why?

Circles: Adding a circle for each launch site on the map.

Why? : in order to give an intuitive insights about where the launch sites are located, and whether are they in proximity to the Equator line, or the coast.

Markers: Adding the launch outcomes for each lunch site on the map. If a launch was successful, a green marker and if it was failed, a red marker

Why? : Now one can easily identify which launch sites have relatively high success rates.

Lines: lines between a launch site to its closest coastline, city, railway, highway.

Why? : Are the launch sites in a close proximity to railways, cities and so on?

GitHub URL:

- https://github.com/Masoud6990/Courcera_Capstone.git (Please see Interactive Map with Folium.ipynb file in the link)

Build a Dashboard with Plotly Dash

What plots/graphs have been added to the dashboard:

Following cases are presented as an interactive pie chart and a scatter plot showing the outcomes of the lunches vs. Payload Mass (kg), while the Payload could be varied by the user:

- Success count for all sites, Success vs. Failed Launches for CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40

- **Why?**

Performing interactive visual analytics on SpaceX launch data in real-time helps to obtain insights to answer the following questions:

- Which site has the largest successful launches?
- Which site has the highest launch success rate?
- Which payload range(s) has the highest launch success rate?
- Which payload range(s) has the lowest launch success rate?
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

- **GitHub URL:**

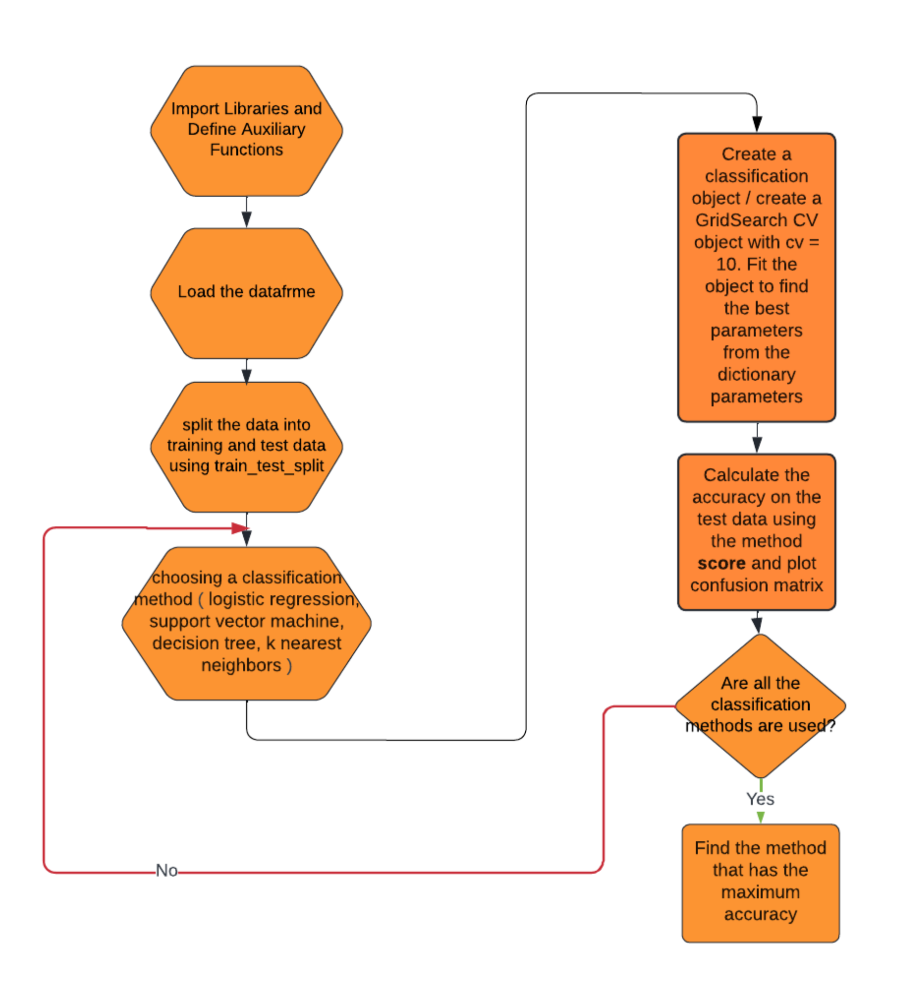
https://github.com/Masoud6990/Courcera_Capstone.git (Please see Dashboard with Plotly Dash.ipynb file in the link. Please consider that this code is written Skills Network Labs (SN Labs) and thus should be run in the same environment. The GitHub cannot run the code)

Predictive Analysis (Classification)

- **GitHub URL:**

https://github.com/Masoud6990/Courcera_Capstone.git

(Please see Predictive Analysis_Classification.ipynb file in the link.



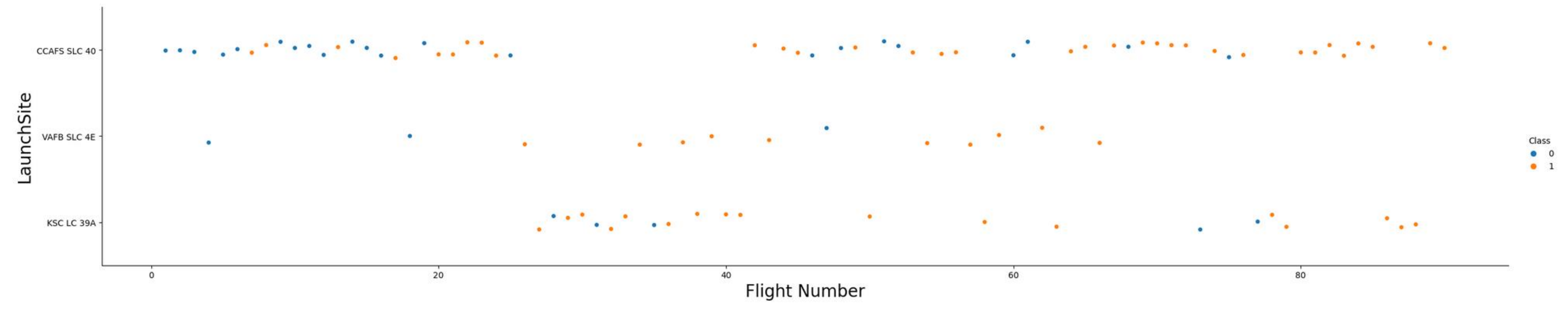
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

