IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Francisco Lodi B.
April 10, 2023

Outline

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

Executive Summary

- Summary of methodologies and results
  - The methodologies used mainly are web scraping with API, JSON, Request and BeautifulSoup, then the treatment and cleaning of data in Pandas and basic Exploratory Analysis with MySql and scatter graphs with Pandas. The visualization of the main indicators was carried out with Plotly Dash and the mapping with Folium. Finally, a predictive analysis was made using the Sklearn library, addressing SVM models, classification trees and logistic regression, dividing the data into training and test.
  - All the machine learning models had good results over 80% accuracy, the logistic regression model stands out slightly more with 84% accuracy.
  - Launch sites have different success rates. CCAFS LC-40 has a 60% success rate, while KSC LC-39A and VAFB SLC 4E have a 77% success rate.
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

Introduction

SpaceX company focused on rocket launches, only Falcon 9 has a cost of 62 million dollars; other vendors cost more than $165 million each, much of the savings due to SpaceX being able to reuse the first stage. If it is possible to determine whether the first stage will land, the cost of one launch can be determined. This information can be used if an alternative company wants to bid against SpaceX for a rocket launch.

This presentation summarizes the main results of the exploratory analysis of SpaceX's Falcon 9 rocket launches, seeking to predict which are the conditions that allow the greatest number of successful launches to be generated.

What is the causality of each of the characteristics, being able to determine the significance of the launch site, weight of the load and engineering characteristics.

Section 1

Methodology

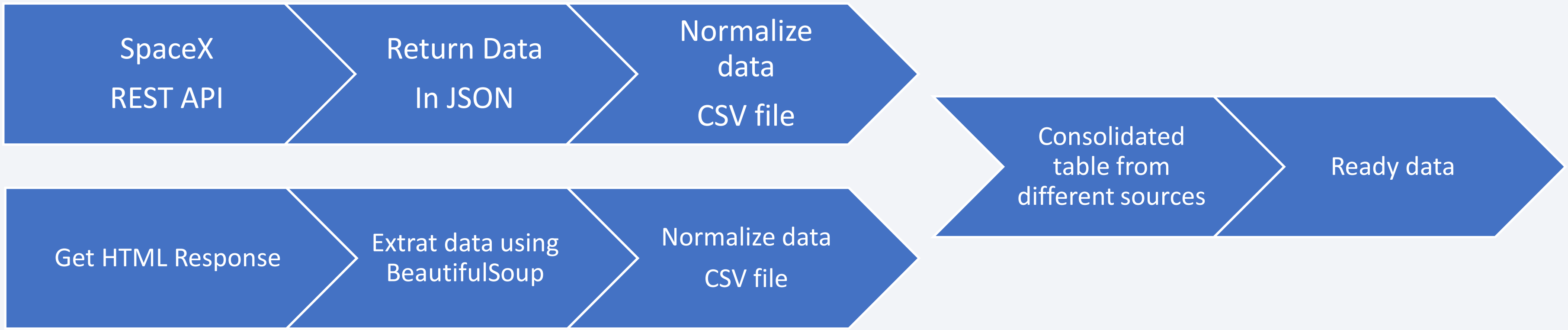
Methodology

Executive Summary

- Data collection methodology:
  - Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Data cleaning of null values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Regression Logistic, SMV, KNN and Decision Tree

Data Collection

- SpaceX launch data that is collected from the SpaceX Rest API.
- This API provides data about launches, including information about the rocket used, the payload delivered, the launch specifications, the landing specifications, and the landing outcome.
- The API starts with api.spacexdata.com/v4/.
- Another source of data for getting Falcon 9 is a web scaraping via BeautifulSoup.



Data Collection – SpaceX API

- GitHub URL: https://github.com/franciscolodi/testrepo/blob/67c1a15d453a5861965458d464084ebbae4a1401/jupyter-labs-spacex-data-collection-api.ipynb

```
# Takes the dataset and uses the rocket column to call the API and append the data to the
def getBoosterVersion(data):
    for x in data['rocket']:
       if x:
        response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
        BoosterVersion.append(response['name'])
```

object for this project.

```
[13]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-D
```

