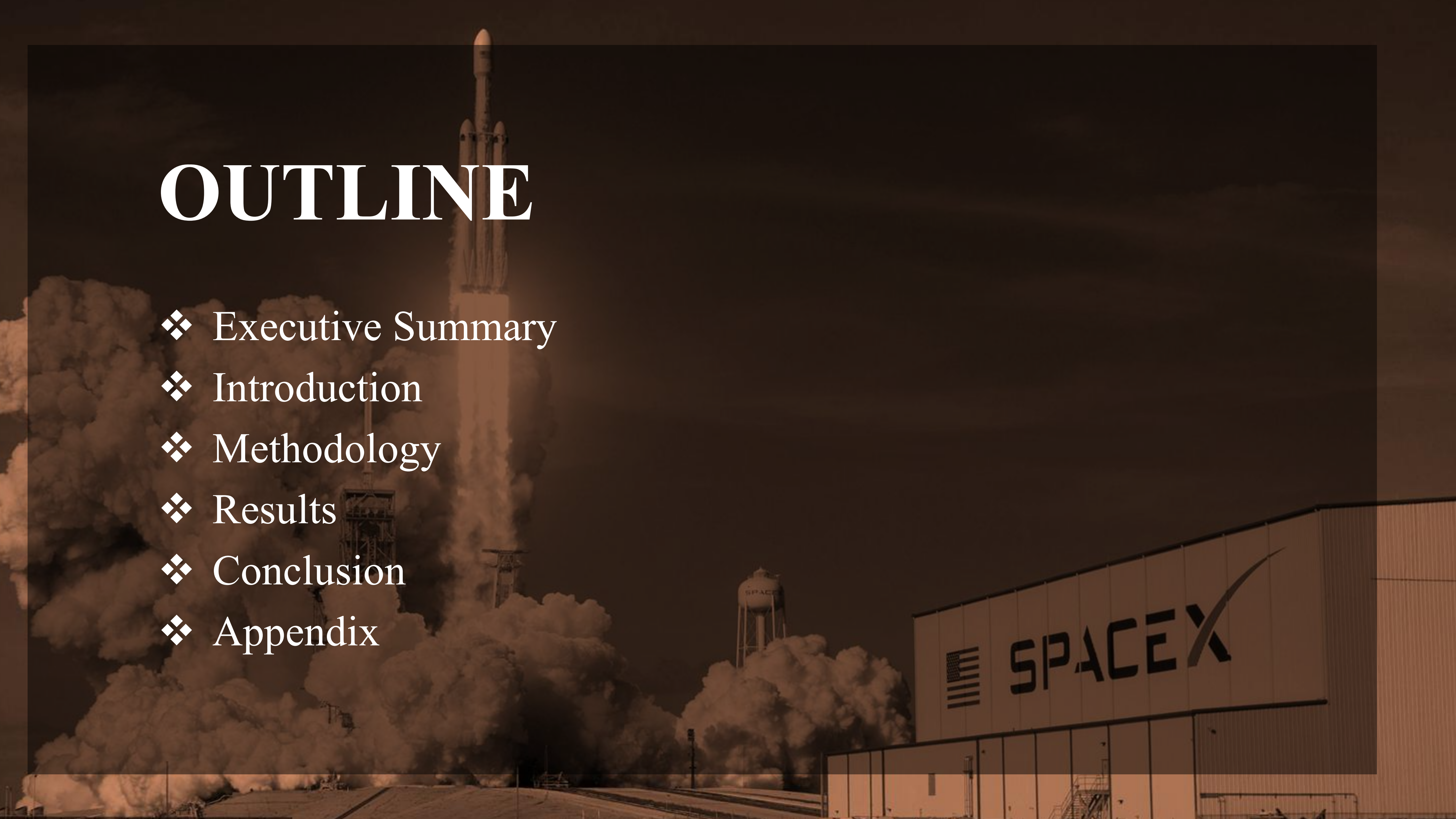
APPLIED DATA SCIENCE CAPSTONE

MUBASHIRALI
June 14, 2025
Founder Code with Bismillah

OUTLINE

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



EXECUTIVE SUMMARY

- In this Applied Data Science capstone project, I took on the exciting challenge of predicting whether the SpaceX Falcon 9 first stage would successfully land or not a real-world scenario that merges the thrill of space exploration with the power of data science.
- My journey began with gathering and preparing the launch data cleaning it, organizing it, and shaping it into a form that could actually "speak" to us. Then, through exploratory data analysis (EDA), I dug deep into patterns and trends hidden in the numbers.
- Using interactive visualizations, I was able to not just analyze, but truly *understand* how different factors like payload mass, orbit type, or booster version might impact the rocket's success or failure.
- Finally, I applied multiple machine learning algorithms to make predictions. Among them, the Decision Tree model stood out, providing the most reliable results in predicting whether a rocket would safely land back or not.
- This project not only improved my technical skills but also gave me a taste of how data science can be used to decode the mysteries of the universe quite literally.

