IBM DATA SCIENCE CAPSTONE PROJECT

Ivan Azpirolea
2026/01/09

Outline

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

Executive Summary

- In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.
- The main steps in this project include:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure.
- It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

METHODOLOGY

Methodology

- Data collection methodology:
  - Gathering data from web APIs and Web Scrapping
- Exploratory data analysis (EDA) and data wrangling:
  - Using Python libraries (Pandas, Numpy) and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

Data Collection – SpaceX API

- The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.

Every missing value in the data is replaced the mean the column that the missing value belongs to. We end up with 90 rows or instances and 17 columns or features.

- Link to Github notebook

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0

Data Collection – Web Scraping

- We gather this data from Wikipedia Web:

The website contains only the data about Falcon 9 launches.

We end up with 121 rows or instances and 11 columns or features.
The picture below shows the first few rows of the data:

- Link to Github notebook

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	1	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0003.18	Failure	4 June 2010	18:45
2	1	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0004.18	No attempt\n	8 December 2010	15:43
3	2	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0005.18	No attempt	22 May 2012	07:44
4	3	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.07B0006.18	No attempt\n	8 October 2012	00:35

Data Wrangling

- We identify and replace, if necessary, missing data, null data, and resolve misassigned data types. An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.

- Link to the notebook.

EDA with Data Visualization

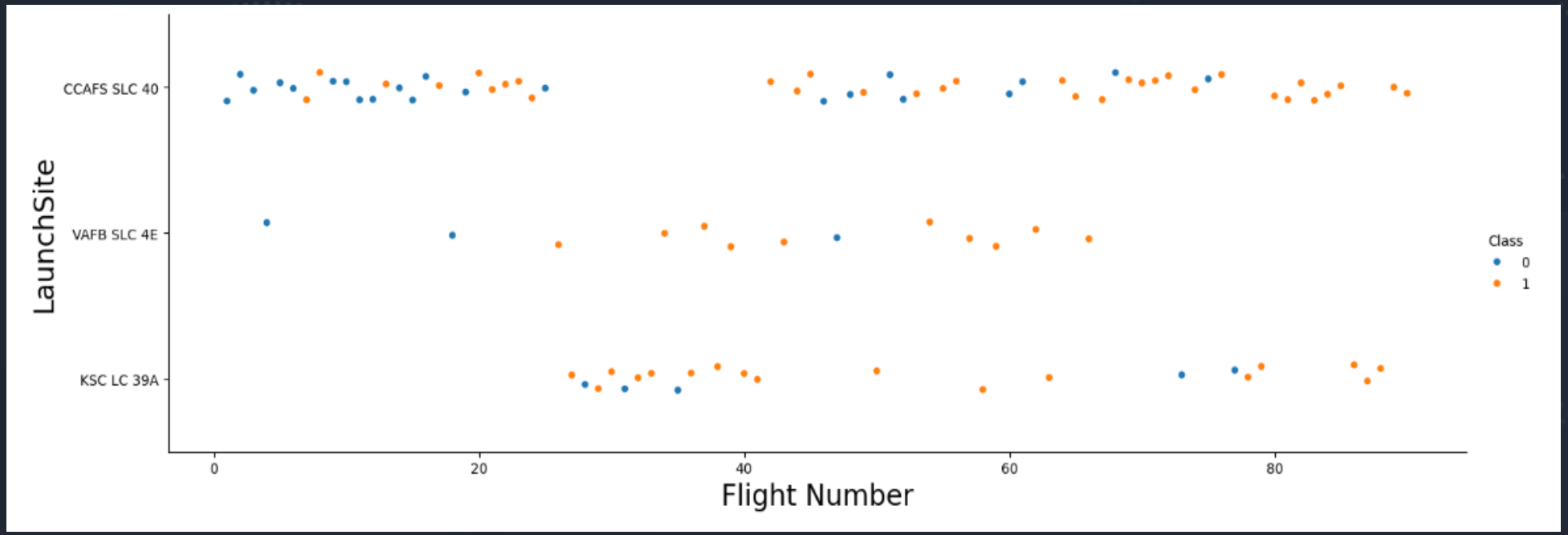
- In this notebook lab we use visualization to perform an exploratory analysis of the data, for:

  - Visualize the relationship between Flight Number and Launch Site
  - Visualize the relationship between Payload Mass and Launch Site
  - Visualize the relationship between success rate of each orbit type
  - Visualize the relationship between Flight Number and Orbit type
  - Visualize the relationship between Payload Mass and Orbit type
  - Visualize the launch success yearly trend