We should see that the request was successfull with the 200 status response code

```
[14]: response.status_code
```

```
[14]: 200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
[17]: # Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

Data Collection - Scraping

- GitHub URL: https://github.com/franciscolodi/testrepo/blob/67c1a15d453a5861965458d464084ebbae4a1401/jupyter-labs-spacex-data-collection-api.ipynb

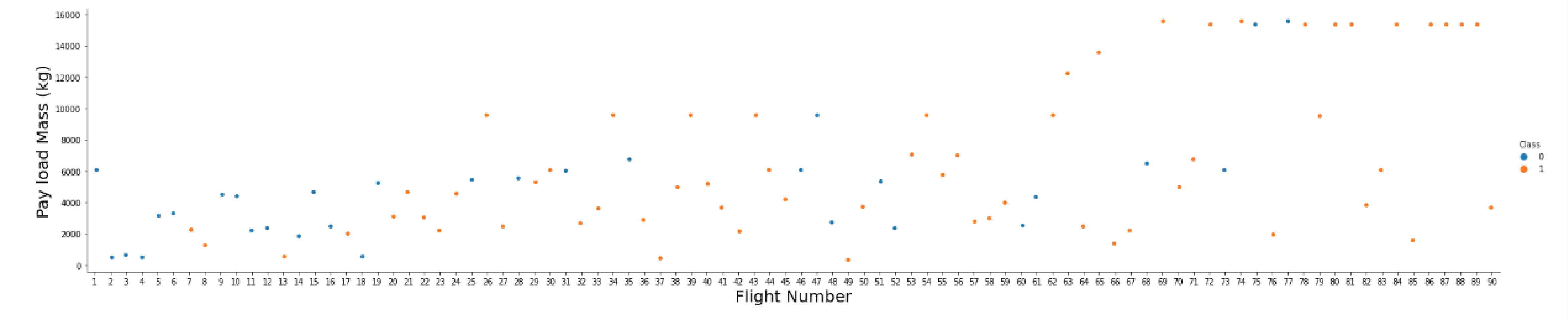
```
]: from bs4 import BeautifulSoup
   url = 'https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches'
   response = requests.get(url)
   soup = BeautifulSoup(response.text, 'html.parser')
   soup.title

]: <title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

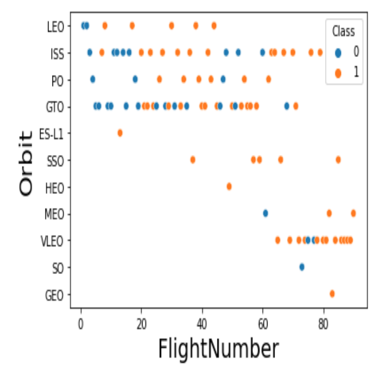
Data Wrangling

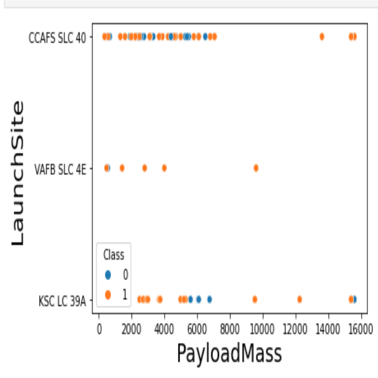
- GitHub URL:

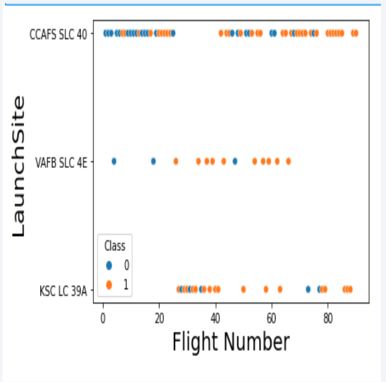
- https://github.com/franciscolodi/testrepo/blob/67c1a15d453a5861965458d464084ebbae4a1401/jupyter-labs-spacex-data-collection-api.ipynb