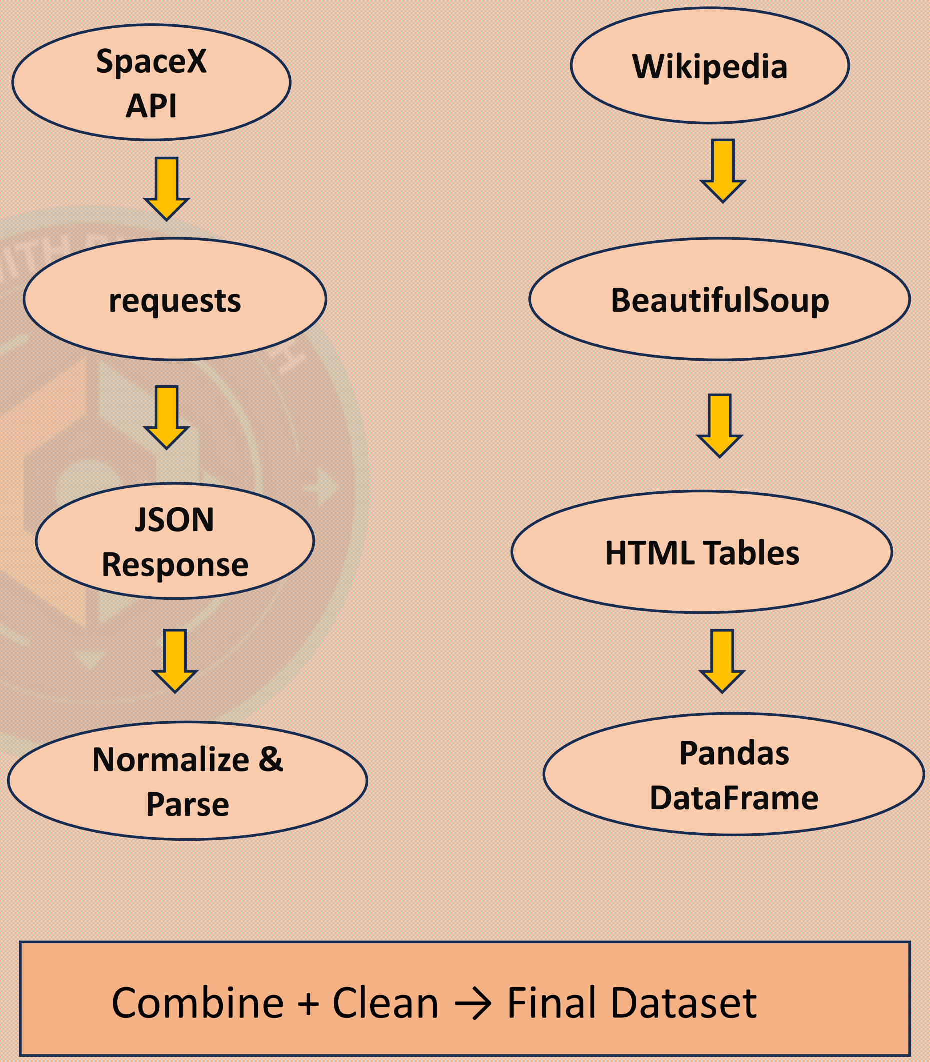
INTRODUCTION

- 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.
- Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean.
- 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

The Complete Methodology includes:

- ➢ Data collection, wrangling, and formatting:
  - SpaceX API
  - Web scraping
- ➢ Exploratory data analysis (EDA):
  - Pandas and NumPy
  - SQL
- ➢ Data visualization, using:
  - Matplotlib and Seaborn
  - Folium
  - Dash
- ➢ Machine learning prediction:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)



➢ Data collection, wrangling, and formatting

SpaceX API

- The API used in the capstone project was provided by Coursera: https://api.spacexdata.com/v4/rockets/
- We used the SpaceX API to get data about different rocket launches. But since our project is only about Falcon 9, we filtered the data to include only Falcon 9 launches.
- Some columns had missing values, so we filled those with the average (mean) of that column.
- In the end, our cleaned dataset had 90 rows (launch records) and 17 columns (features). Below is a preview of the dataset after cleaning.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

➢ Data collection, wrangling, and formatting

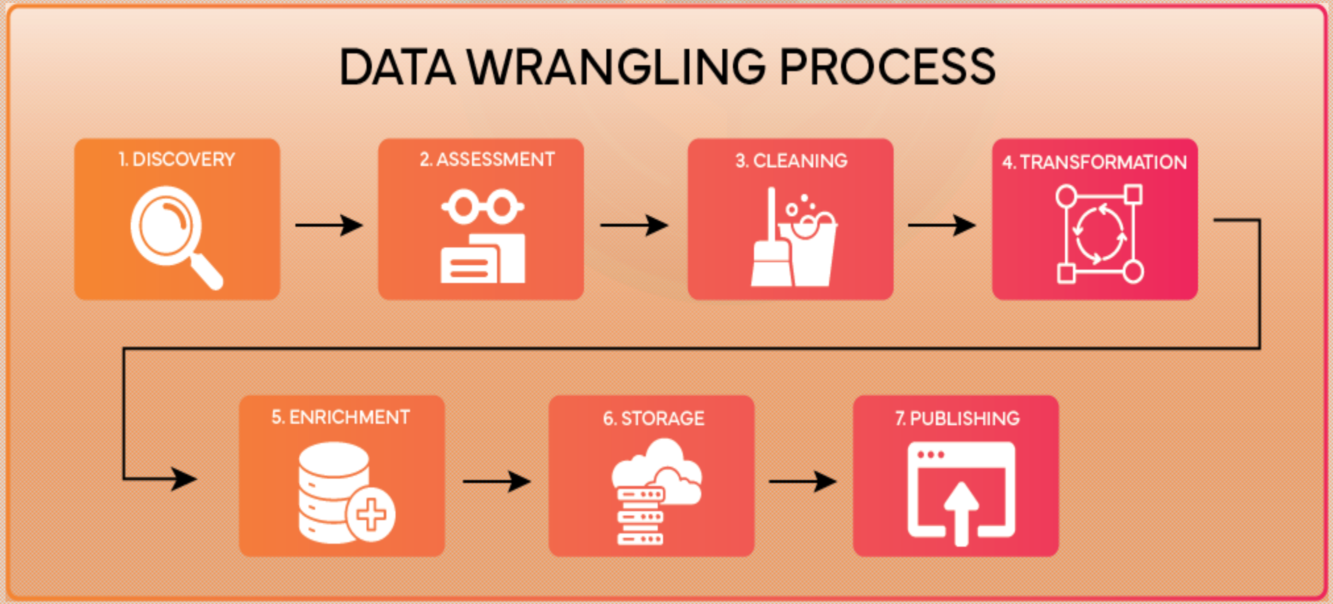
Web Scraping

- The data used was scraped from https://en.wikipedia.org/w/index.php?title=list_of_falcon_9_and_falcon_heavy_launches&oldid=1027686922
- The website provides data specifically related to Falcon 9 launches only.
- After processing the data, we obtained a final dataset with 121 rows (records) and 11 columns (features).
  The table below displays the first few entries of this dataset:

	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 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

➢ Data collection, wrangling, and formatting

- Next, the data was cleaned to remove any missing values, and all categorical features were converted into numerical form using one-hot encoding.
- We also added a new column called ‘Class’, which indicates the outcome of each launch — where ‘0’ means the launch failed, and ‘1’ means it was successful.
- After this processing, the final dataset contained 90 rows (launches) and 83 columns (features)



➢ Exploratory Data Analysis (EDA)

1. Pandas & NumPy Analysis

We used functions from Pandas and NumPy libraries to explore and understand the dataset. Some of the insights we extracted include:

- How many launches took place at each launch site
- How often each orbit type was used
- The count and types of different mission outcomes

2. SQL-Based Exploration

Using SQL queries, we explored the dataset further to answer key questions such as:

- What are the unique launch sites in the dataset?
- How much total payload mass was launched by NASA (CRS) missions?
- What is the average payload mass for the booster version **F9 v1.1**?