- Link to the notebook

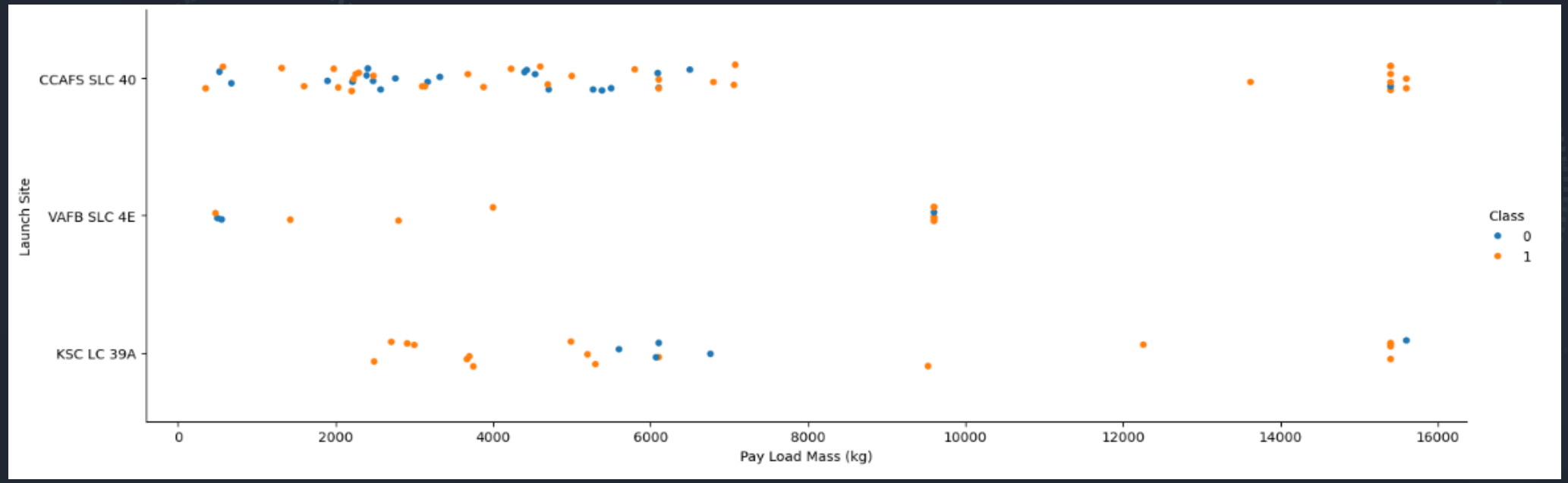
Relationship between Flight Number and Launch Site

- We observed that most of the Launches take place in CCAFS SLC 40 Launch Spot



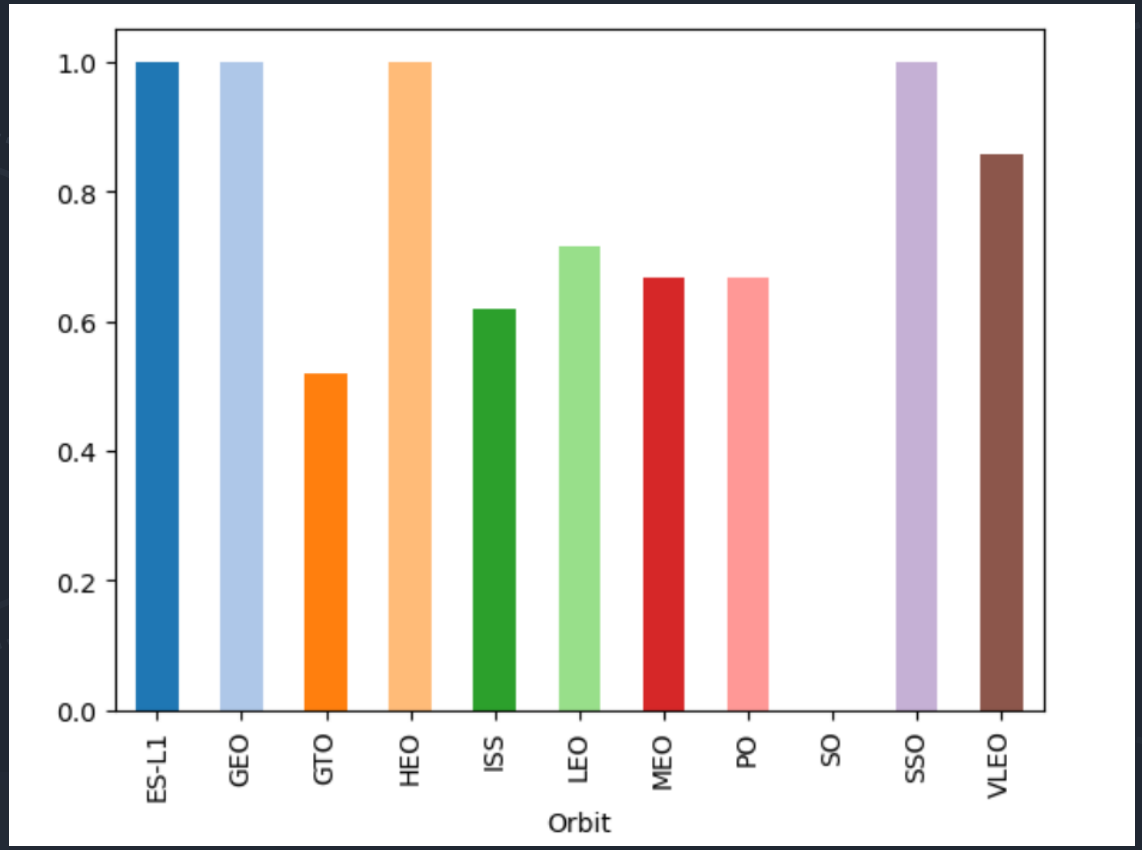
Relationship between Payload Mass and Launch Site

- If we observe this chart we will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).