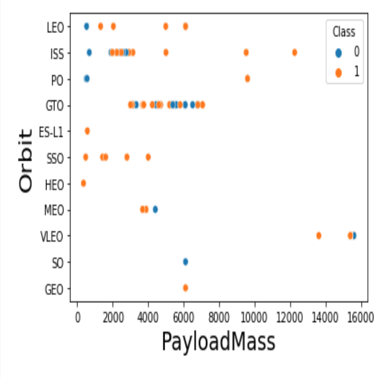
EDA with Data Visualization

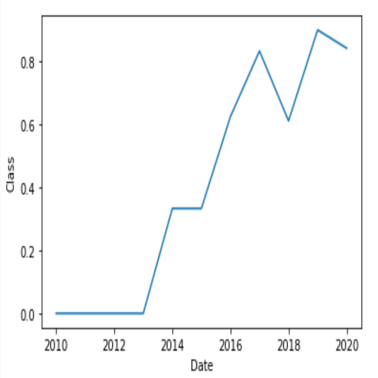


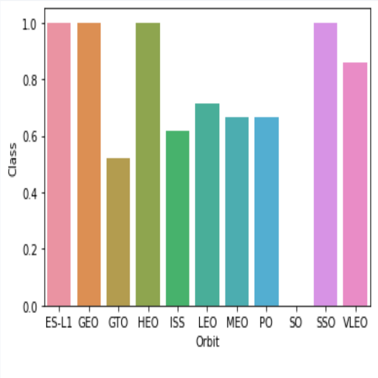












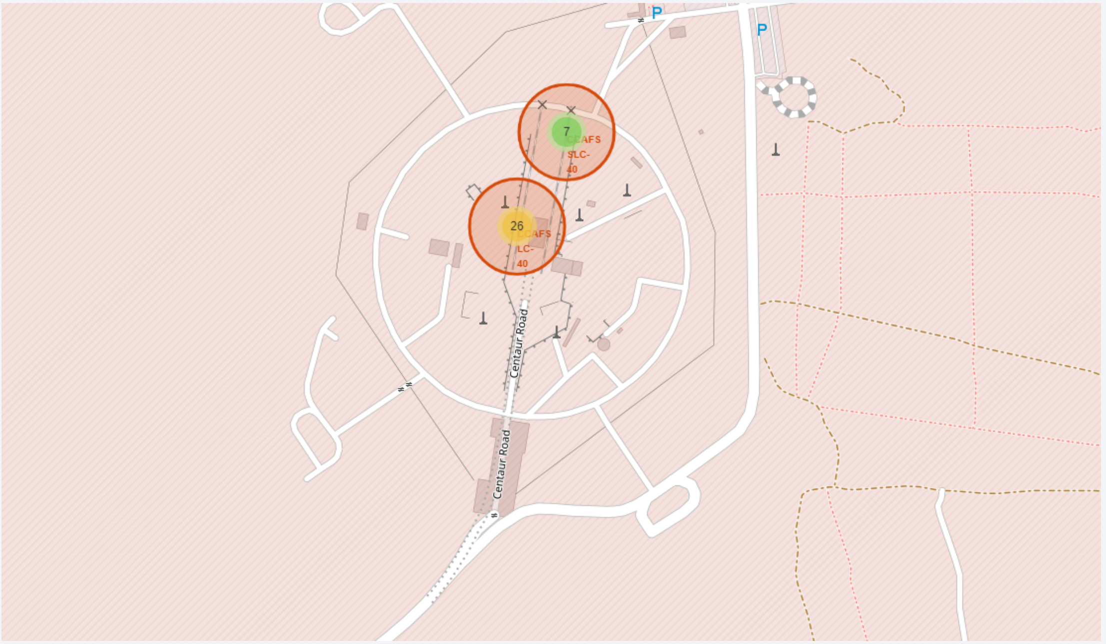
- GitHub URL:
- https://github.com/franciscolodi/testrepo/blob/2ea7940457ec00ca67e380046d162a86add4ac5c/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Displaying the names of the unique launch sites in the space mission.
- Displaying 5 records where launch sites begin with the string ‘KSC.
- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad booster versions, launch_site for the months in year 2017.
- Ranking the count of successful landing _outcomes between the date 2010 06 04 and 2017 03 20 in descendingorder.

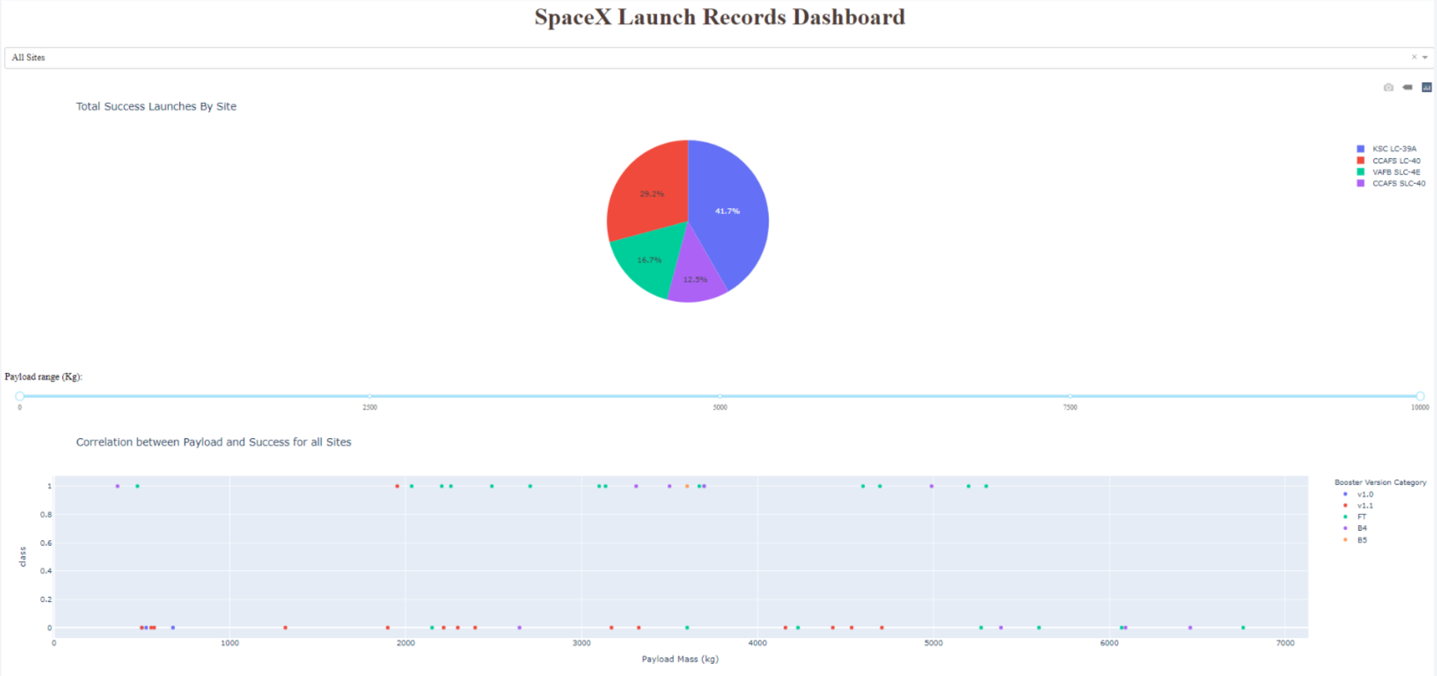
GitHub URL:
https://github.com/franciscolodi/testrepo/blob/6afc9ff3cff0fdffb0cff7f818cc6a8741ca8f9d/Copia_de_DB0201EN_PeerAssign_v5_SQLite.ipynb