➢ Data Visualization

1. Matplotlib and Seaborn

We used Matplotlib and Seaborn to create different types of plots like scatter plots, bar charts, and line charts etc.

These visualizations helped us understand key relationships in the data, such as:

- How launch site varies with flight number
- How payload mass is distributed across launch sites
- How orbit type is related to success rate of launches

2. Interactive Mapping with Folium

Folium was used to create interactive maps for deeper spatial understanding.

With Folium, we were able to:

- Plot all launch sites on a map
- Highlight which launches were successful or failed at each site
- Show distances from launch sites to nearby locations like cities, highways, and railways

➢ Data Visualization

3. Interactive Dashboard with Dash

We used **Dash** to build an interactive web dashboard where users can explore the data using a dropdown menu and a range slider.

The dashboard includes:

- A **pie chart** that shows the total number of successful launches from each launch site
- A **scatter plot** that reveals the relationship between **payload mass** and **mission outcome** (success or failure) for different launch sites

➢ Machine Learning Prediction

We used the Scikit-learn library to build and evaluate multiple machine learning models. The entire prediction process followed these key steps:

1. **Standardized** the data for better model performance
2. **Split** the dataset into training and testing sets
3. **Trained four different models:**
   - Logistic Regression
   - Support Vector Machine (SVM)
   - Decision Tree
   - K-Nearest Neighbors (KNN)
4. **Fitted** each model on the training data
5. **Tuned hyperparameters** to find the best settings for each model
6. **Evaluated performance** using:
   - **Accuracy scores**
   - **Confusion matrix** for each model

RESULTS

The results of our analysis are divided into five main sections:

1. EDA using SQL
2. EDA using Matplotlib & Seaborn visualizations
3. Interactive maps using Folium
4. Interactive dashboard using Dash
5. Predictive analysis using machine learning models

Throughout all the charts and graphs:

- Class 0 indicates a failed launch
- Class 1 indicates a successful launch

Exploratory Data Analysis (EDA) using SQL

We retrieved the names of all unique launch sites involved in the SpaceX missions to understand how many different sites were used.

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

Then, we filtered the data to show only those launch sites that begin with 'CCA' revealing 5 specific records matching this condition.

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Exploratory Data Analysis (EDA) using SQL

Total payload mass transported by boosters for NASA (CRS) missions

Total payload mass by NASA (CRS)
45596

Mean payload mass delivered by the F9 v1.1 booster version

Average payload mass by Booster Version F9 v1.1
2928

The date on which the first successful landing on a ground pad was recorded

Date of first successful landing outcome in ground pad
2015-12-22

Exploratory Data Analysis (EDA) using SQL

Names of boosters that successfully landed on a drone ship and carried a payload between 4000 kg and 6000 kg

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

Total count of missions categorized by success and failure outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Exploratory Data Analysis (EDA) using SQL

List of booster versions with the maximum recorded payload capacity

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Exploratory Data Analysis (EDA) using SQL

Booster versions, launch site names, and failed drone ship landings that occurred during the year 2015

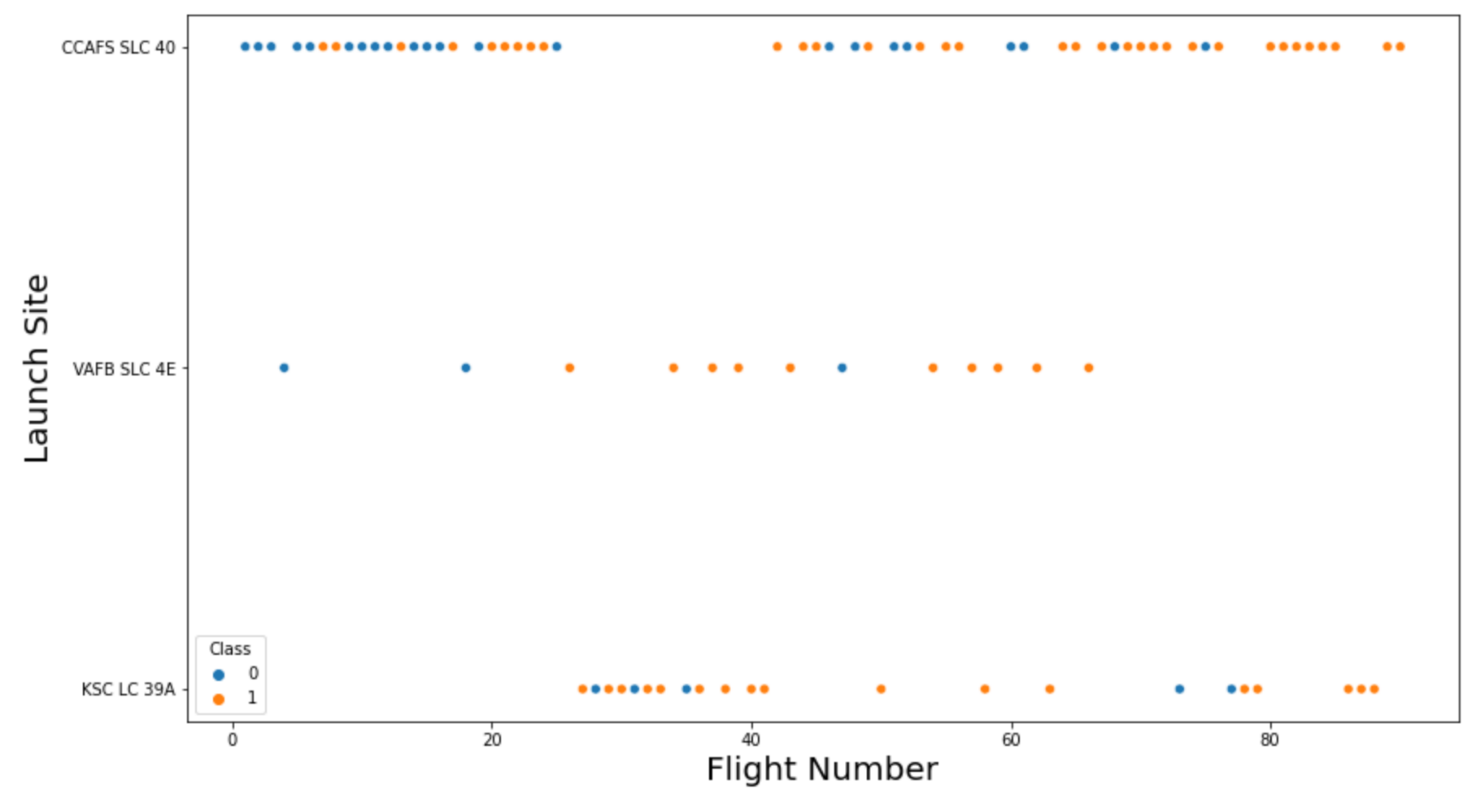
DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