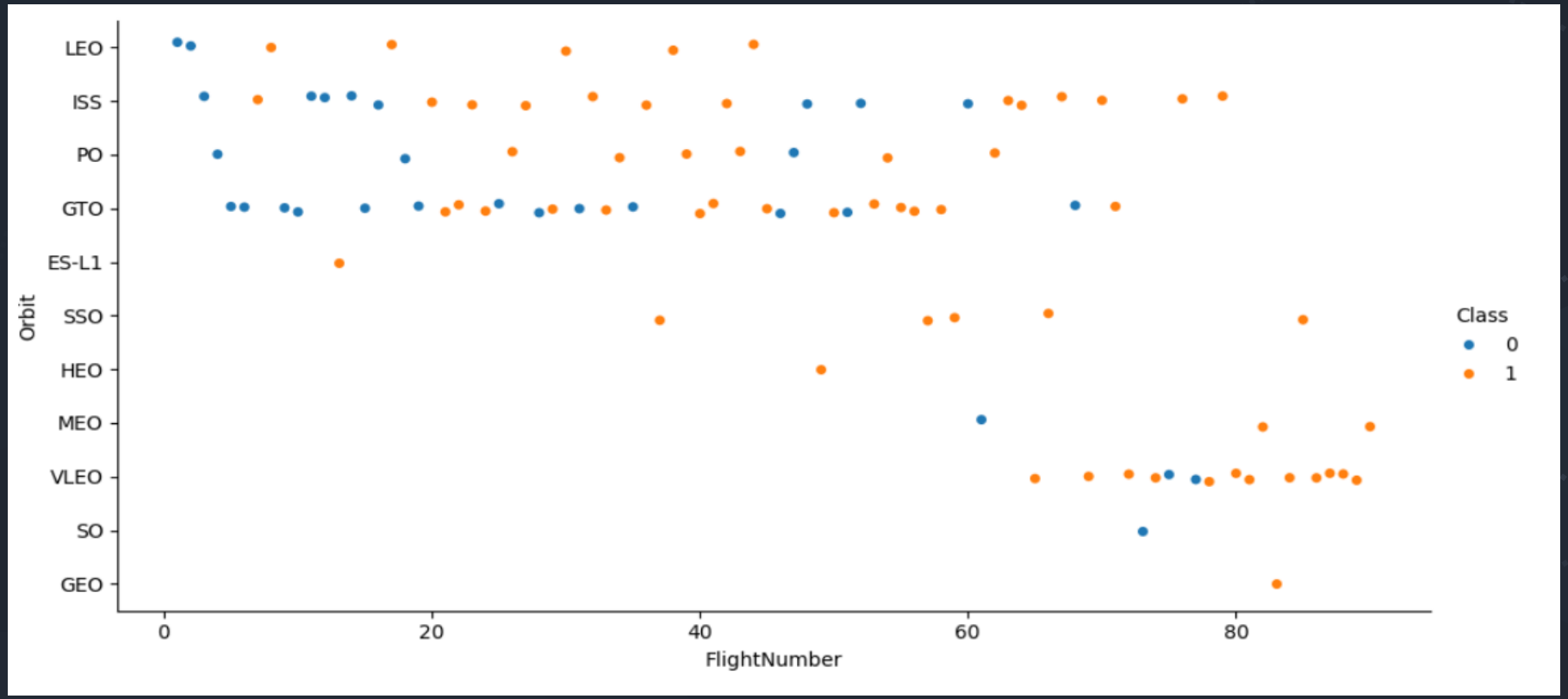
Relationship between Success rate of each Orbit type

- We can observe that ES-L1, GEO, HEO and SSO orbits have the most success rate.



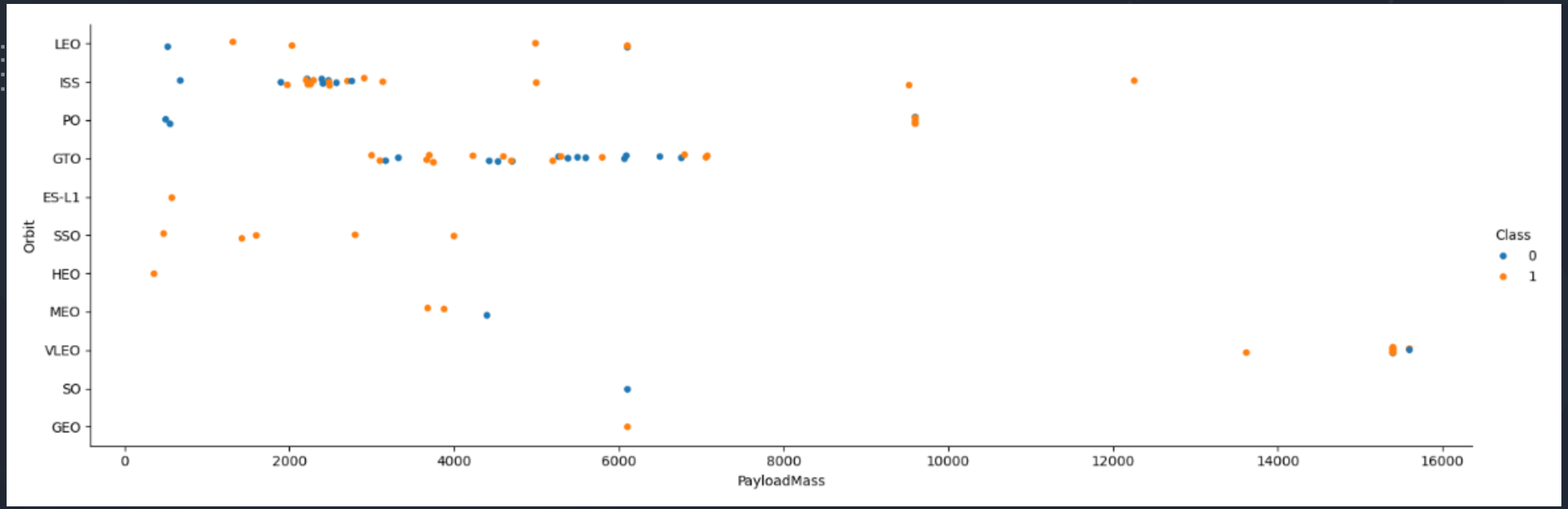
Relationship between Flight number and Orbit type