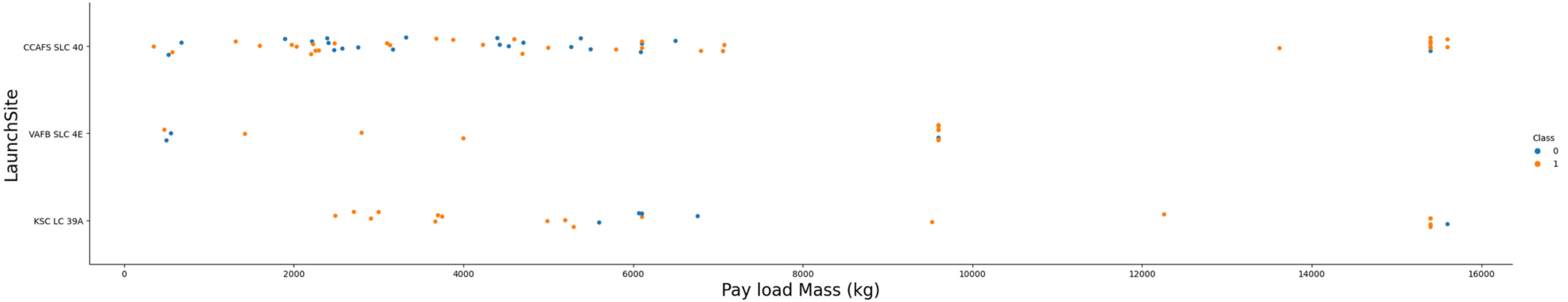
Insights drawn from EDA

Flight Number vs. Launch Site



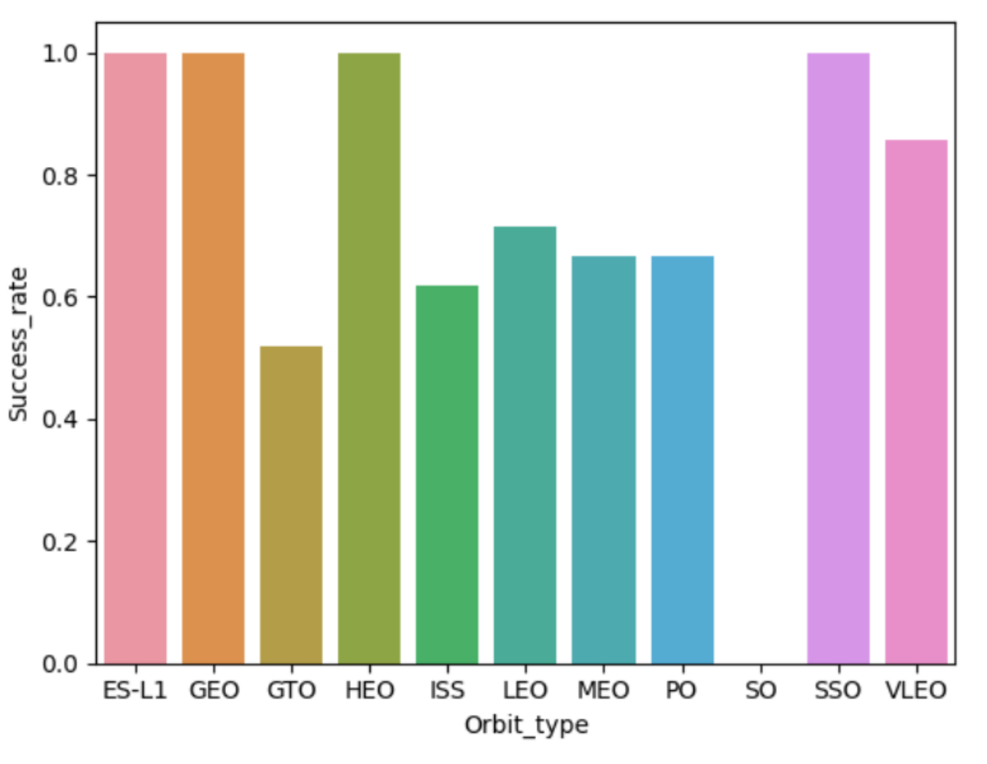
- In the above plot, it can be seen that as the Flight Numbers increases, the first stage is more likely to land successfully, in all three lauch sites. It can also be observed that the CCAFS SLC 40, has the largest number of launches in general.

Payload vs. Launch Site



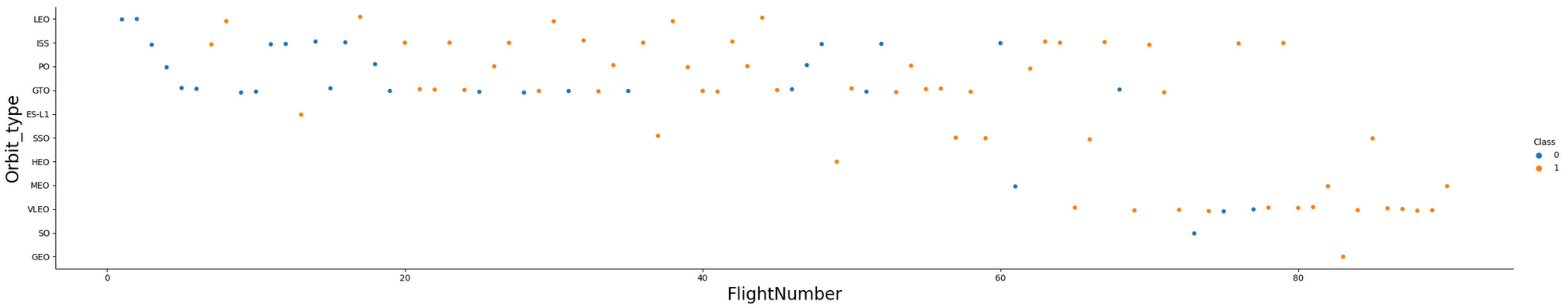
- It can be seen that the number of launches with lighter Payloads are higher. There is no evident relationship between the Pay Load mass and the outcome of the the first stage. Additionally, one can see that, for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



