Data Science Capstone Project

By: Cheong Jin Hui

Table Of Content

Executive Summary

Project flow

- Collected and scraped data from SpaceX API and SpaceX Wikipedia Page
- Explored and visualised data using SQL, folium maps and dashboards.
- Predictive analysis on data using machine learning to determine if the first stage of Falcon 9 will land successfully
- Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.

Introduction

Project context and background

SpaceX is the most successful company in the commercial space age, making space travel more affordable. The company advertises Falcon 9 rocket launches on its website for $62 million, while other providers charge upwards of $165 million. This is because SpaceX can reuse the first stage of its rockets, which saves a significant amount of money.

If we can predict whether the first stage will land, we can determine the cost of a launch. We can do this by using public information and machine learning models.

By predicting whether the first stage will land, we can help SpaceX to reduce the cost of its launches and make space travel even more affordable.

Methodology

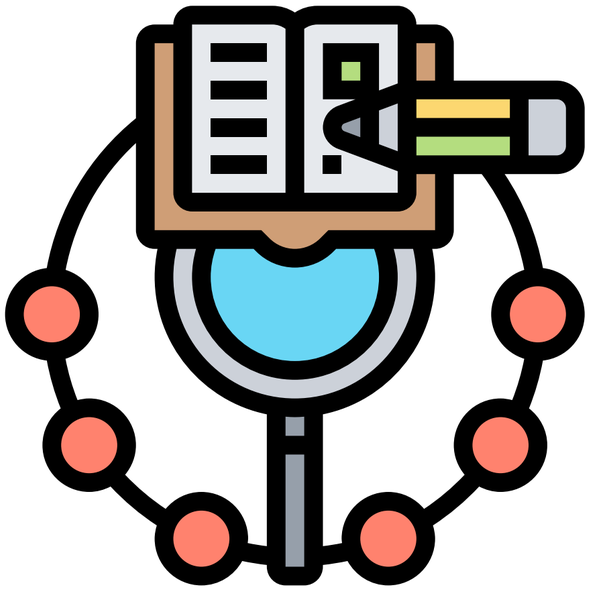
Data collection
- Use SpaceX Rest API
- Web Scrapping from Wikipedia

Data Wrangling
- Filter and fill up missing data
- Use One Hot Encoding to prepare the data to a binary classification

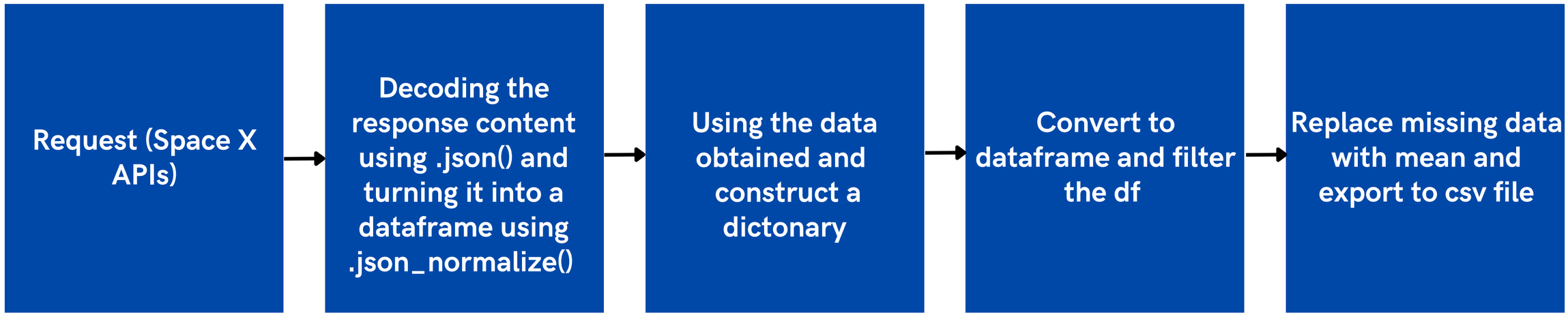
Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models
- Building, tuning and evaluation of classification models to get the best results