- You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



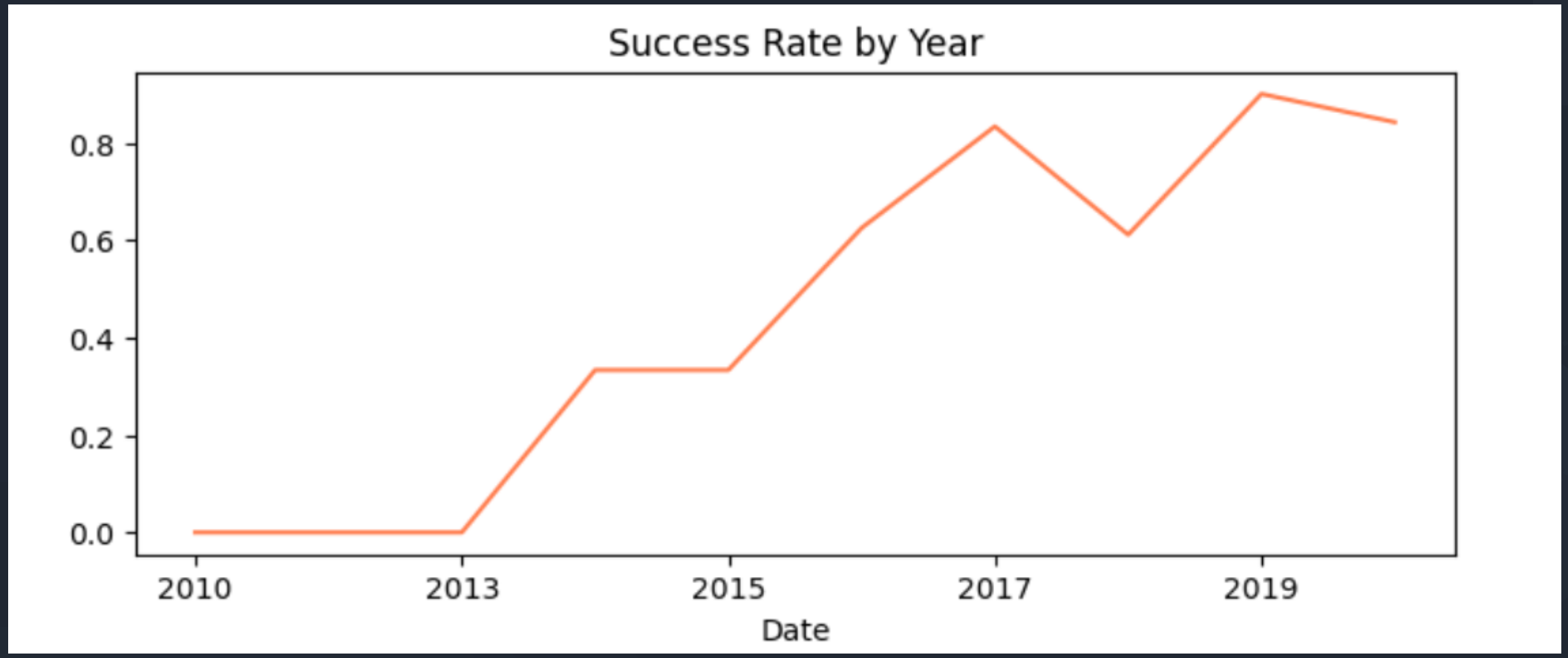
Relationship between Payload Mass and Orbit type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch success yearly trend

- You can observe that the success rate since 2013 kept increasing till 2020



EDA with SQL

- We use SQL queries for complete the following tasks:

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

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

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

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

  5- List the date when the first successful landing outcome in ground pad was achieved.

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

  7- List the total number of successful and failure mission outcomes

  8- List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

  9- 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.

  10- Rank 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.

- Link to the notebook lab.

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

- Names of the unique launch sites in the space mission

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

Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

SUM(PAYLOAD_MASS__KG_)
99980

- Total payload mass carried by boosters launched by NASA (CRS)

- Average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

MIN(Date)
2015-12-22

- Date when the first successful landing outcome in ground pad was achieved.

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Outcomes
101

- Total number of successful and failure mission outcomes

- Boosters that have carried the maximum payload mass

COUNT(*)
12

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

- Records for failed launches in drop ship in months from year 2015.

- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

COUNT(landing_outcome)
31

Build an Interactive Map with Folium

- In this lab, my task was to create markers, indicators, and visualizations, and added to a folium map, to complete the following tasks:

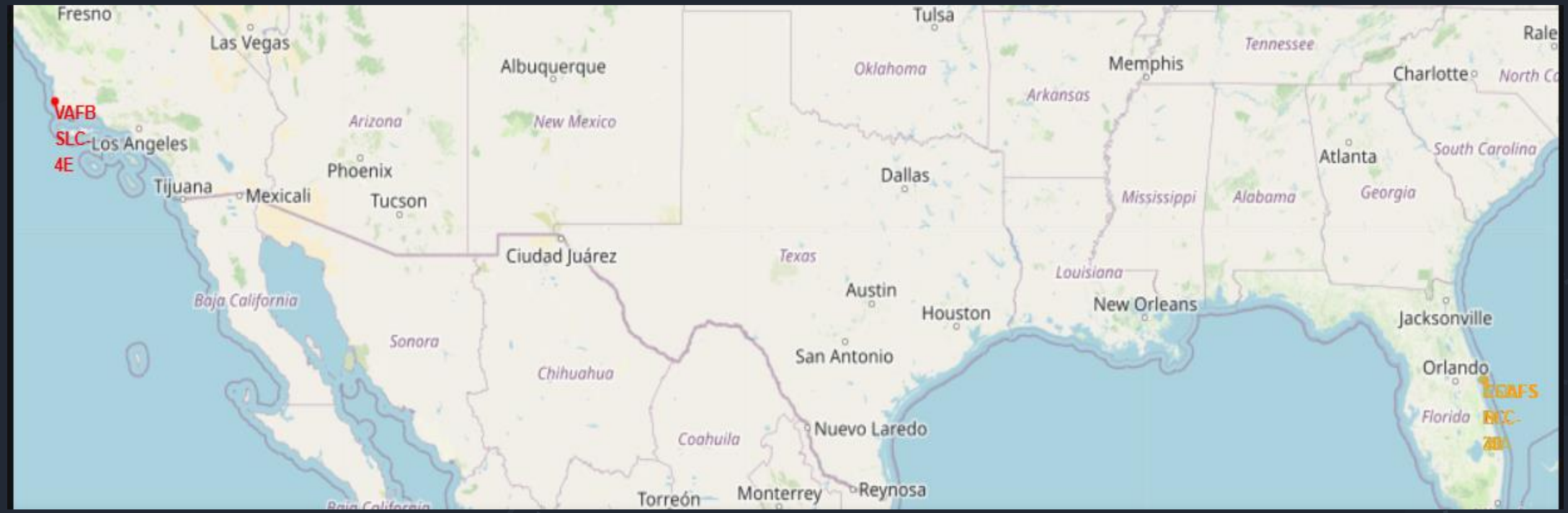
1- Mark all launch sites on a map

2- Mark the success/failed launches for each site on the map

3- Calculate the distances between a launch site to its proximities.

- Link to the notebook.

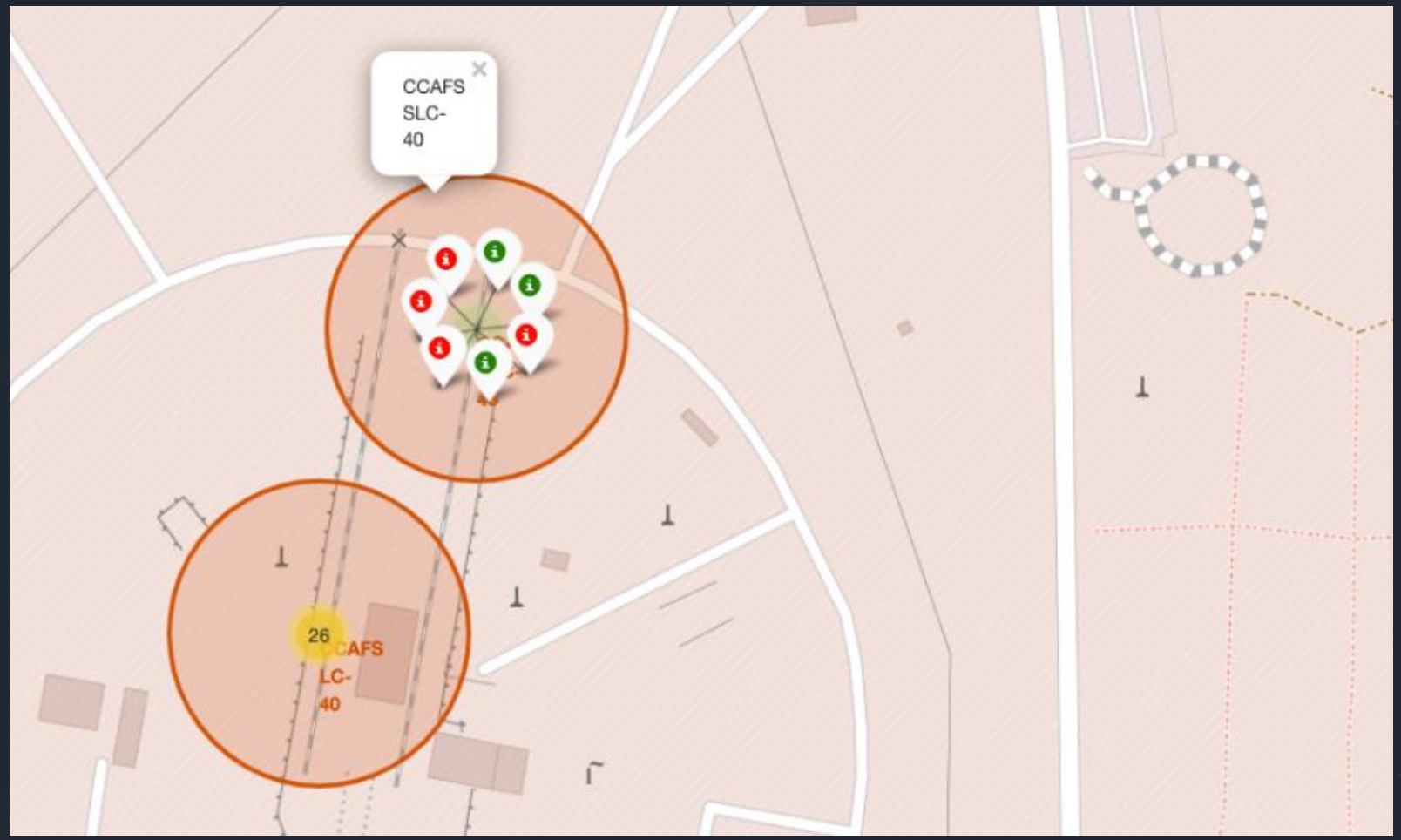
Launch Sites



- We can observe that all launch sites are in very close proximity to the coast

Success/failed launches for each site

- From the color-labeled markers in marker clusters, we should be able to easily identify which launch sites have relatively high success rates.



Distances between a launch site to its proximities.