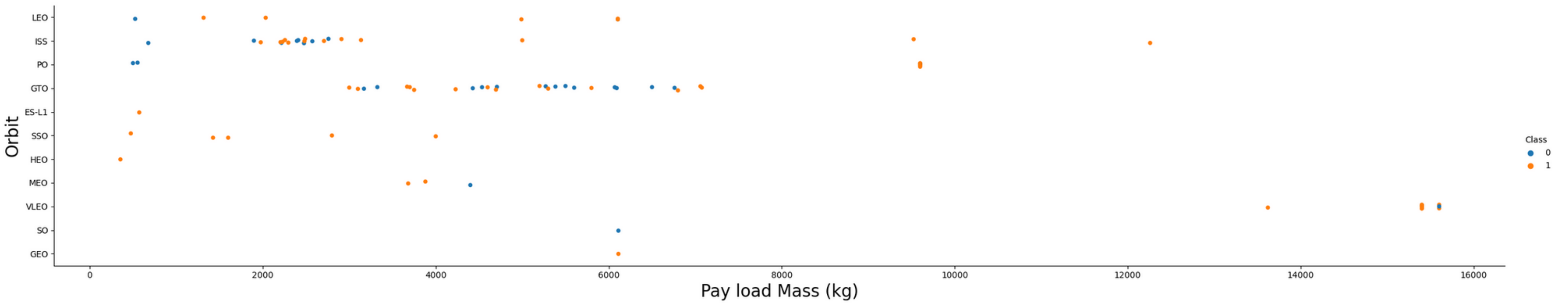
- It can be seen that 4 Orbit types have the best success rate, however, SO, has zero success rate.

Flight Number vs. Orbit Type



- It can be seen that most laucnhes are carried out in 5 Orbits namely, LEO, ISS, PO, GTO, and VLEO. Moreover, the success rate in the LEO orbit appears to be related to the number of flights, however, there seems to be no relationship between flight number and success rate in other orbits.

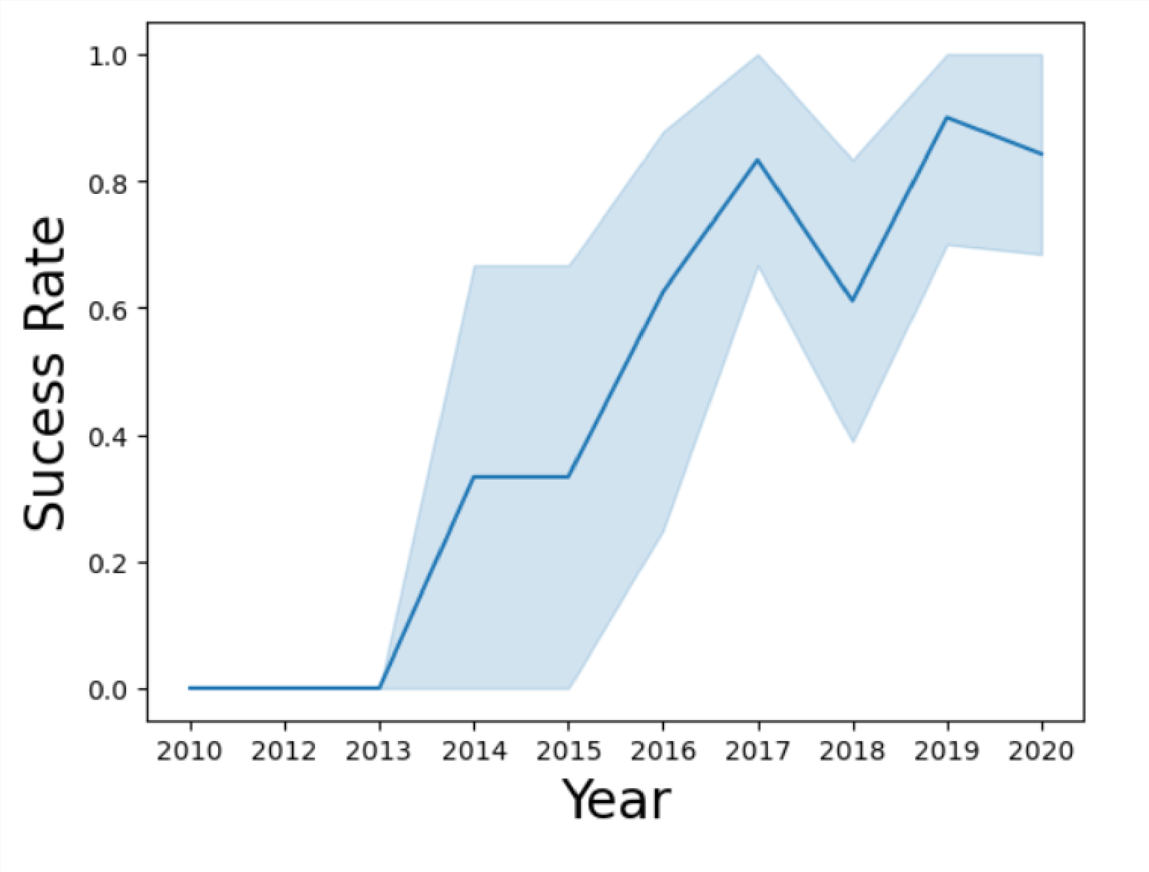
Payload vs. Orbit Type



- It can be observed that having heavy payloads has a positive effect on the success rate of the lunches for Polar, LEO and ISS. However, for GTO one cannot distinguish this effect, since both positive and negative landing (unsuccessful mission) are present by increasing the Payload.

Launch Success Yearly Trend

- One can see that after 2013 the success rate has started to increase continuesly except for the year 2014-2015. From 2017 to 2018 there is an updown trend, however it has started to improve from 2018 to almost 90%. From 2019 to 2020 the success rate is slightly decreased.



All Launch Site Names

- Finding the names of the unique launch sites:
- Query:

```
%sql select DISTINCT "Launch_Site" from SPACEXTABLE
```

- Results and explanation:

Using “DISTINCT” helps to get The unique names.

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

Launch Site Names Begin with 'CCA'

- Finding 5 records where launch sites begin with `CCA`:
- Query:

```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
```

- Results and explanation:

using % in 'CCA%' helps to find every launch site that begins with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculating the total payload carried by boosters from NASA:
- Query:

```
%sql select sum("PAYLOAD_MASS__KG_") as 'total payload_NASA(CRS)'  from SPACEXTABLE where Customer='NASA (CRS)'
```

- Results and explanation:

  Using the function “sum” we could sum up all payloads.