Data collection - SpaceX API

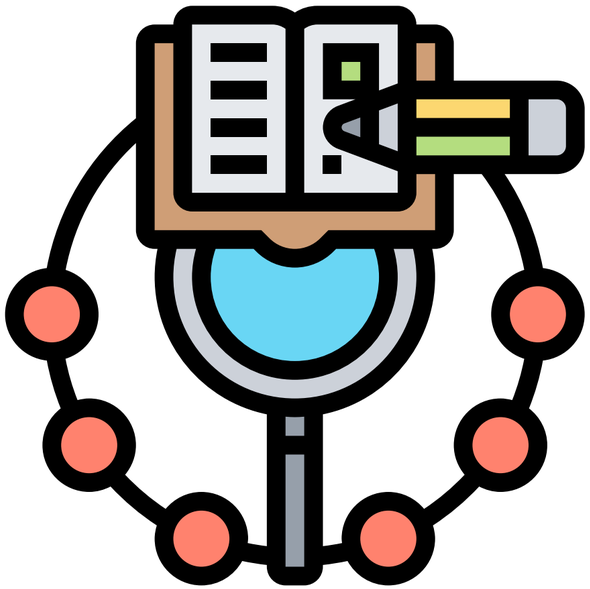


Data collection - Web Scraping

Data Wrangling

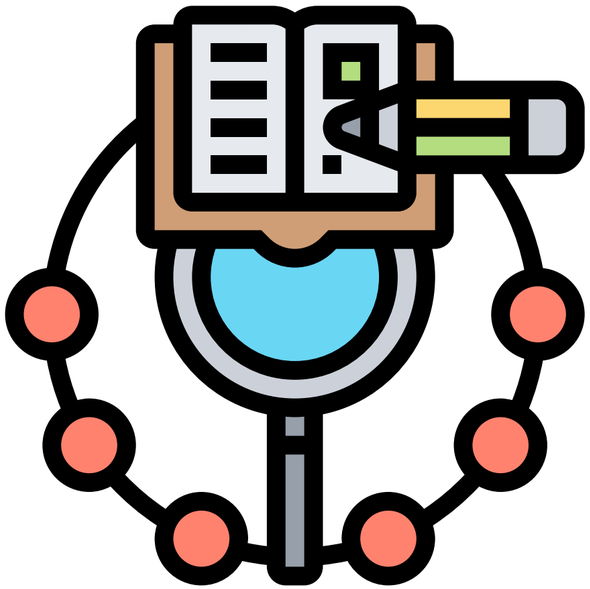
- The data set contains information about the success or failure of SpaceX booster landings. The data set includes information about the location of the landing, such as the ocean or a ground pad.

- The outcomes are converted into training labels, with "1" indicating a successful landing and "0" indicating an unsuccessful landing. This information can be used to train a machine learning model to predict whether a booster will land successfully.

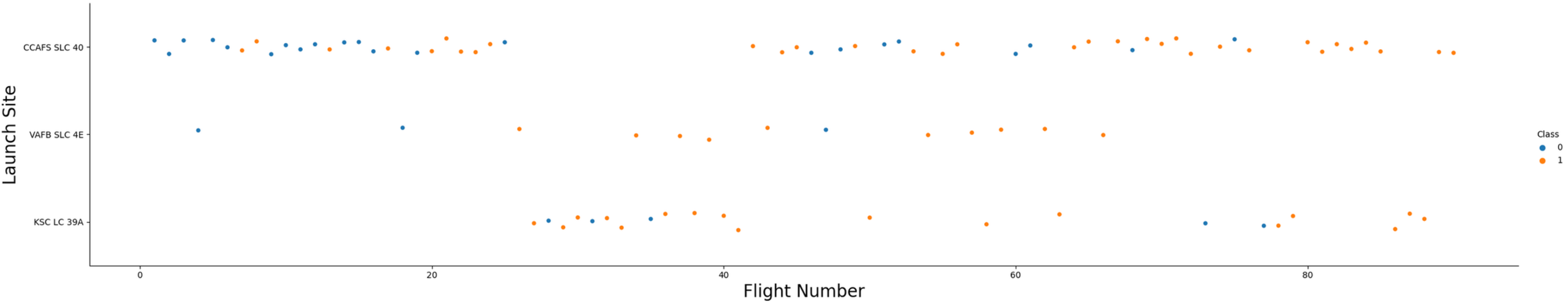


EDA with data visualization

- In this segment we plot various scatter plots to explore the relationship of various variables
- If a relationship exists, they could be used in machine learning model.
- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend



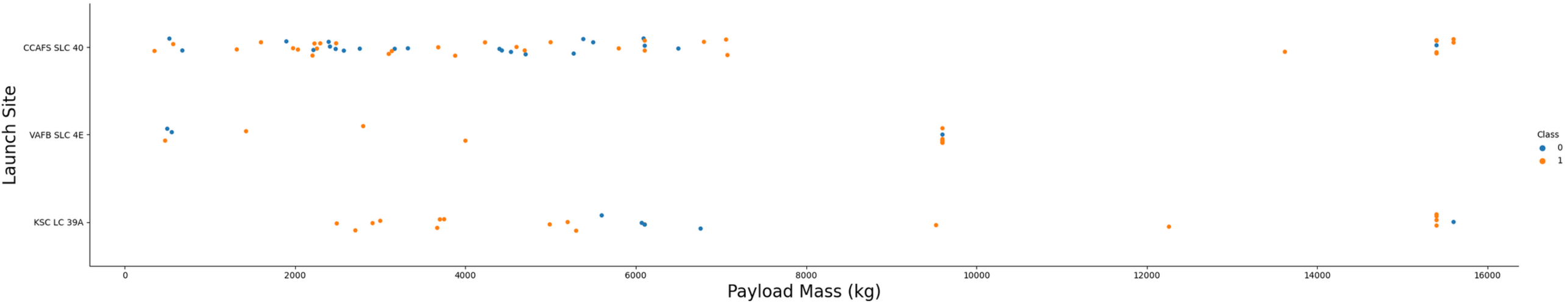
EDA with data visualization (results)



Explanation (Flight Number vs Launch Site)

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.

EDA with data visualization (results)

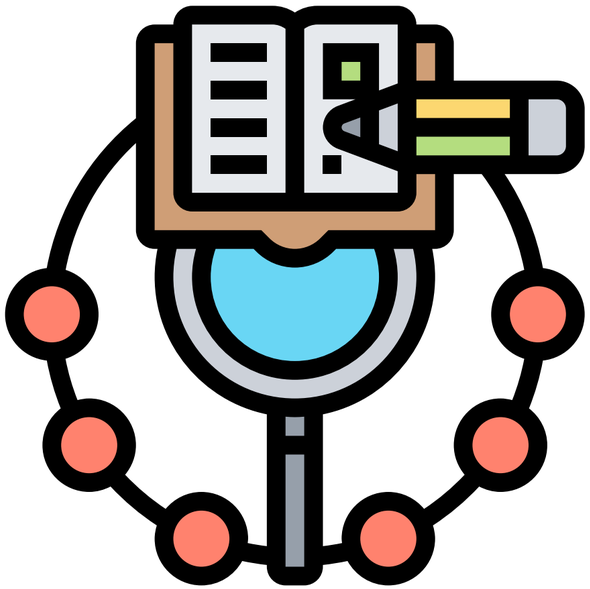


Explanation (Payload vs. Launch Site)

- For each launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successfull.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

EDA with SQL

- In this segment we performed various sql queries as we loaded data into the IBM DB2 Database
- We queried the data using python
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes



EDA with SQL (results)

Display the names of the unique launch sites in the space mission

```
In [14]: %sql select distinct launch_site from SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

EDA with SQL (results)

Display 5 records where launch sites begin with the string 'CCA'

In [38]:
```
%sql select * from SPACEXTBL where launch_site like "CCA%" limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[38]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

EDA with SQL (results)

Display the total payload mass carried by boosters launched by NASA (CRS)

In [32]:
```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload from SPACEXTBL where Customer like "NASA%";
```

```
 * sqlite:///my_data1.db
Done.
```

Out[32]:

total_payload
99980.0

EDA with SQL (results)