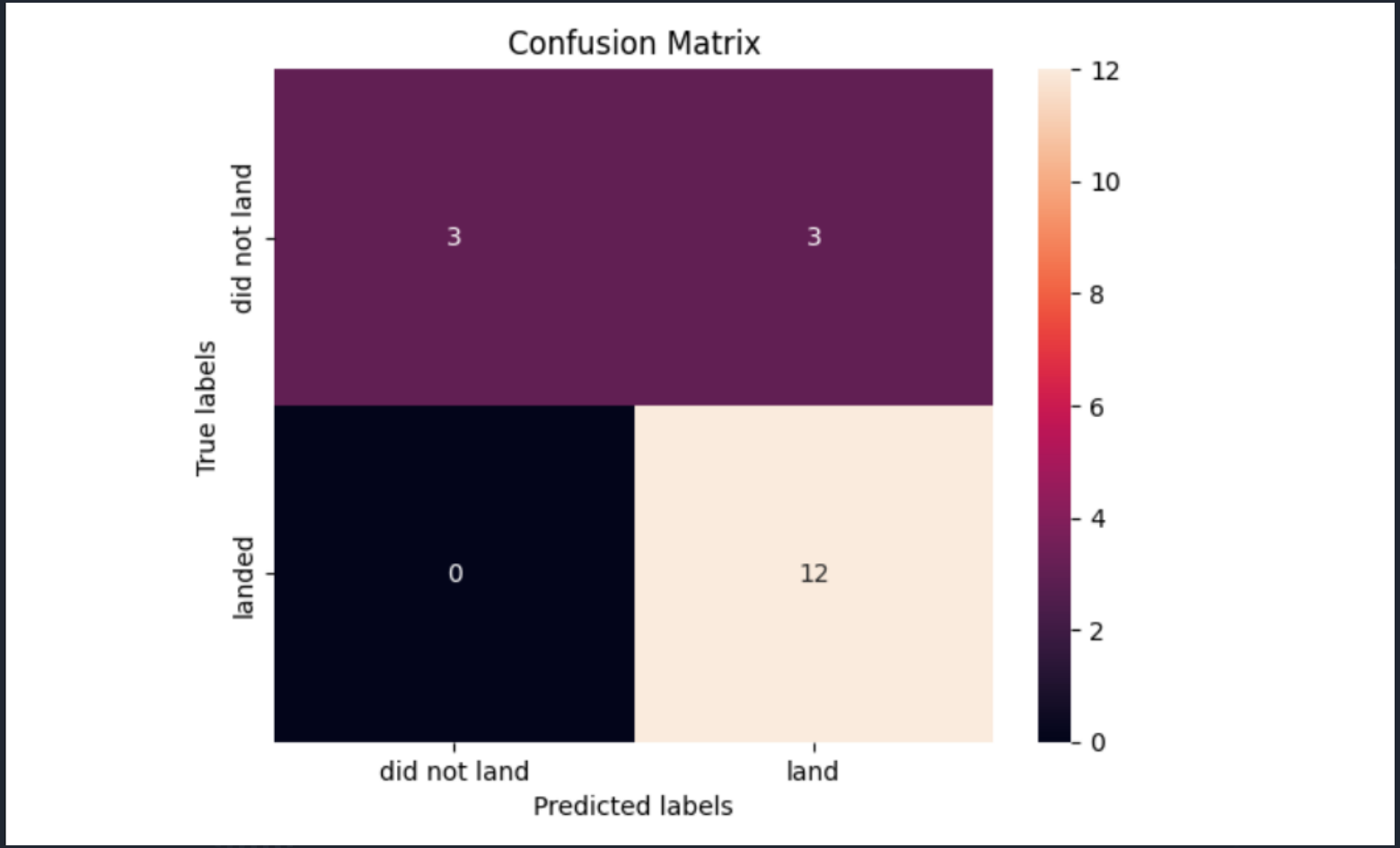
- If we zoom in to a launch site and explore its proximity to see if you can easily find any railway, highway, coastline, etc.

Predictive Analysis (Classification)

Functions from the Scikit-learn library are used to create our machine learning models.