Build an Interactive Map with Folium



- Map markers have been added to the map with the goal of finding an optimal location for successful launches.
- GitHub URL: https://github.com/franciscolodi/testrepo/blob/dbf05425ed3eda7e38e30a2bcc7b921b1635a2c0/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

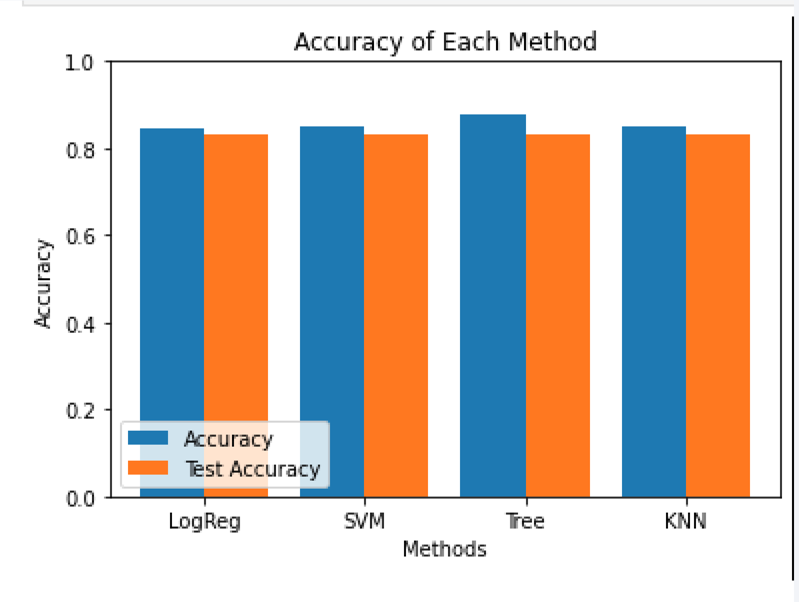


A Plotly Dash app was built for users to perform interactive visual analysis on SpaceX launch data on real time

GitHub URL:
https://github.com/franciscolodi/testrepo/blob/dbf05425ed3eda7e38e30a2bcc7b921b1635a2c0/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

All the machine learning models had good results over 80% accuracy, the logistic regression model stands out slightly more with 84% accuracy.



GitHub URL:

- https://github.com/franciscolodi/testrepo/blob/86ffb21084ff90f82d0be68403dec0946f9ba296/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