Display average payload mass carried by booster version F9 v1.1

In [35]:
```
%sql select avg(PAYLOAD_MASS__KG_) as avg_payload from SPACEXTBL where Booster_Version like "F9 v1.1%";
```

```
 * sqlite:///my_data1.db
Done.
```

Out[35]:

avg_payload
2534.6666666666665

EDA with SQL (results)

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [48]:
```
%% sql select booster_version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000
and 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[48]:

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

EDA with SQL (results)

List the total number of successful and failure mission outcomes

In [58]:
```
%sql select Mission_Outcome, count(*) as total_number from SPACEXTBL group by Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[58]:

Mission_Outcome	total_number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

EDA with SQL (results)

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

In [60]:
```
%sql select distinct booster_version from SPACEXTBL  where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

Out[60]:

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

EDA with SQL (results)

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

In [67]:
```
%%sql

SELECT
  substr(Date, 4, 2) AS Month,
  Landing_Outcome AS Failure_Landing_Outcomes_In_Drone_Ship,
  Booster_Version,
  Launch_Site
FROM
  SPACEXTBL
WHERE
  substr(Date, 7, 4) = '2015'
  AND Landing_Outcome = 'Failure (drone ship)';
```

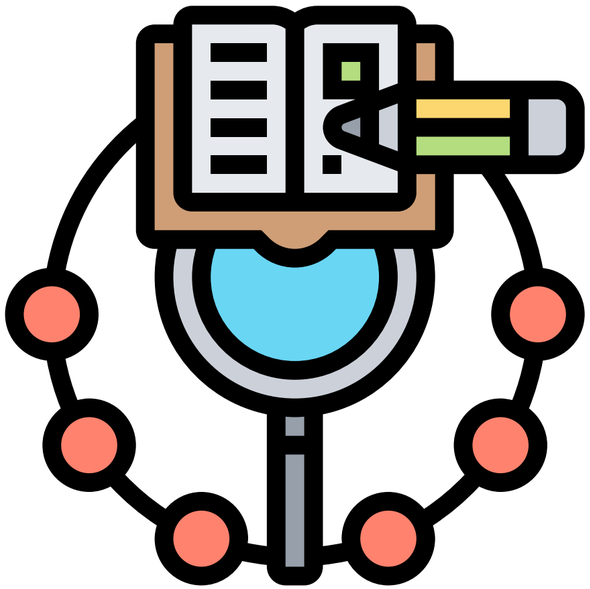
```
 * sqlite:///my_data1.db
Done.
```

Out[67]:

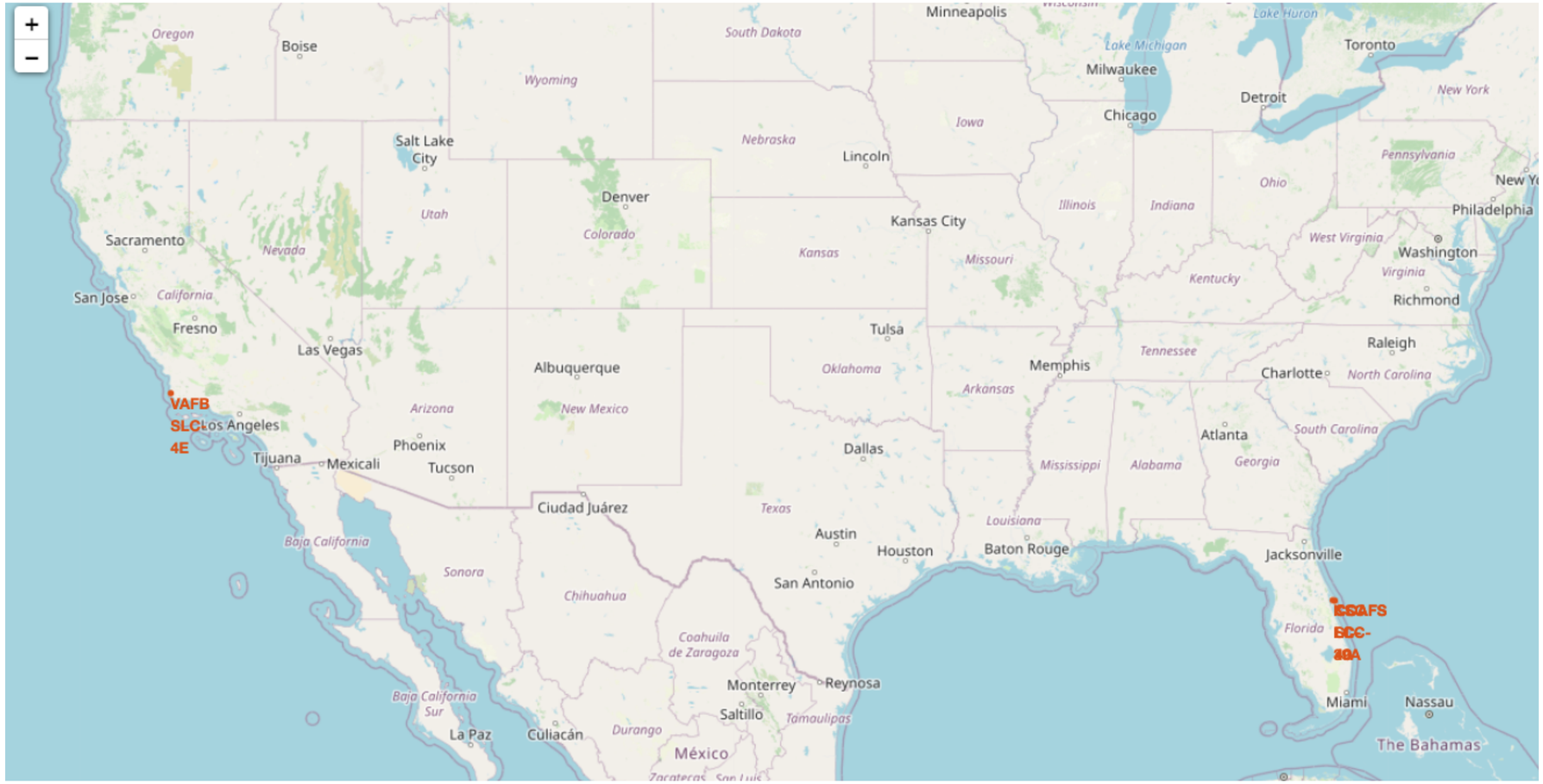
Month	Failure_Landing_Outcomes_In_Drone_Ship	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Build an interactive map with Folium

- Creates a map that shows the locations of all launch sites.
- The markers are colored green for successful launches and red for failed launches.
- Add lines to show the distances between the launch sites and their proximities, such as railways, highways, coastlines, and closest cities.