Number of landing outcomes recorded between June 4, 2010, and March 20, 2017, sorted in descending order

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

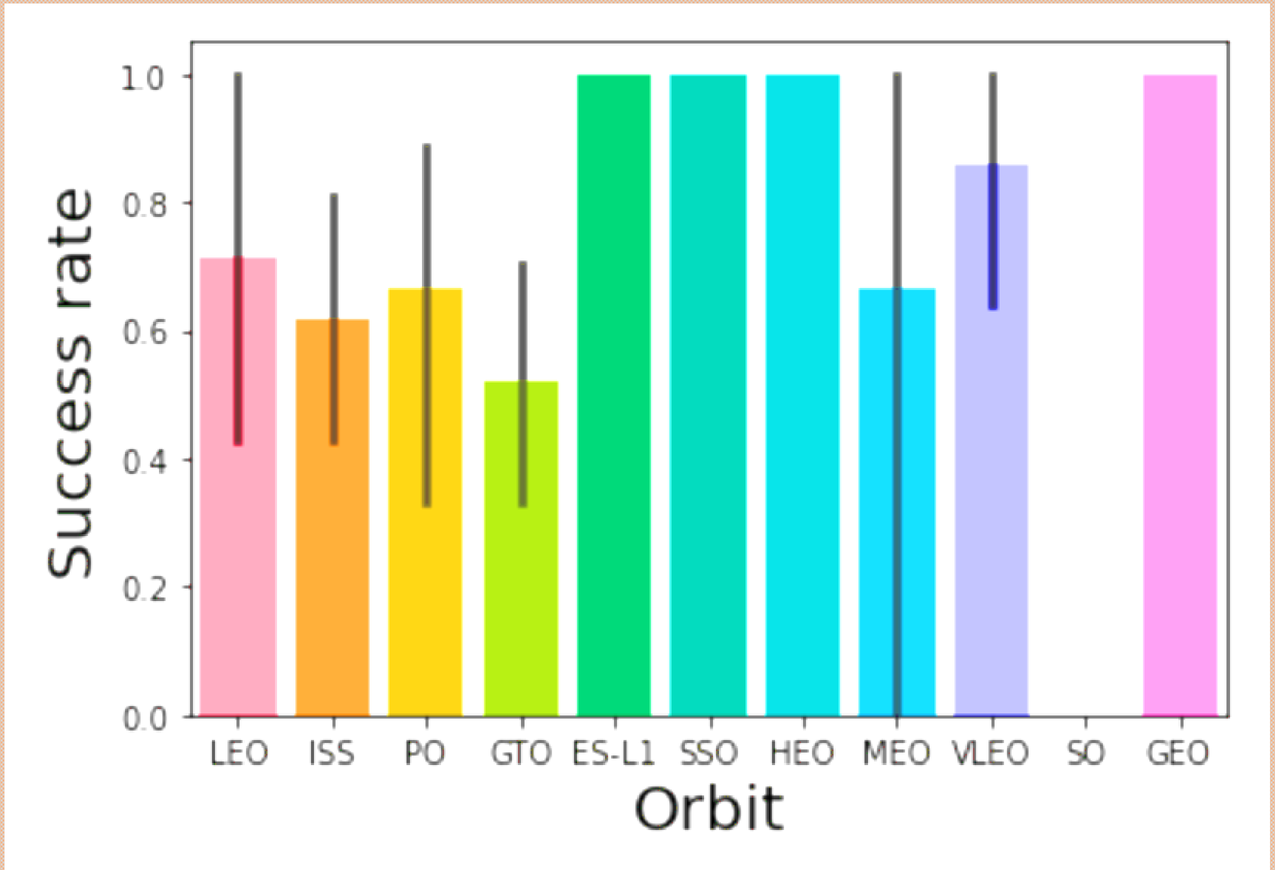
Matplotlib and Seaborn (EDA with Visualization)

Analyzing how flight numbers vary across different launch sites



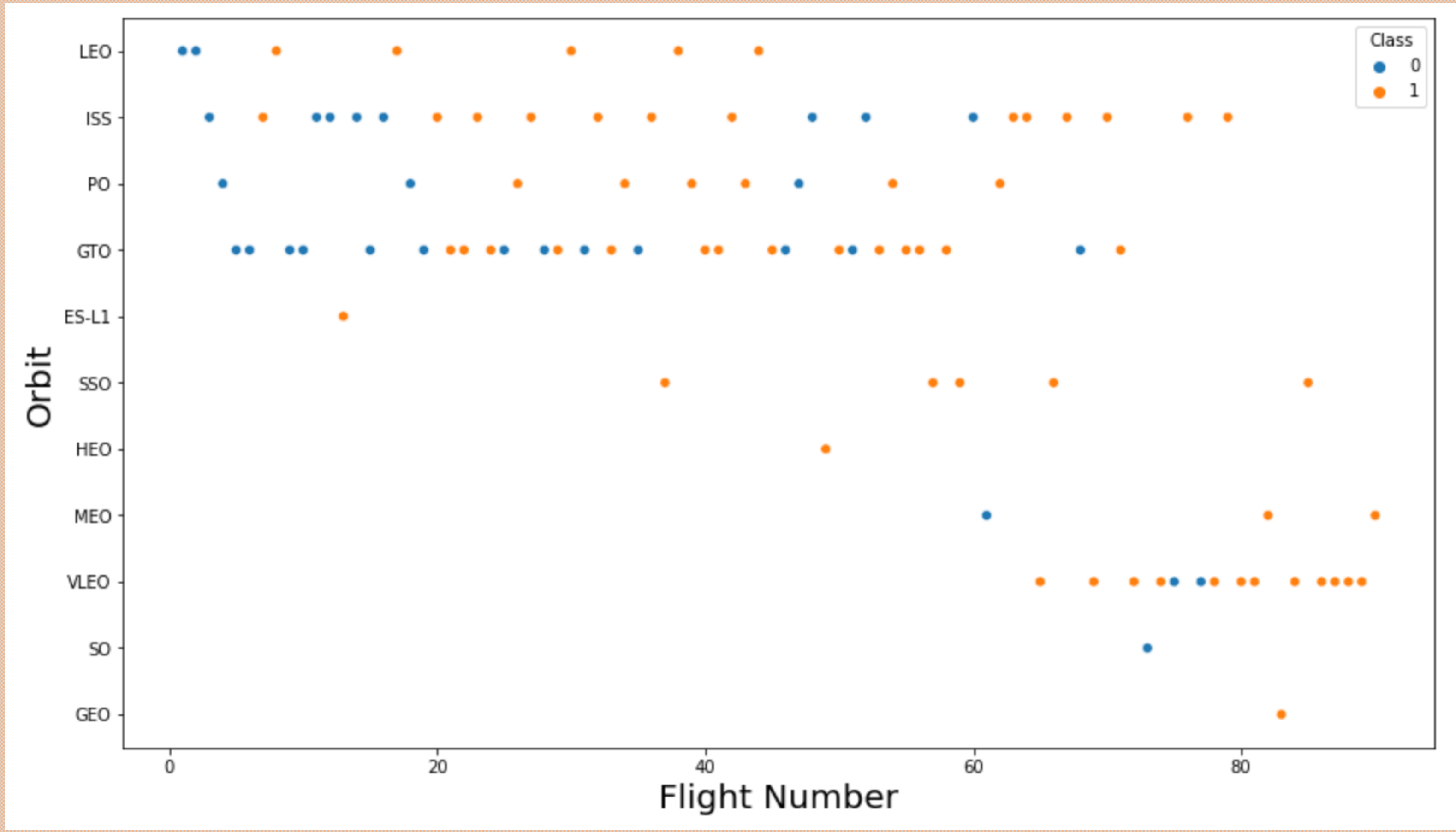
Matplotlib and Seaborn (EDA with Visualization)