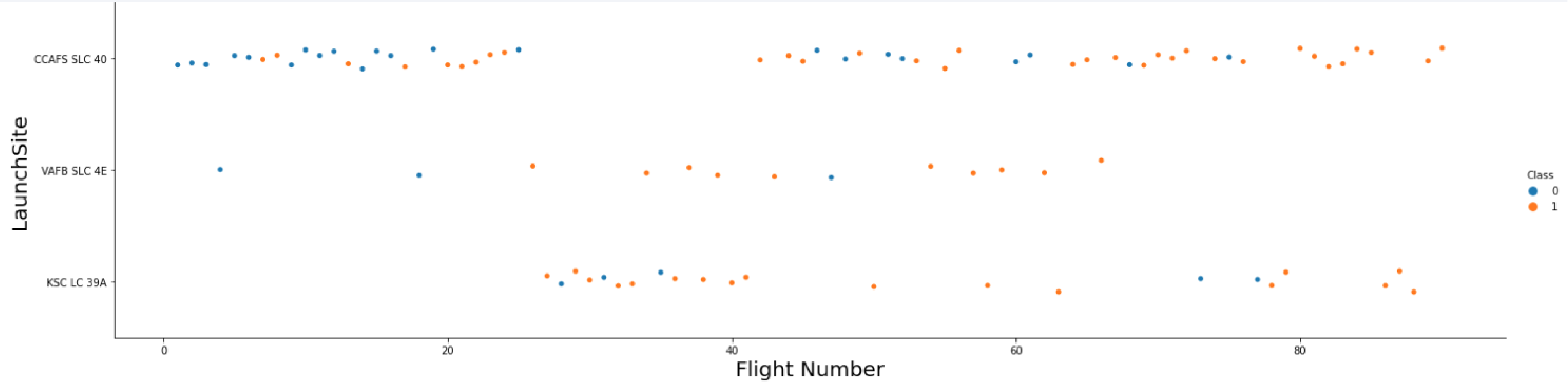
Results

- The SVM, KNN, DT and Logistic Regression models give good results in terms of prediction accuracy for this data set.
- Low weight payload performs better than heavier payloads.
- SpaceX lunch success rates are directly proportional to the time in years that they will eventually perfect the lunches.
- KSC LC 39A had the most successful launches of all the sites.
- Orbit GEO, HEO, SSO, ES L1 has the best success rate.

Section 2

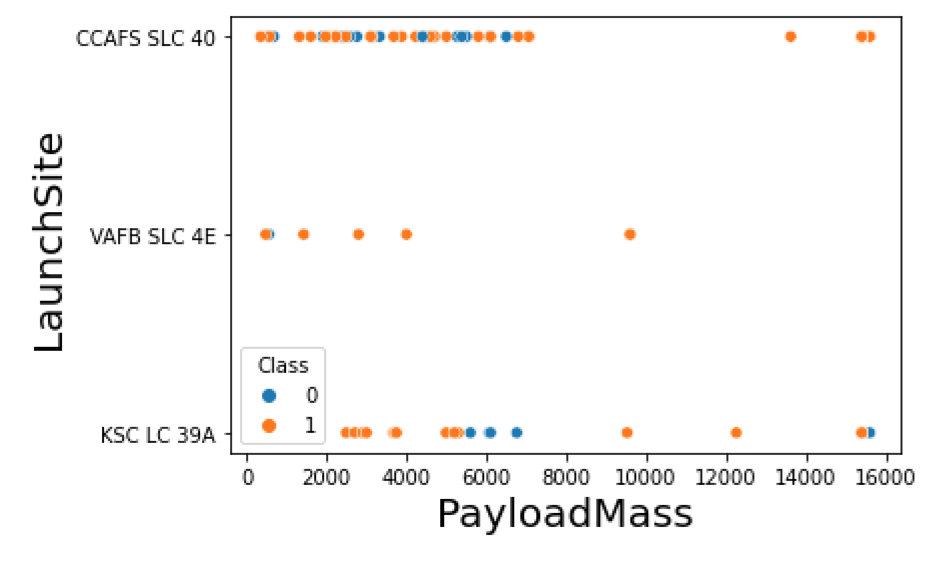
Insights drawn from EDA

Flight Number vs. Launch Site



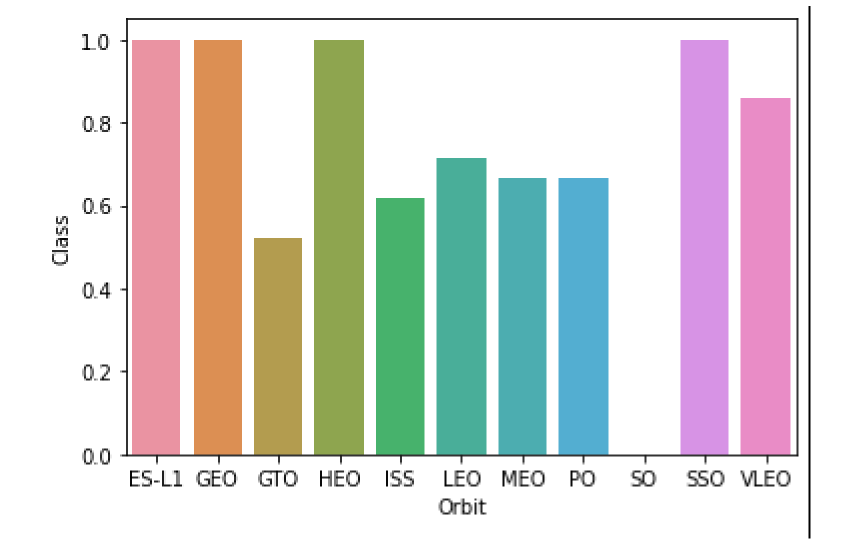
- VAFB SLC 4E has a higher success rate
- Launches from the site of CCAFS SCL 40 are significantly higher than launches form other sites.