- This information can be used to identify which launch sites have relatively high success rates and to understand the factors that may contribute to launch success or failure.

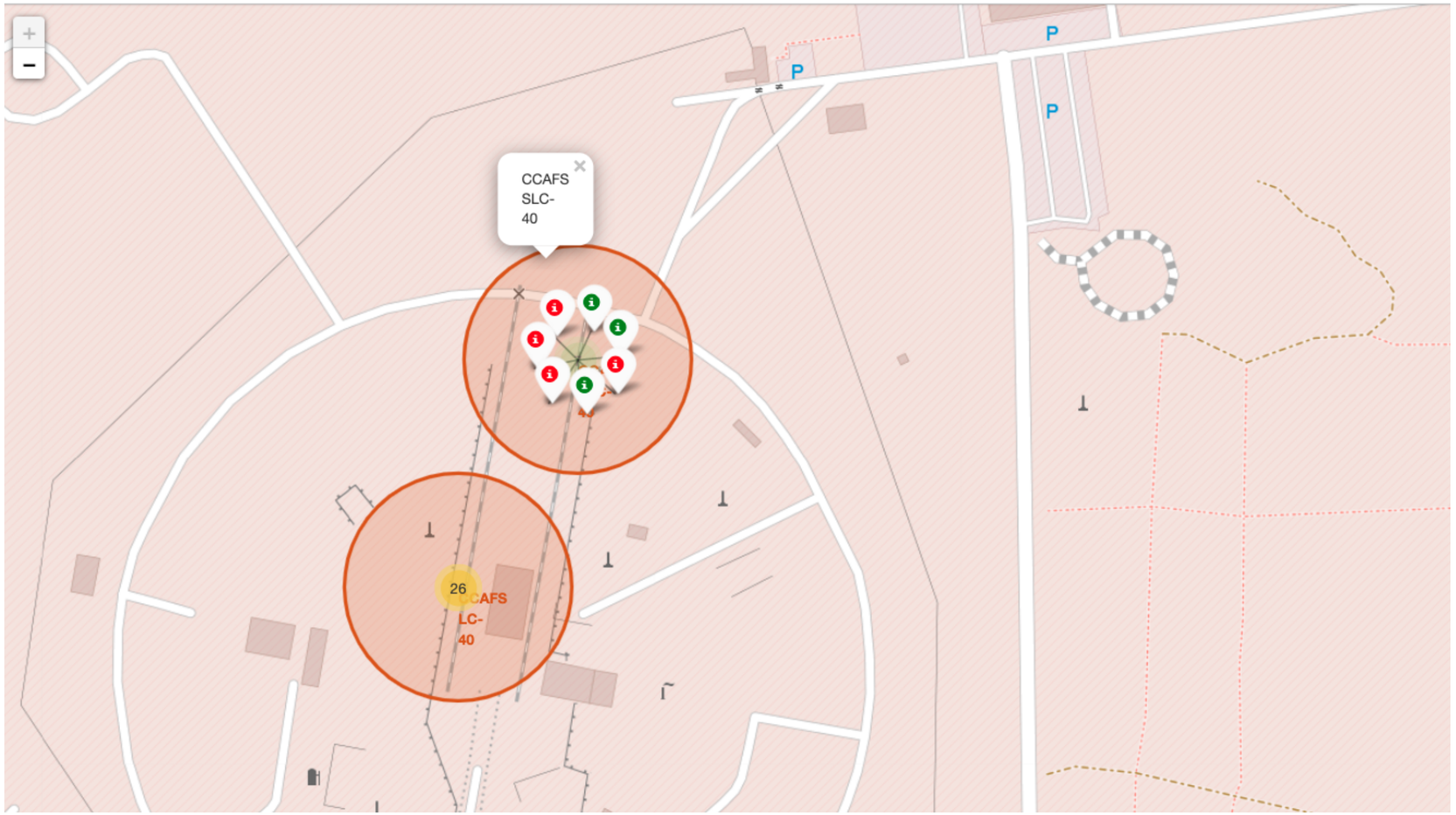


Interactive map (results)



- Most of the launch site are close to the equator line.
- All launch site are close proximity to the coast, in case of engine failure and explosion, the damnage will be reduced.

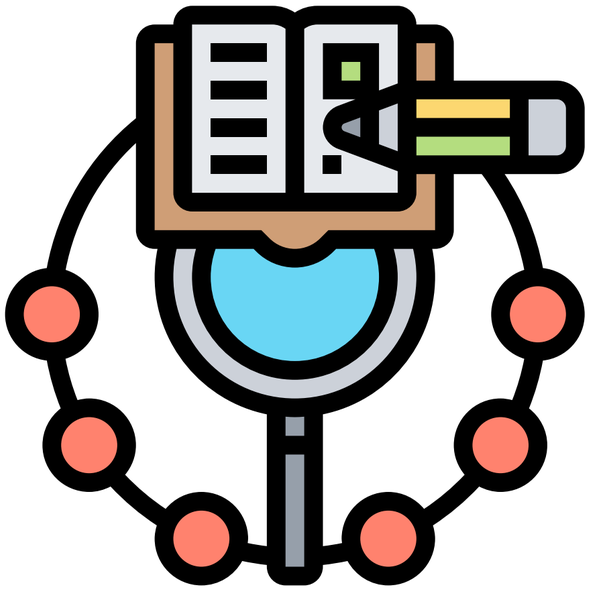
Interactive map (results)



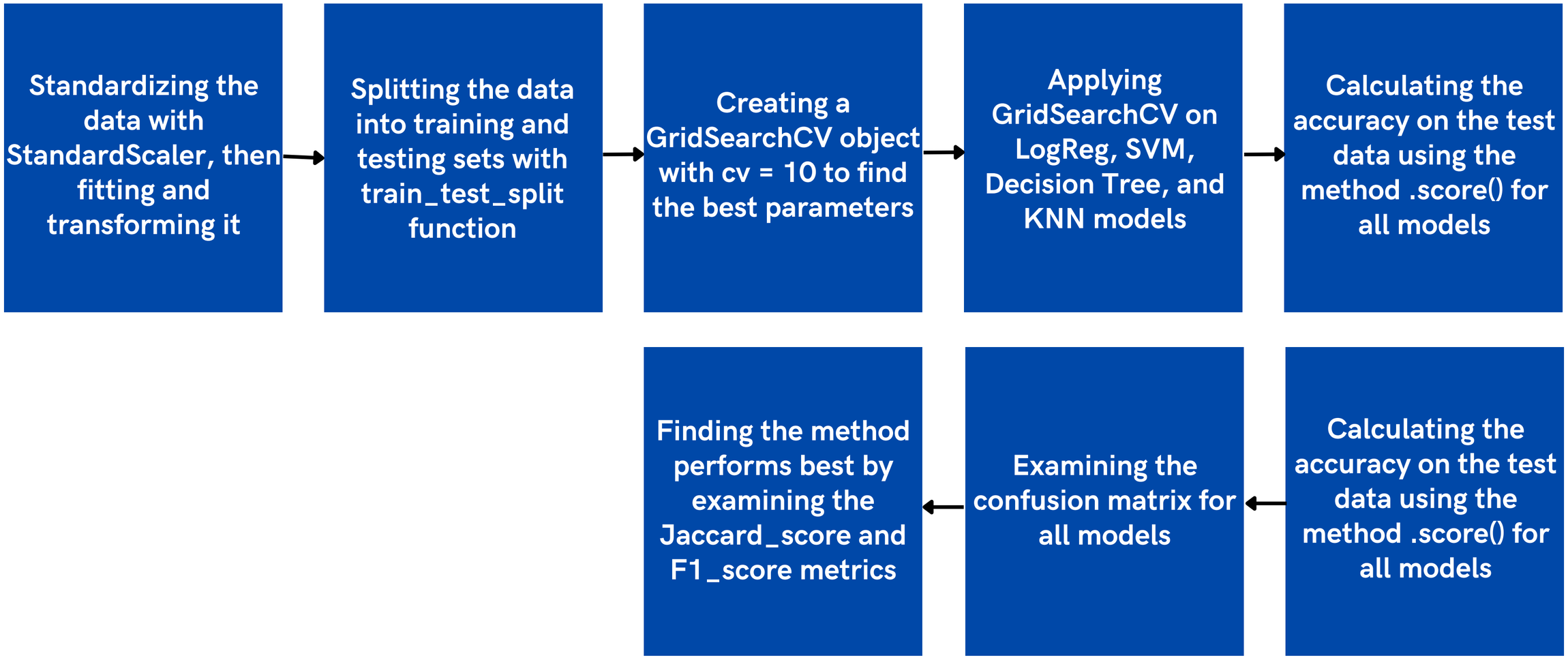
- Green Marker = Successful Launch
- Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - This feature allows users to select a specific launch site from a list.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - This feature displays a pie chart that shows the total number of successful launches for all sites and the success vs. failed counts for a specific site, if selected.
- Slider of Payload Mass Range:
  - This feature allows users to select a specific range of payload masses.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
  - This feature displays a scatter chart that shows the correlation between payload mass and launch success rate for different booster versions.



Predictive analysis



Predictive analysis (results)

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.846154	0.800000
F1_Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

- Since Decision Tree has the highest F1 Score, it means that it is the best model

Conclusion

- Decision Tree is the best algorithm to predict this dataset
- Launches with a low payload mass are more successful than launches with a larger payload mass.
- Most launch sites are located near the equator and close to the coast.
- The success rate of launches has increased over the years.
- KSC LC-39A has the highest success rate of all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

Thanks