Analyzing the link between mission outcome success and the type of orbit used



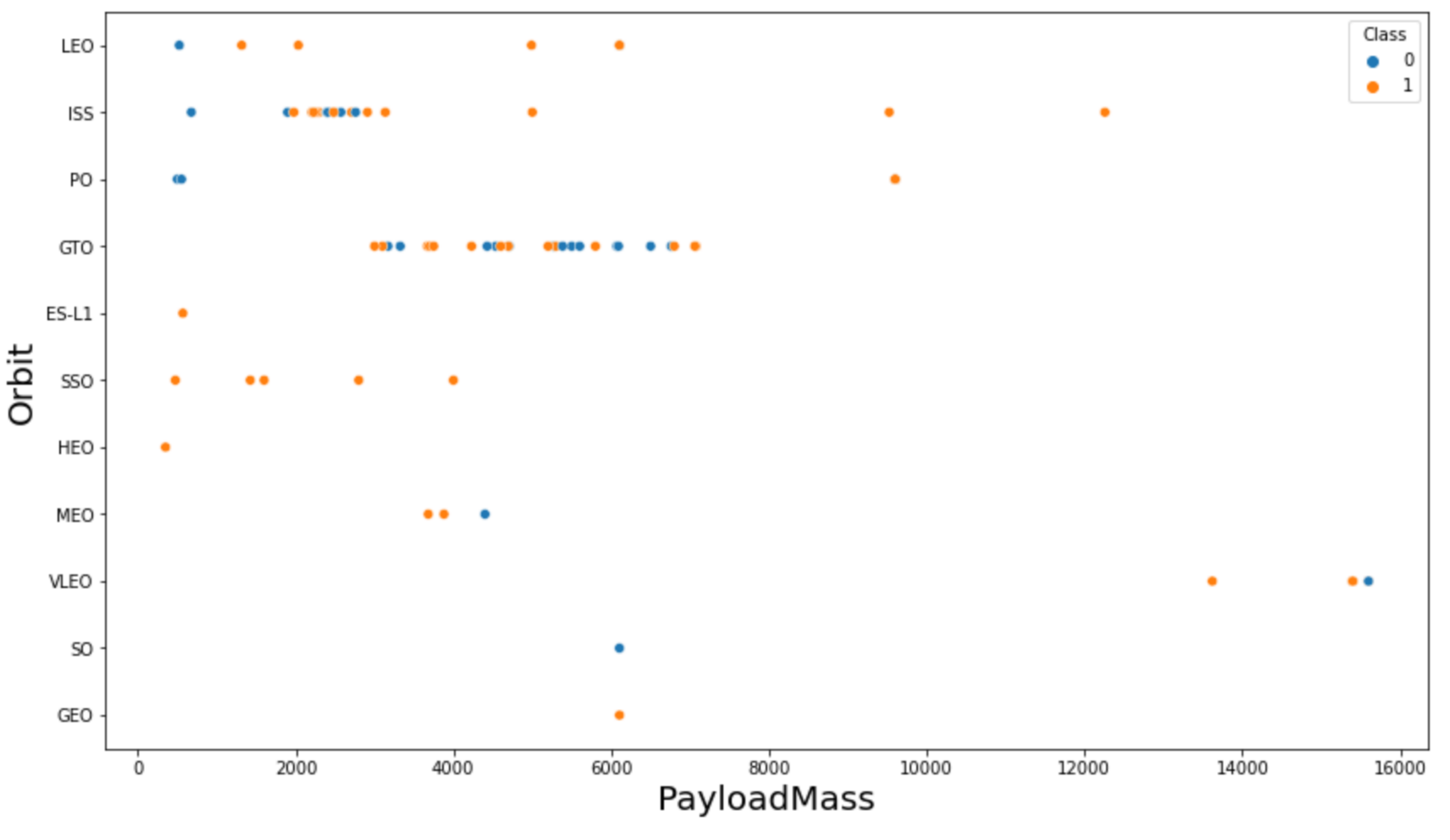
Matplotlib and Seaborn (EDA with Visualization)