Payload vs. Launch Site



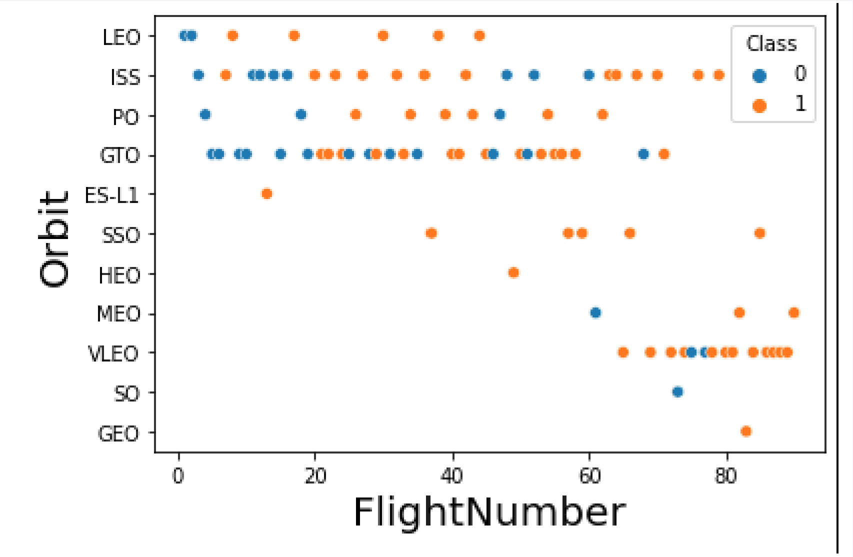
- Payload vs. The launch site scatter point chart you will find for the VAFB-SLC launch site no rockets are launched for a heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type



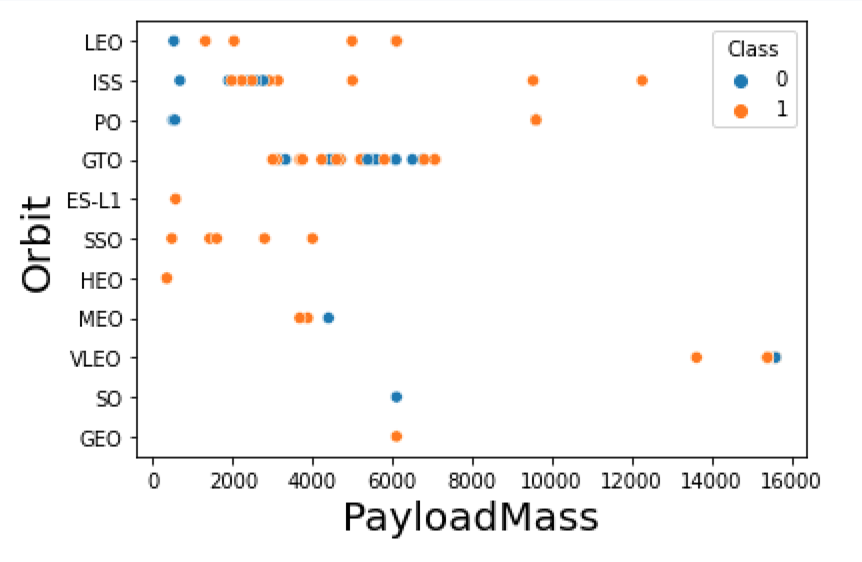
- ES-L1, GEO, HEO, SSO are the highest success rate.

Flight Number vs. Orbit Type



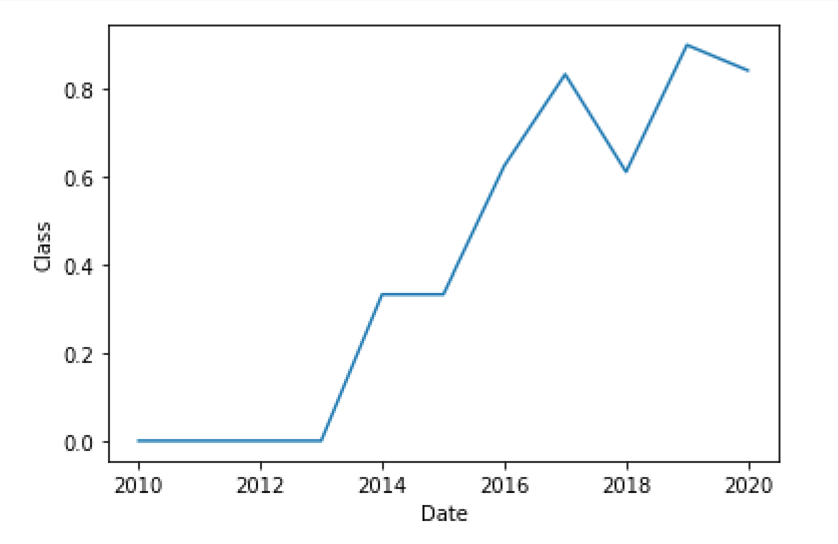
- LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Po, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- We can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
#1 Display the names of the unique launch sites in the space mission
%sql select distinct  Launch_Site FROM SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

```
#2 Display 5 records where launch sites begin with the string 'CCA'
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE '%CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
#3 Display the total payload mass carried by boosters launched by NASA (CRS)
%sql SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
#Display average payload mass carried by booster version F9 v1.1
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
#5List the date when the first succesful landing outcome in ground pad was acheived.
%sql SELECT Date FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)'
```