[9]:

total payload_NASA(CRS)
45596

Average Payload Mass by F9 v1.1

- Calculating the average payload mass carried by booster version F9 v1.1
- Query:

```
%sql select avg("PAYLOAD_MASS__KG_") as 'average_payload_versio_F9_v1.1' from SPACEXTABLE where
Booster_Version like 'F9 v1.1%'
```

- Results and explanation:

  - here we need to use the ‘avg’ function, meanwhile filtering the Booster versions starting with ‘F9 v1.1’ which is accomplished by ‘like’ method.

[10]:

average_payload_versio_F9_v1.1
2534.6666666666665

First Successful Ground Landing Date

- Finding the dates of the first successful landing outcome on ground pad

- Query:

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome ='Success (ground pad)'
```

- Results and explanation:

  - here we need to use the 'min' function on the Date of the landings while filtering the data for the finding the successful attempts.

[11]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query:

%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome ='Success (drone ship)' and "PAYLOAD_MASS__KG_">4000 and "PAYLOAD_MASS__KG_"<6000

- Results and explanation:

  - the results are filtered for the successful outcomes on the drone ship and certain amount of the payload using ‘where’ and ‘and’

[12]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculating the total number of successful and failure mission outcomes
- Query:

```
%sql select SUM(CASE WHEN Landing_Outcome like '%Success%' THEN 1 ELSE 0 END) as
'No._successful_outcome', SUM(CASE WHEN Landing_Outcome like '%Failure%' THEN 1 ELSE 0 END) as
'No._Failure_outcome' from SPACEXTABLE
```

- Results and explanation:

  - We needed to first turn the success and failure outcomes to numbers so we could sum up them.

[13]:

No._successful_outcome	No._Failure_outcome
61	10

Boosters Carried Maximum Payload

- Listing the names of the boosters which have carried the maximum payload mass

- Query:

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_
in (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE )
```

[14]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- Results and explanation:

  - here we needed to use a sub-query to first find the launches with the maximum payload mass, so we can then extract the Booster version from this list.

2015 Launch Records

- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query:

```
%sql select substr(Date, 1, 4) as Year, substr(Date, 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE  where Landing_Outcome='Failure (drone ship)' and substr(Date, 1, 4)= '2015'
```

- Results and explanation:

  - using the function 'substr()', we could reach our needed information for year and month out of the Date.

[15]:

Year	Month	Landing_Outcome	Booster_Version	Launch_Site
2015	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Ranking 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
- **Query**:
- %sql SELECT Date, Landing_Outcome, COUNT(*) as Count FROM SPACEXTABLE where Date >='2010-06-04' and Date <='2017-03-20' GROUP BY Landing_Outcome Order by Count Desc
- **Results and explanation:**

  - Here we used the ‘count’ and ‘group by’ methods to count the number of landing outcomes and ordered them in descending order using “ Order by “ and ‘Desc’.

[16]:

Date	Landing_Outcome	Count
2012-05-22	No attempt	10
2015-12-22	Success (ground pad)	5
2016-08-04	Success (drone ship)	5
2015-10-01	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2015-06-28	Precluded (drone ship)	1
2010-08-12	Failure (parachute)	1

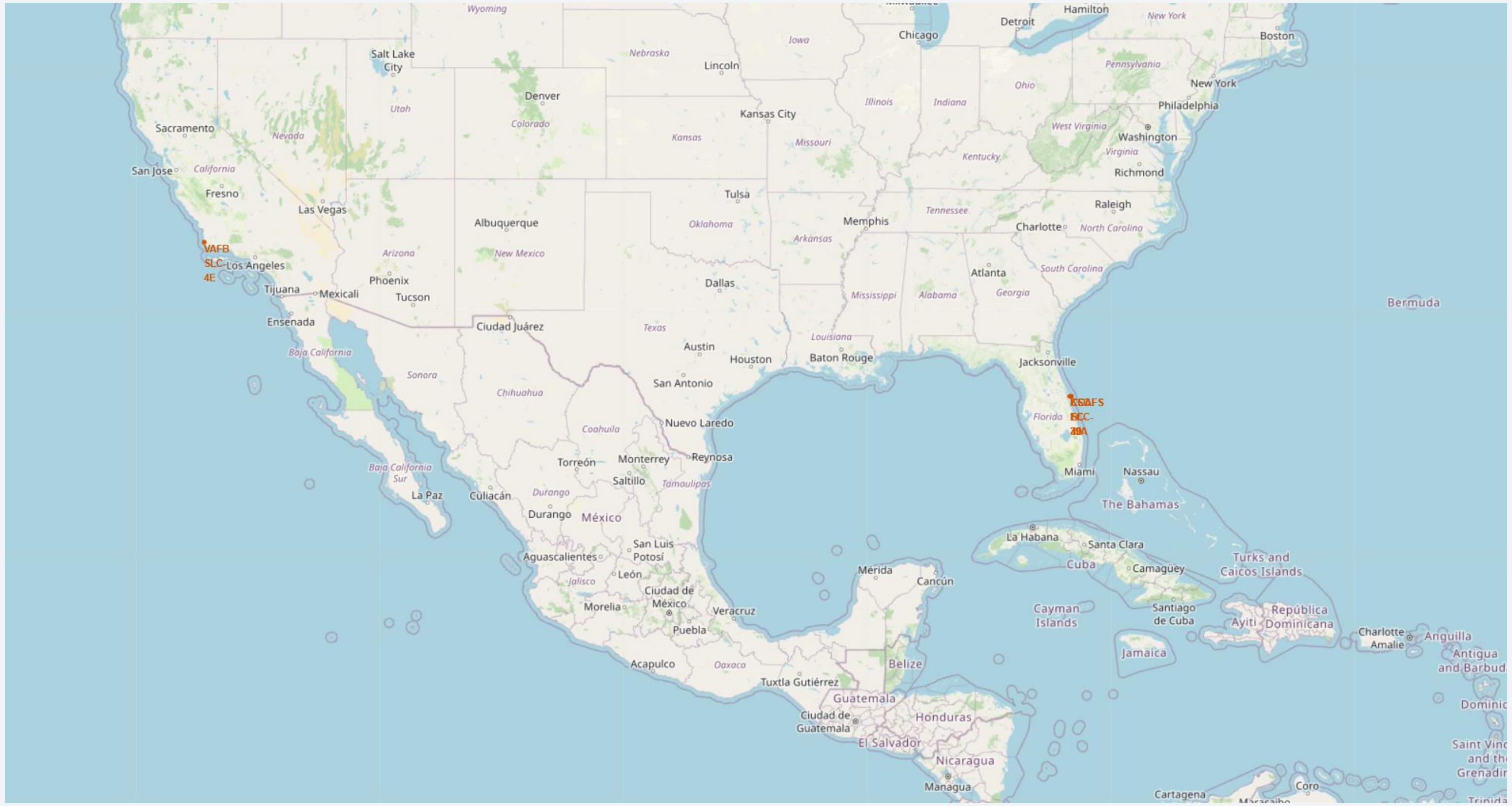
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- It could be observed that all the launch sites are located in the close proximity to the coast and the Equator line.