Analyzing how orbit types are distributed across different flight numbers



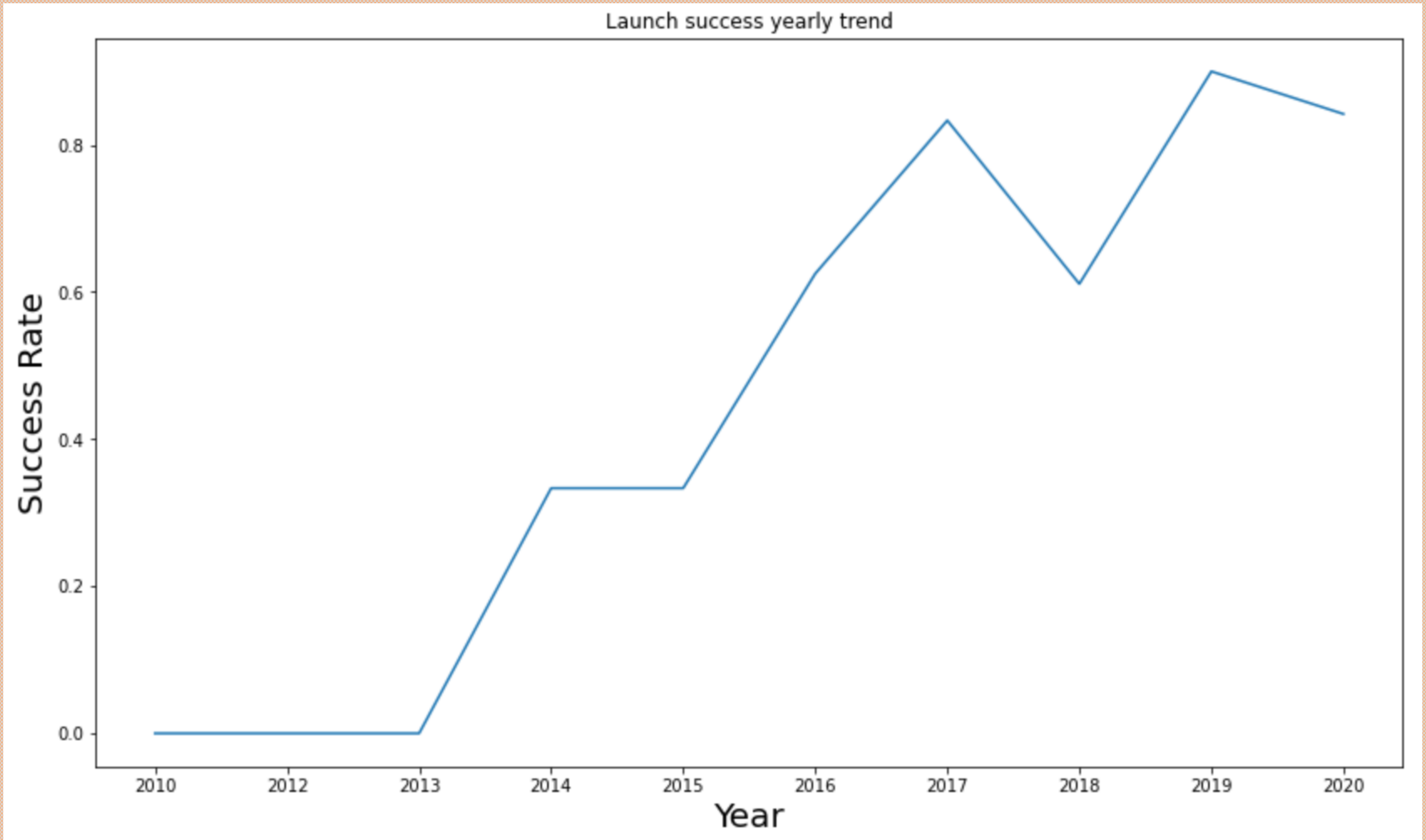
Matplotlib and Seaborn (EDA with Visualization)

Analyzing the connection between the weight of the payload and the orbit targeted



Matplotlib and Seaborn (EDA with Visualization)

The launch success yearly trend



➢ FOLIUM

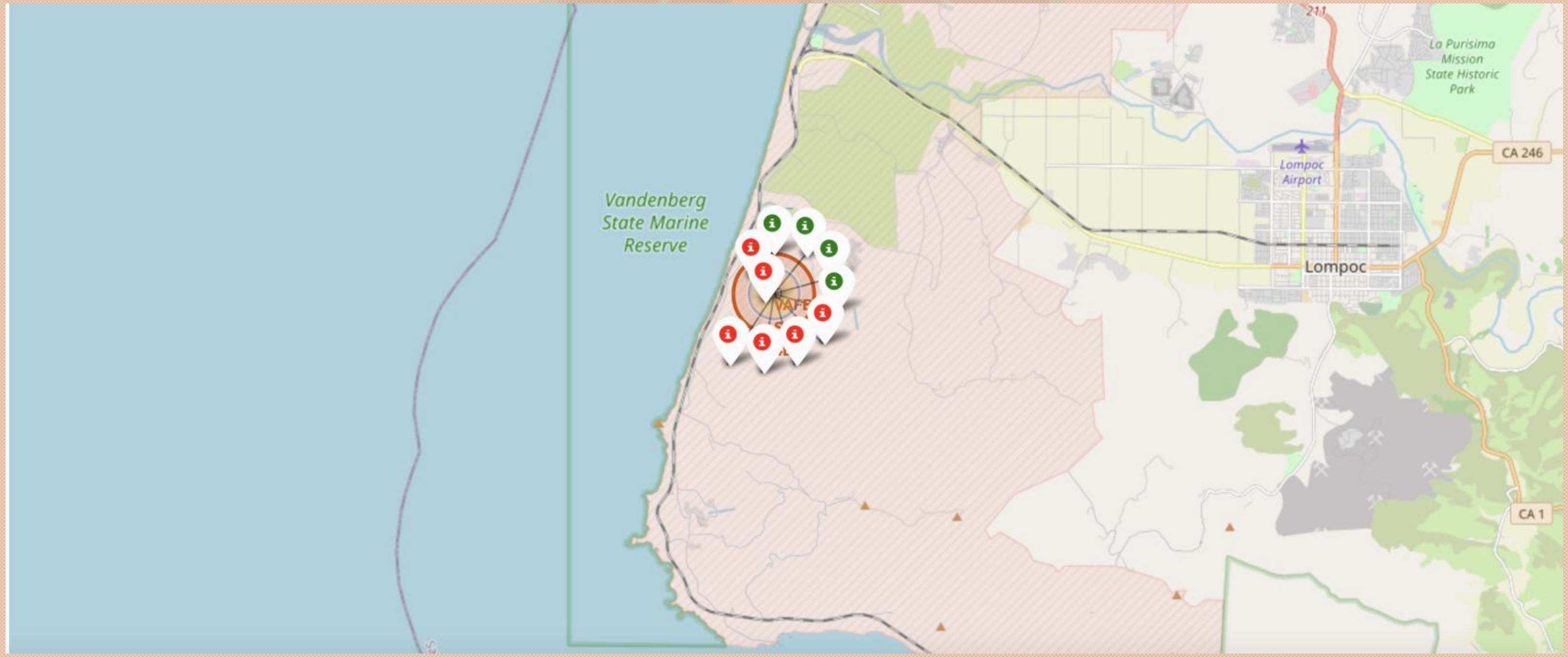
All the launch sites on map

➢ FOLIUM

Each launch site is marked on the map, showing both successful and failed missions.
When you **zoom in** on a specific site, you'll notice **green and red markers**:

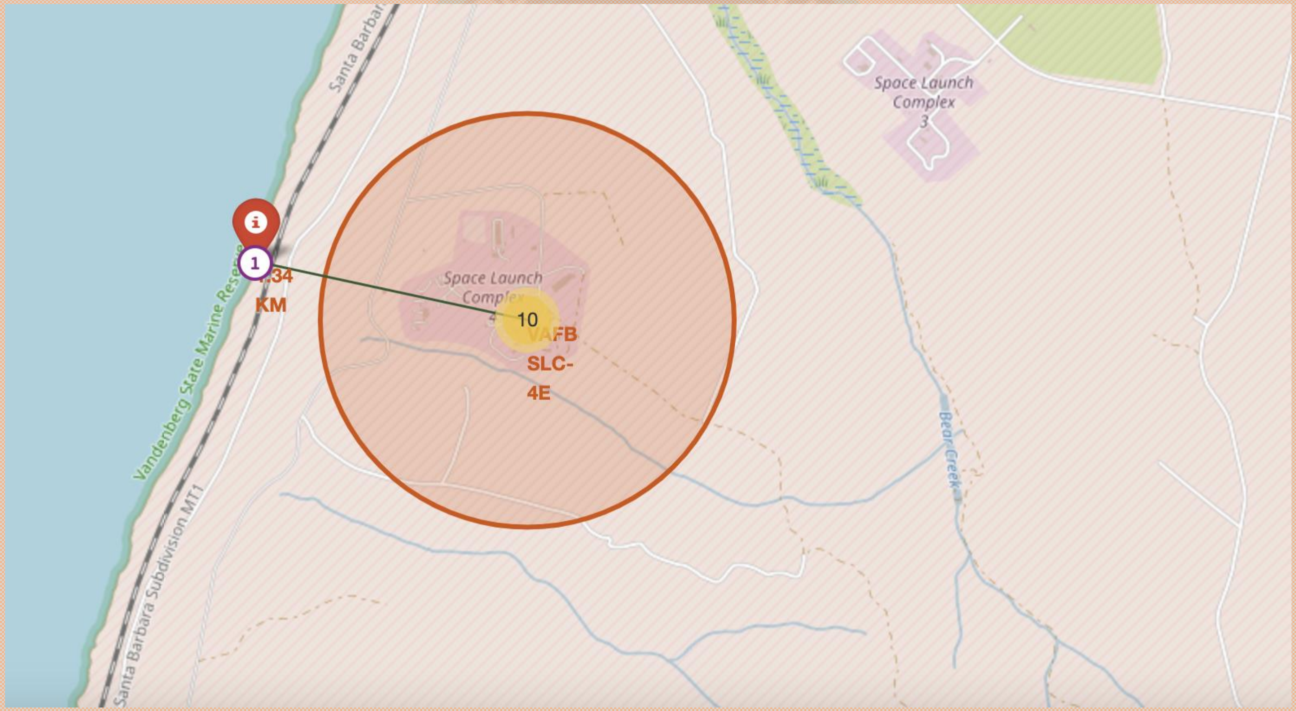
- **Green markers** indicate **successful launches**
- **Red markers** indicate **failed launches**

This visual helps clearly compare the performance of each site based on past outcomes.

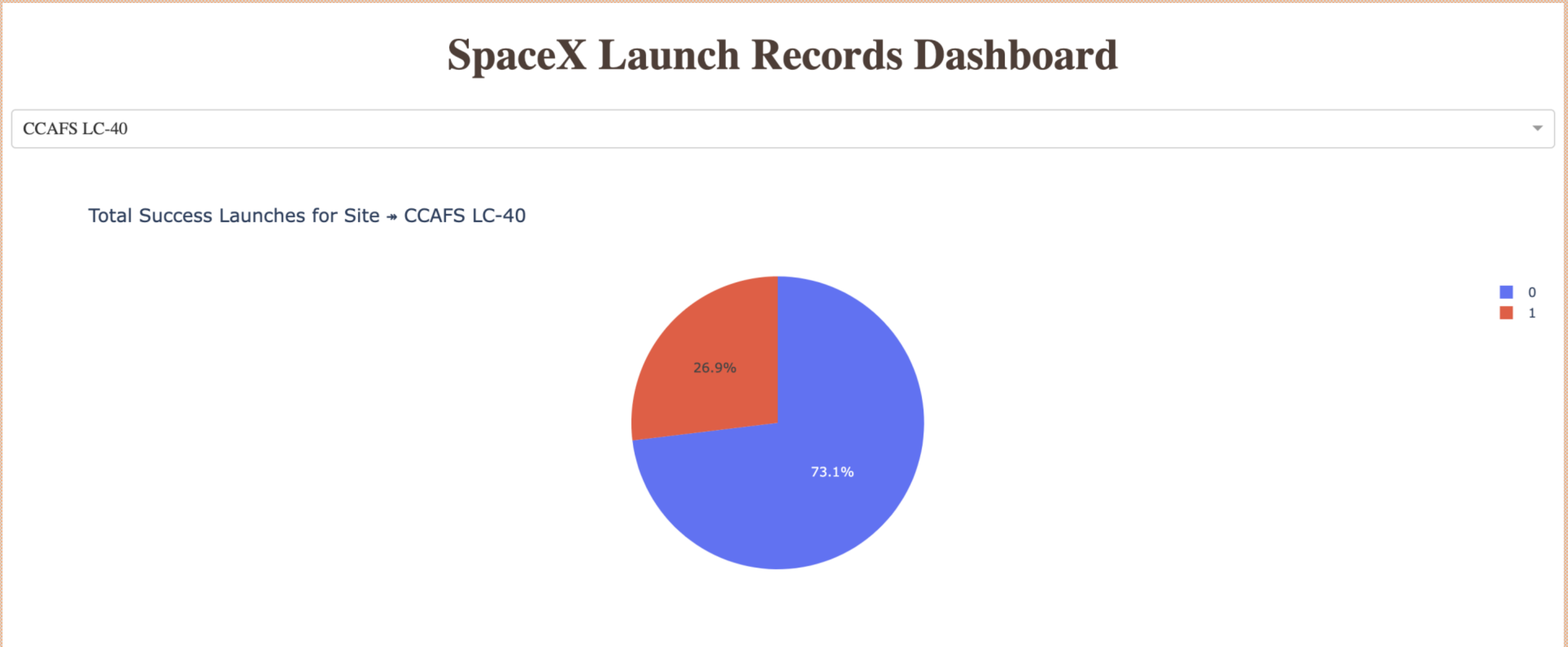


➢ FOLIUM

We calculated the distances from each launch site to nearby key locations including the nearest city, railway line, and highway.
The image below highlights one such example:
It shows the distance from the VAFB SLC-4E launch site to the nearest coastline.