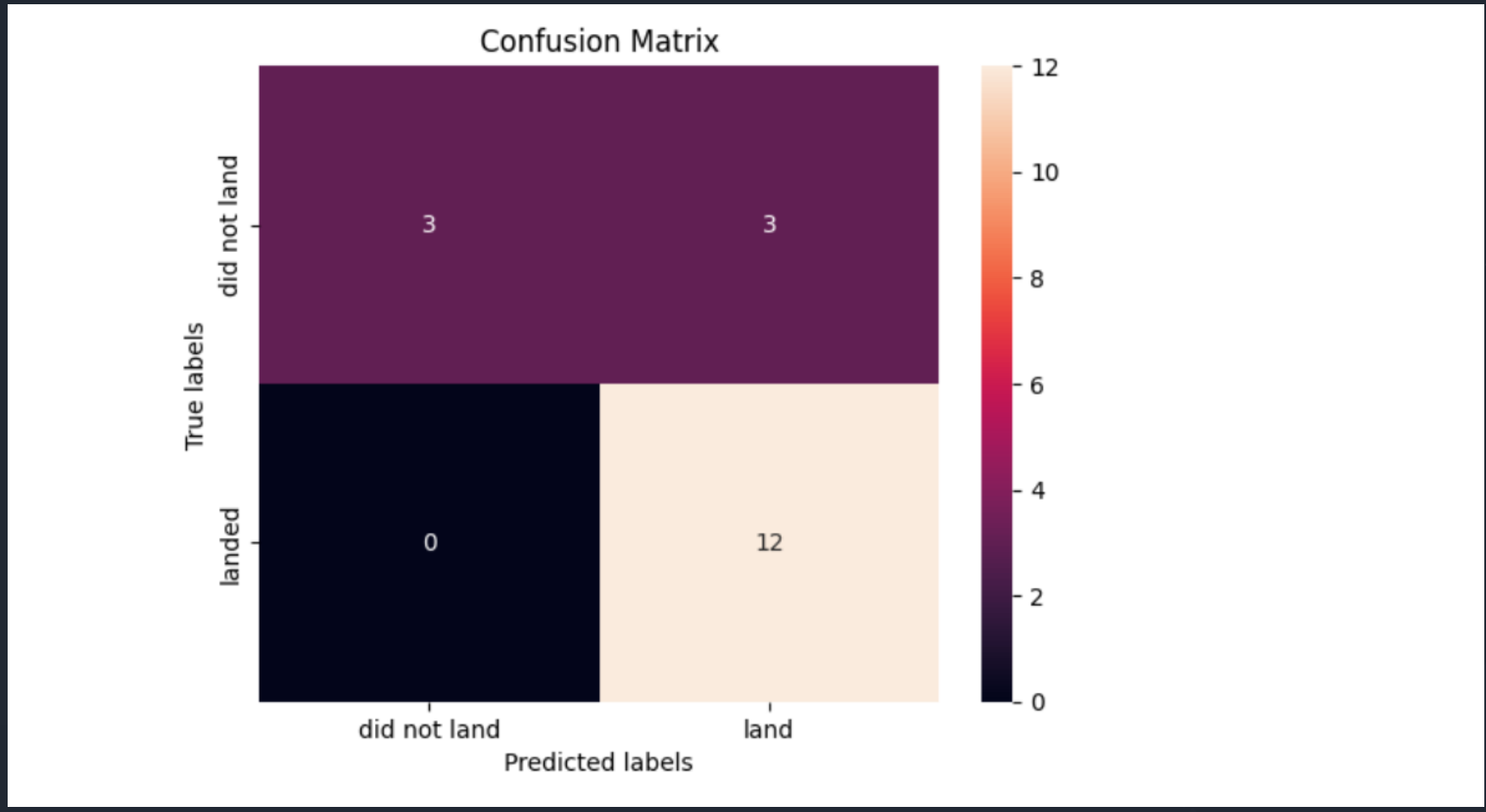
- The machine learning prediction phase include the following steps:
- Standardizing the data
- Splitting the data into training and test data
- Creating machine learning models, which include:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K nearest neighbors (KNN)
- Fit the models on the training set
- Find the best combination of hyperparameters for each model
- Evaluate the models based on their accuracy scores and confusion matrix

Logistic Regression - Confusion Matrix



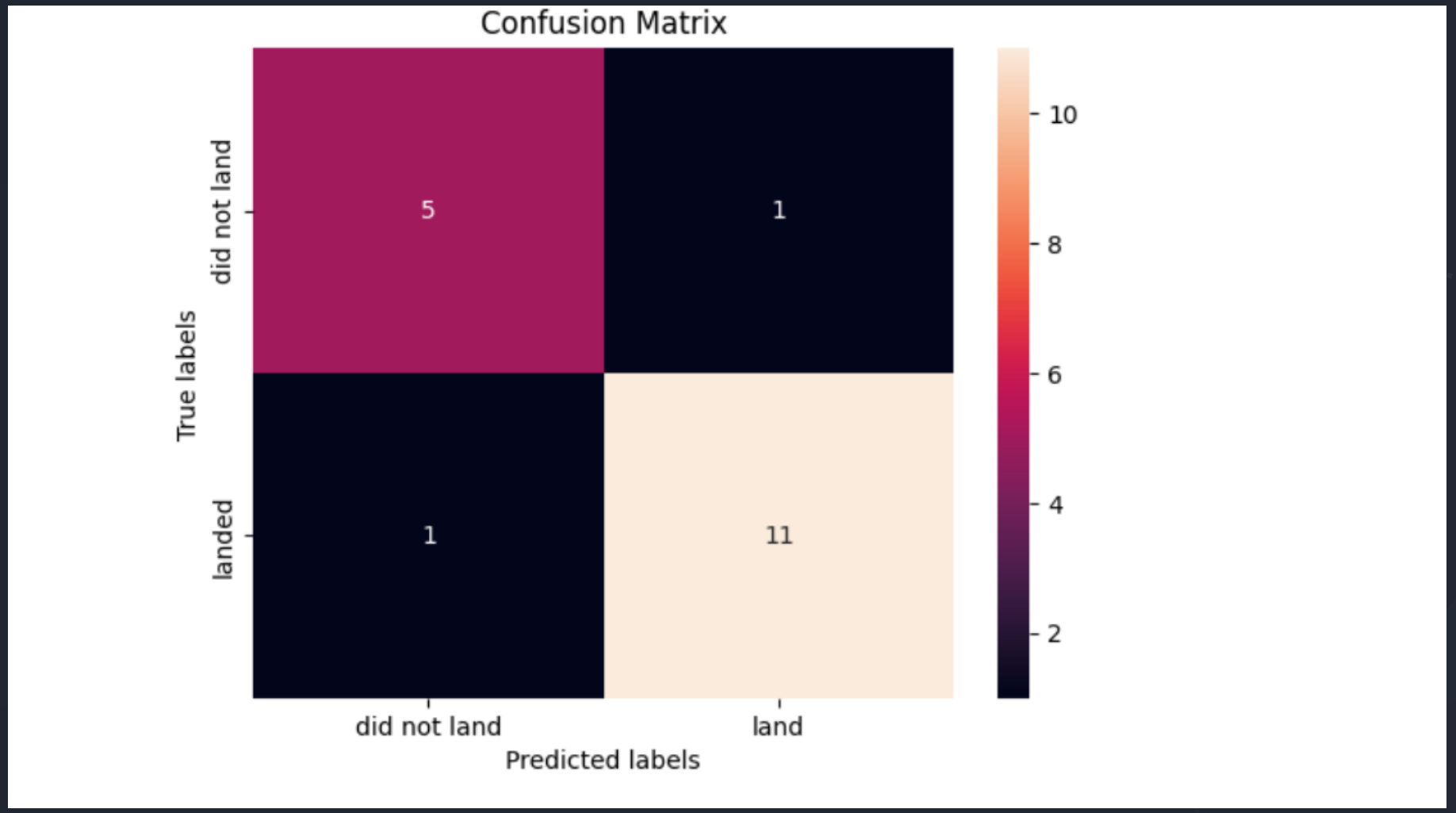
- logreg_cv best score: 0.8464285714285713
- Accuracy score on test set: 0.8333333333333334