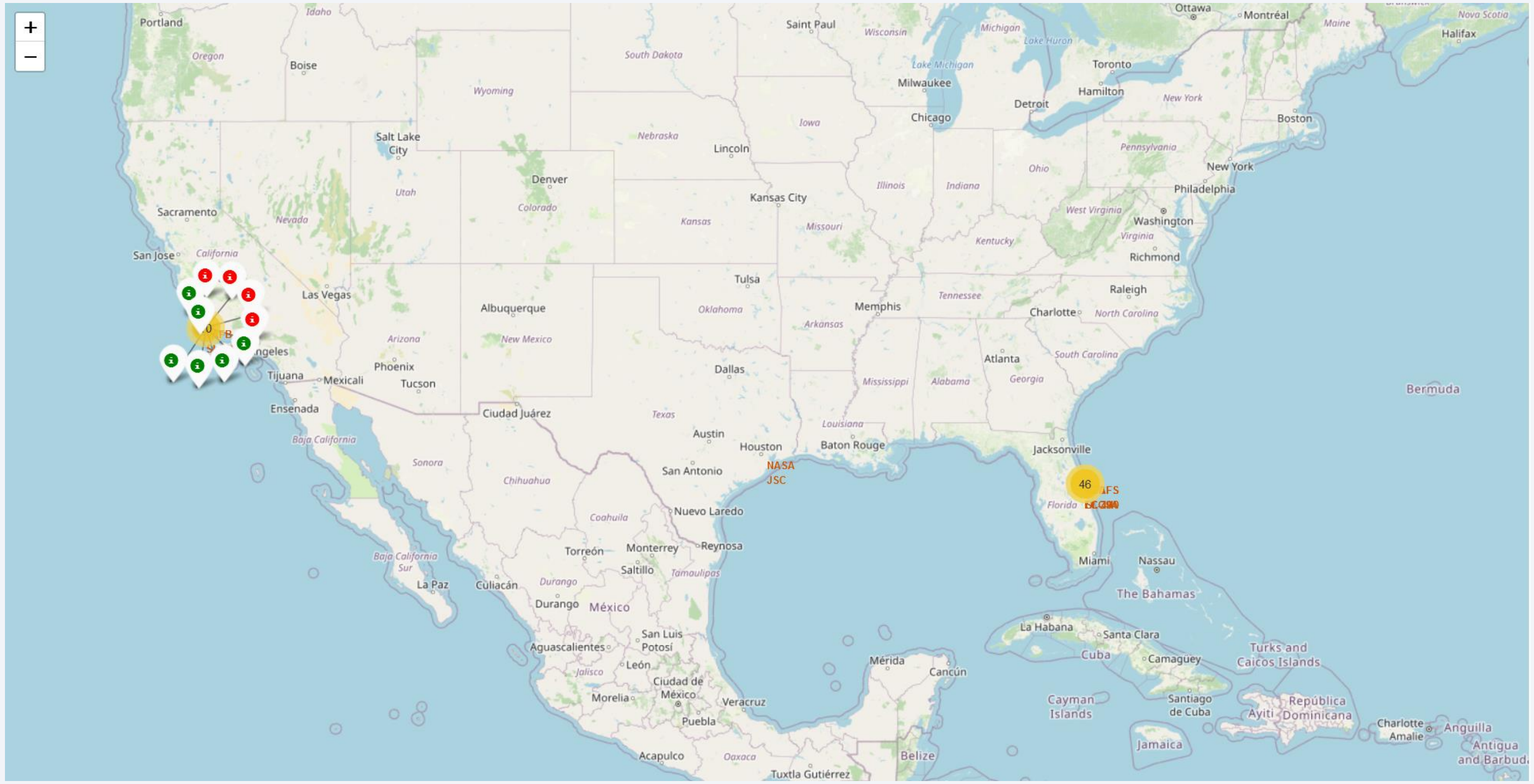
launch sites' location on global map



<Folium Map Screenshot 2>

❖ From the color-labeled markers in marker clusters, one can easily identify which launch sites have relatively high success rates. (By clicking on the location of the launch, the markers pop-up. For one of the sites this has been done in the picture)

The color-labeled launch outcomes on the map



<Folium Map Screenshot 3>

❖ From this figure, one can figure out, if the location of Launch sites are chosen to be in a certain distance to railways, highways, coastline and cities.