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

Date
22-12-2015
18-07-2016
19-02-2017
01-05-2017
03-06-2017
14-08-2017
07-09-2017
15-12-2017
08-01-2018

Successful Drone Ship Landing with Payload between 4000 and 6000

```
#6List 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 from SPACEXTBL WHERE "Landing _Outcome"='Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND "PAYLOAD_MASS__KG_" < 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
#7List the total number of successful and failure mission outcomes
%sql select Mission_Outcome, count('Mission_Outcome') from SPACEXTBL GROUP BY "Mission_Outcome"
```

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

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

Boosters Carried Maximum Payload

```
#8List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
%sql select Booster_Version from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)  from SPACEXTBL)
```

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

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

2015 Launch Records

```
#9List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
#Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

%sql select Date, Booster_Version, Launch_Site FROM SPACEXTBL WHERE "Landing _Outcome"='Failure (drone ship)' AND substr(Date,7,4)='2015'
```

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

Date	Booster_Version	Launch_Site
10-01-2015	F9 v1.1 B1012	CCAFS LC-40
14-04-2015	F9 v1.1 B1015	CCAFS LC-40

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

```
#10Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
%sql select * FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%' AND substr(Date,7,4)>'2010' AND substr(Date,7,4)<'2017'
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
08-04-2016	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
27-05-2016	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)

Section 3

Launch Sites Proximities Analysis

Rocket launch sites CCAFS LC-40 and CCAFS SLC-40

Rocket launch sites CCAFS LC-40 and CCAFS SLC-40
The color-labeled launch outcomes on the map

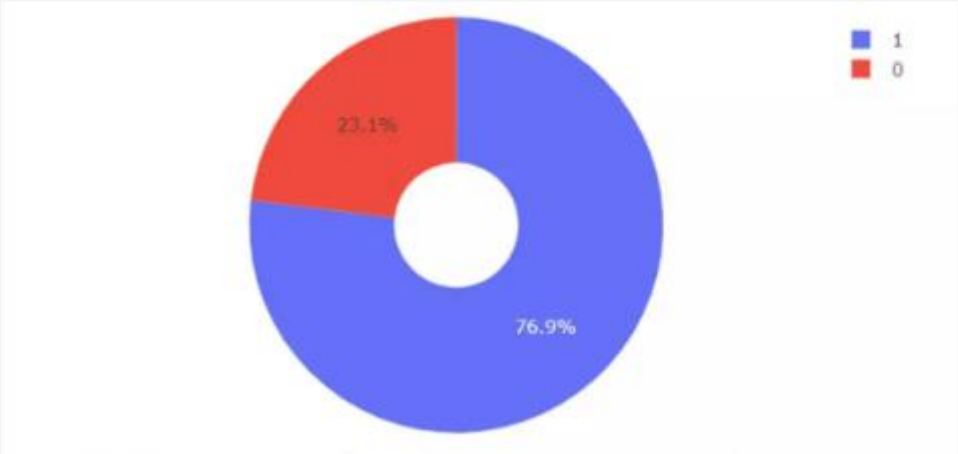
Distance between the coastline point and the launch site

Section 4

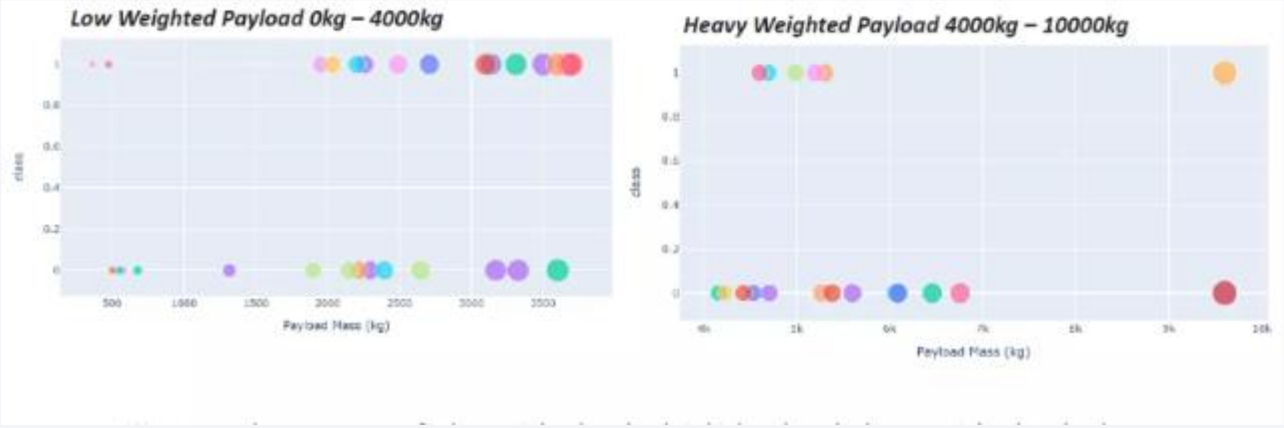
Build a Dashboard with Plotly Dash

Total Success Launches by all sites

Success rate by site



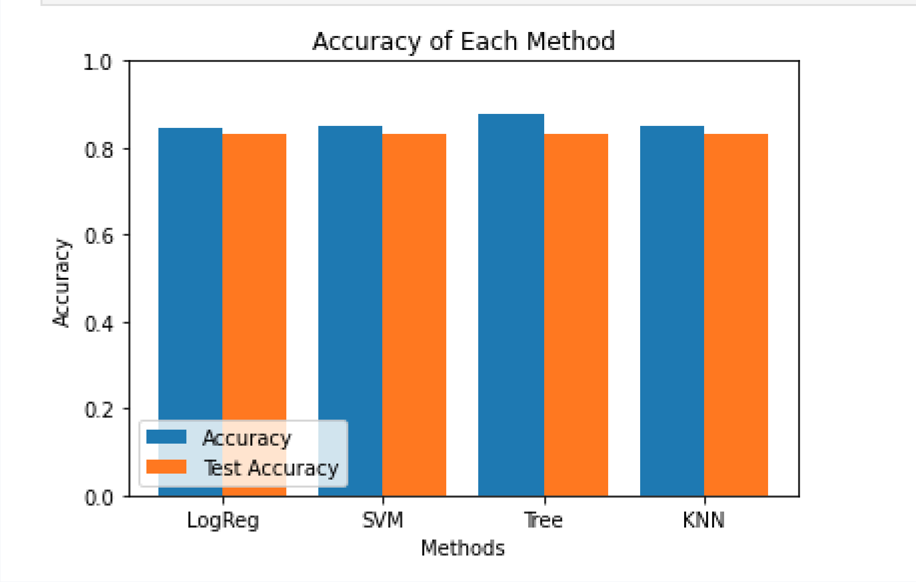
Payload vs launch outcome



Section 5

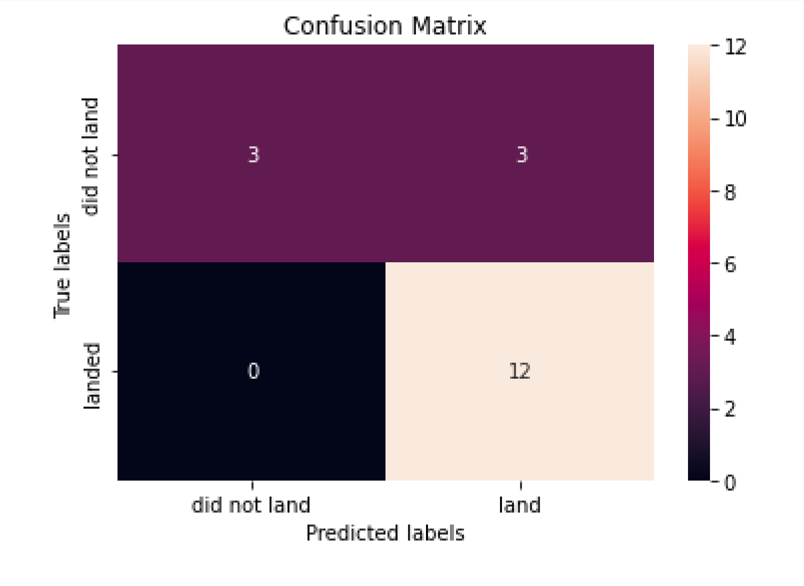
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

In the logistic regression confusion matrix, the twelve successful landings that were predicted are true, it only fails to predict 3 successful landings that are not.



Conclusions

The work presented was developed mainly in Jupyter Lab and Google collaborated with data extracted from NASA, the libraries used are Pandas, Sklearn, Seaborn, Numpy, Matplotlib, Folium, Sqlite3 and Ploly.

Under the proposed models the main results were

- The SVM, KNN, DT and Logistic Regression models give good results in terms of prediction accuracy for this data set.
- Low weight payload performs better than heavier payloads.
- SpaceX lunch success rates are directly proportional to the time in years that they will eventually perfect the lunches.
- KSC LC 39A had the most successful launches of all the sites.
- Orbit GEO, HEO, SSO, ES L1 has the best success rate.

Finally, it would be interesting to continue modifying the parameters of the machine learning models to see how the model varies and its effectiveness.

Thank you!