Support Vector Machine (SVM) - Confusion Matrix



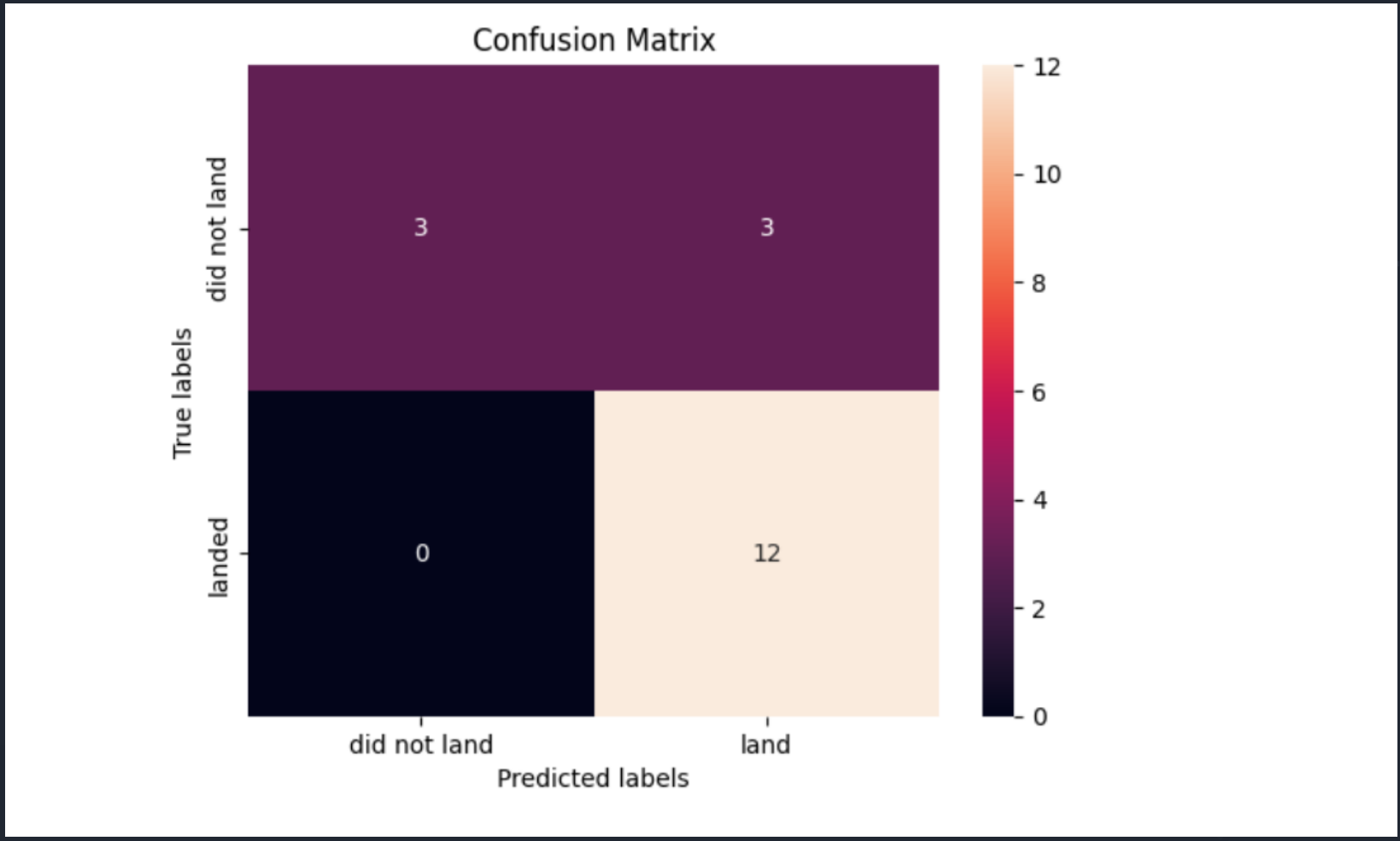
- svm_cv best score: 0.8482142857142856
- Accuracy score on test set: 0.8333333333333334

Decision Tree - Confusion Matrix



- tree_cv best score: 0.8892857142857142
- Accuracy score on test set: 0.8333333333333334

K Nearest Neighbors (KNN) - Confusion Matrix



- knn_cv best score: 0.8482142857142858
- Accuracy score on test set: 0.8333333333333334

DISCUSSION

- Putting the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix when tested on the test set.
- Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:

1. Decision tree (tree_cv best score: 0.889285714285714 2)
2. K nearest neighbors, KNN (knn_cv best score: 0.848214285714285 8)
3. Support vector machine, SVM (svm_cv best score: 0.848214285714285 6)
4. Logistic regression (logreg_cv best score: 0.846428571428571 3)

CONCLUSION

- In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.
- Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.
- Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed.

Thank you!