Distance of launch sites to their proximities

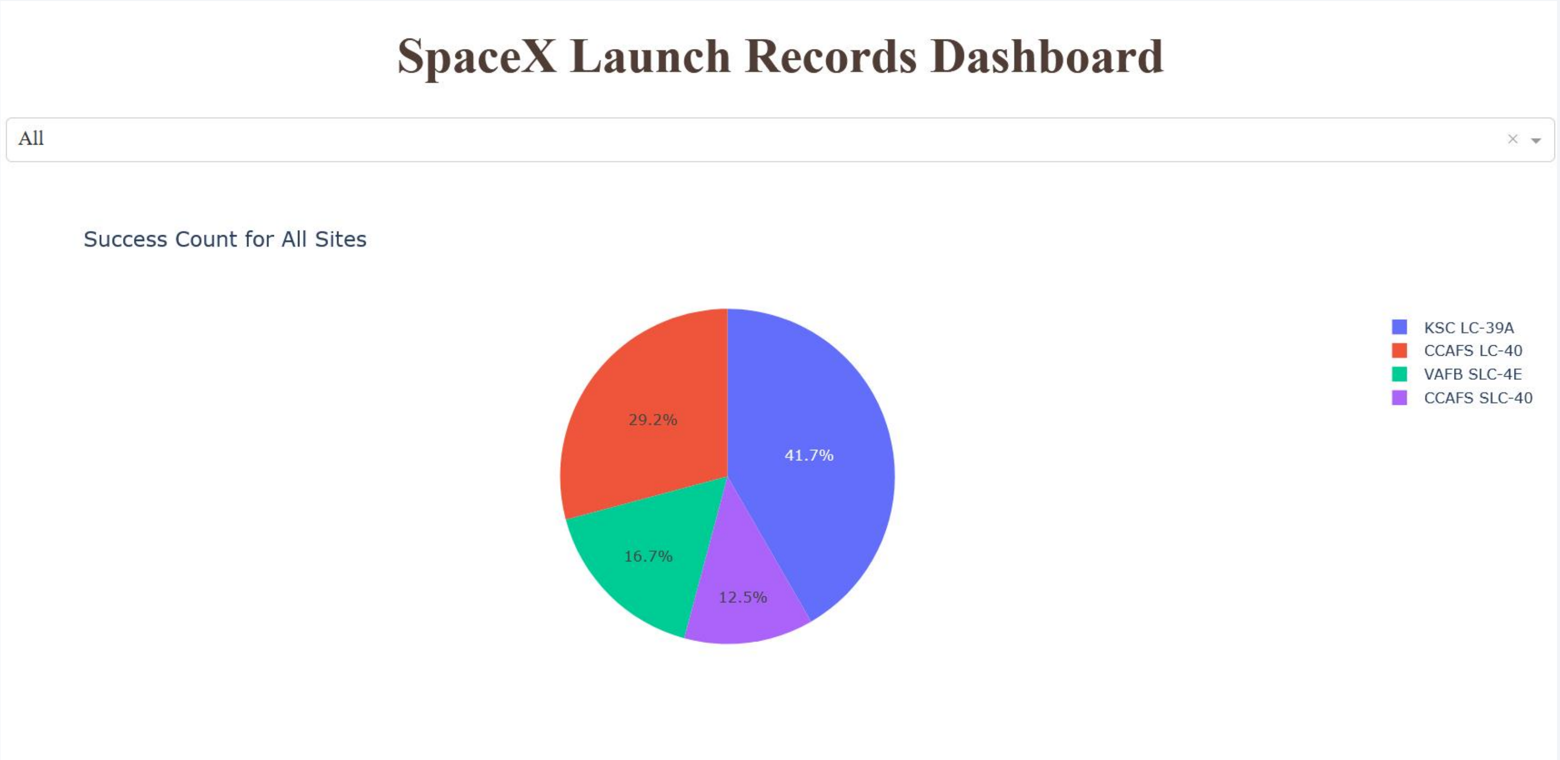
Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

launch success count for all sites in a piechart

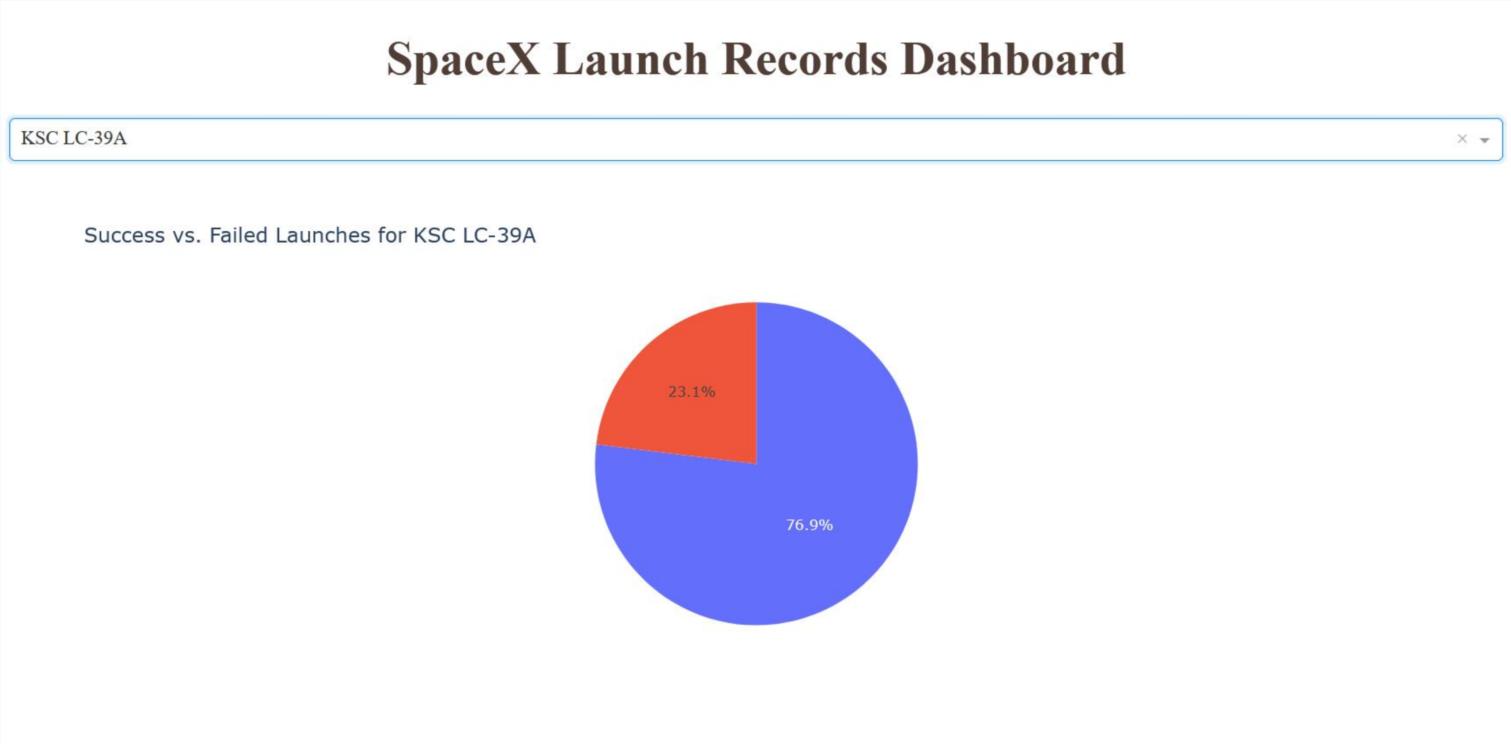
- ❖ It can be observed that KSC LC-39A has the highest success rate, while the CCAFS SLC- 40 has the lowest success rate among all sites.



<Dashboard Screenshot 1>

Piechart for the launch site with highest launch success ratio

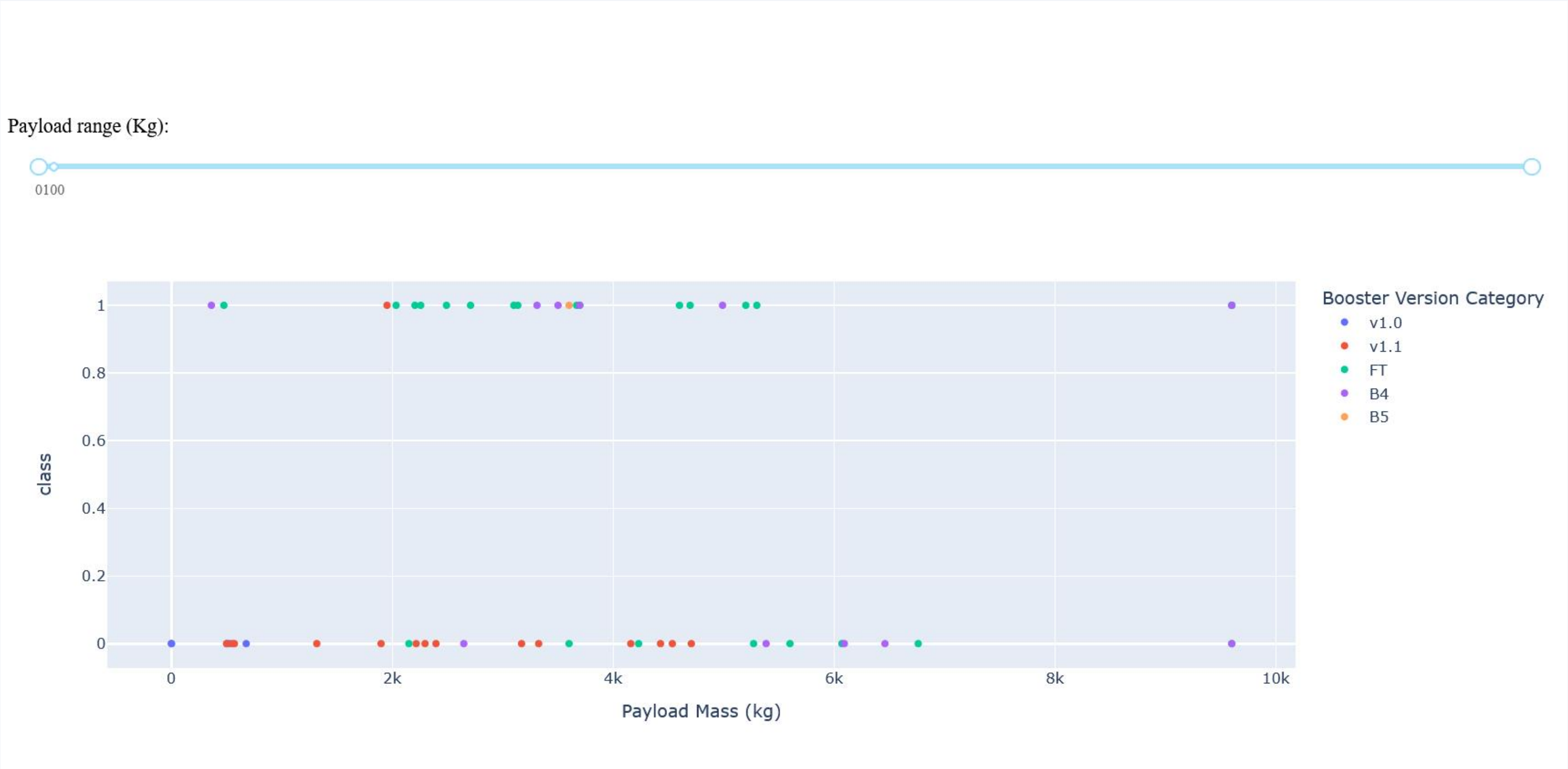
❖ From the Pie Chart it can be observed that in KSC LC-39A, which has the highest success rate between launch sites, 76.9% of launches were successful and 23.1% failed.



<Dashboard Screenshot 1>

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

❖ From the presented Chart it can be observed that 2K to 6K payload range Has the highest success rate, due to more outcomes classified as 1. Moreover, it could be observed that "FT" Boster hast the highest success rate among different Booster versions.

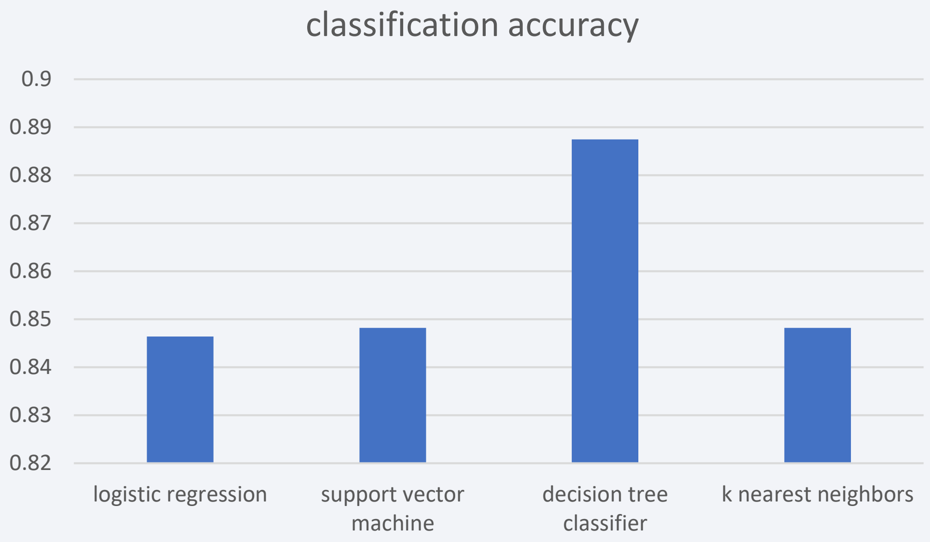


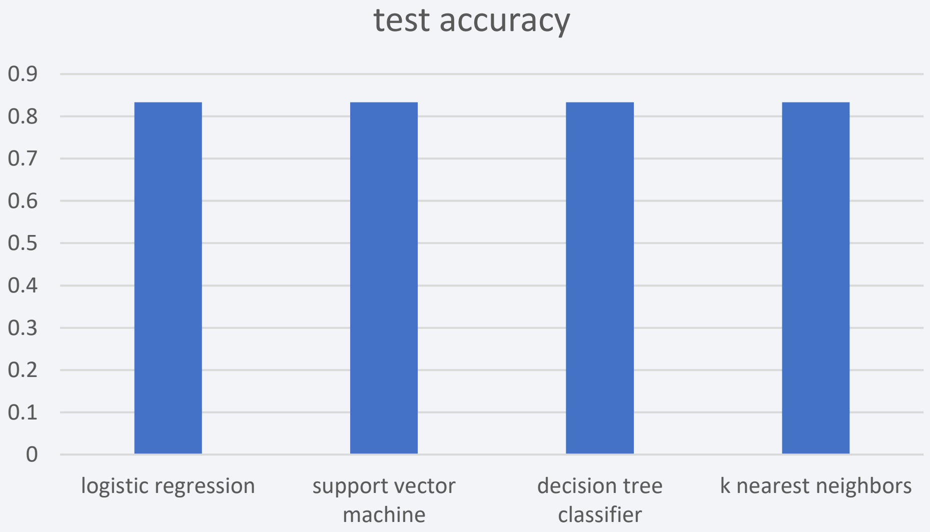
Section 5

Predictive Analysis (Classification)

Classification Accuracy

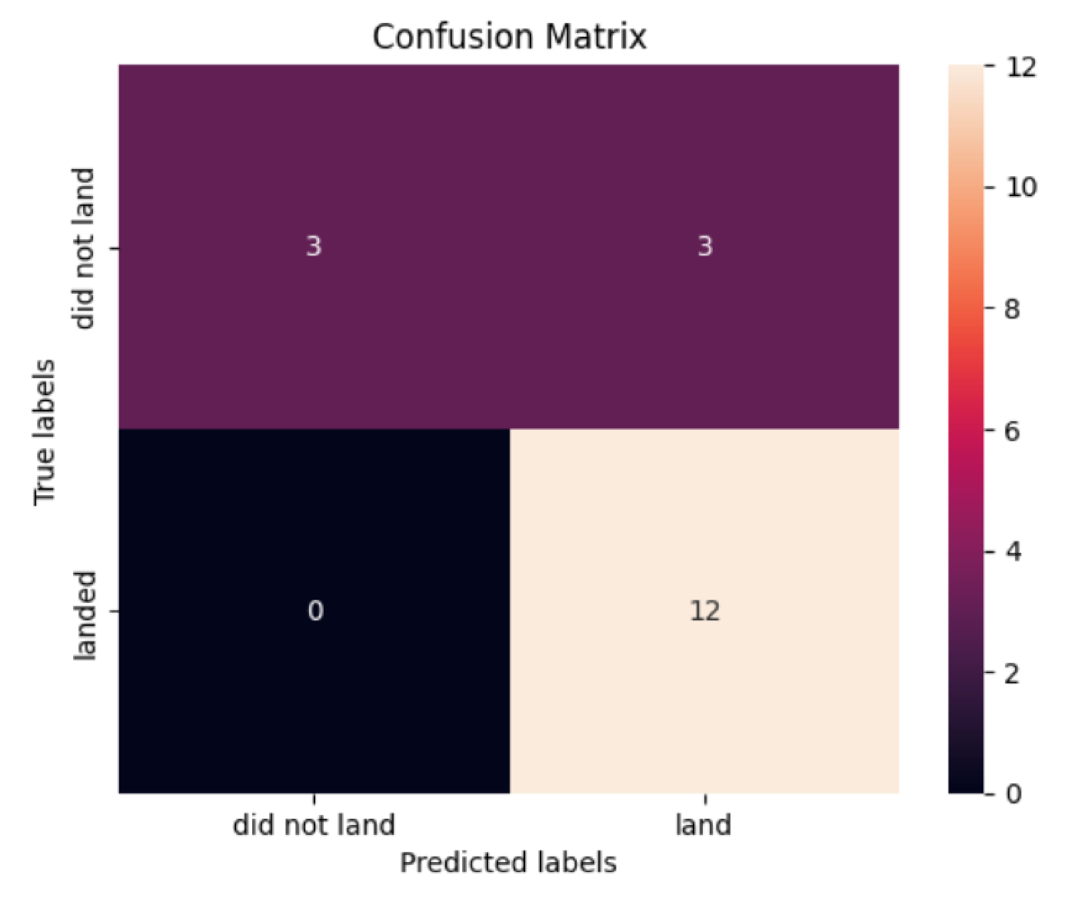
- According to the prsented figures, the “ decision tree“ has the best calssification accuracy, however the performance of the classifiers on the unseen data, test data, is same.





Confusion Matrix

❖ It can be observed that from 18 test samples, 15 samples have been classified correctly, while only 3 ''landed'' outputs are classified wrongly as '' did not land'', so the overall accuracy equals: 15/18=0,834.



Conclusions

- As the flight numbers increases, the first stage is more likely to land successfully.
- 4 Orbits have the best success rate, however, SO, has zero success rate.
- Most launches are carried out in 5 Orbits namely, LEO, ISS, PO, GTO, and VLEO.
- Having heavy payloads has a positive effect on the success rate of the lunches for Polar, LEO and ISS Orbits.
- All the launch sites are located in the close proximity to the coast and the Equator line.
- Between lunch sites the KSC LC-39A has the highest success rate, while the CCAFS SLC- 40 has the lowest success rate.
- In KSC LC-39A, which has the highest success rate between launch sites, 76.9% of launches were successful and 23.1% failed.
- The payloads in 2K to 6K range has the highest success rate.
- "FT" Boster hast the highest success rate among different Booster versions.
- From 2013 the success rate has started to increase continuesly except for the year 2014-2015. From 2017 to 2018 there is an updown trend, however it has started to improve from 2018 to almost 90%. From 2019 to 2020 the success rate is slightly decreased.

Thank you!