➢ DASH



SpaceX Launch Records Dashboard
CCAFS LC-40
Total Success Launches for Site ↠ CCAFS LC-40
26.9%
73.1%
0
1

➢ DASH

The scatterplot below displays the data when the payload mass is filtered between 2000 kg and 8000 kg. In this plot:

- **Class 0** = Failed Launches ( )
- **Class 1** = Successful Launches ( )

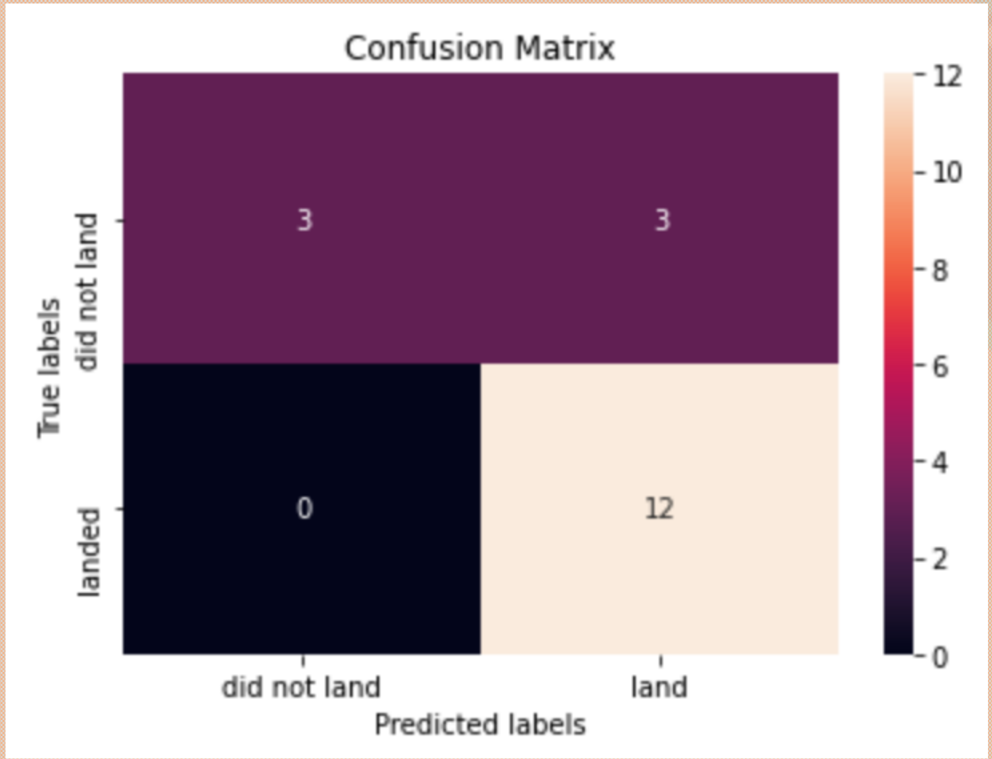
This visualization helps in understanding how payload weight may relate to the launch outcome

➢ PREDICTIVE ANALYSIS

Logistic Regression Results
Best Score from GridSearchCV: 0.846
Accuracy on Test Data: 0.833
Confusion Matrix:

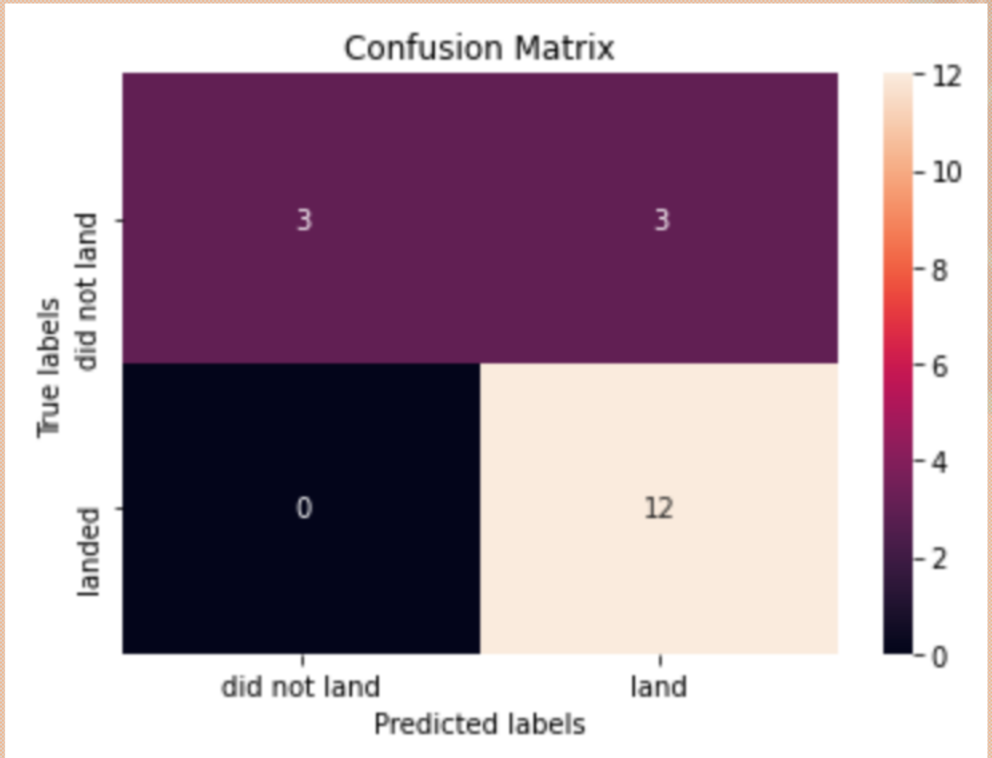
➢ PREDICTIVE ANALYSIS

Decision Tree Results
Best Score from GridSearchCV: 0.889
Accuracy on Test Set: 0.833
Confusion Matrix:



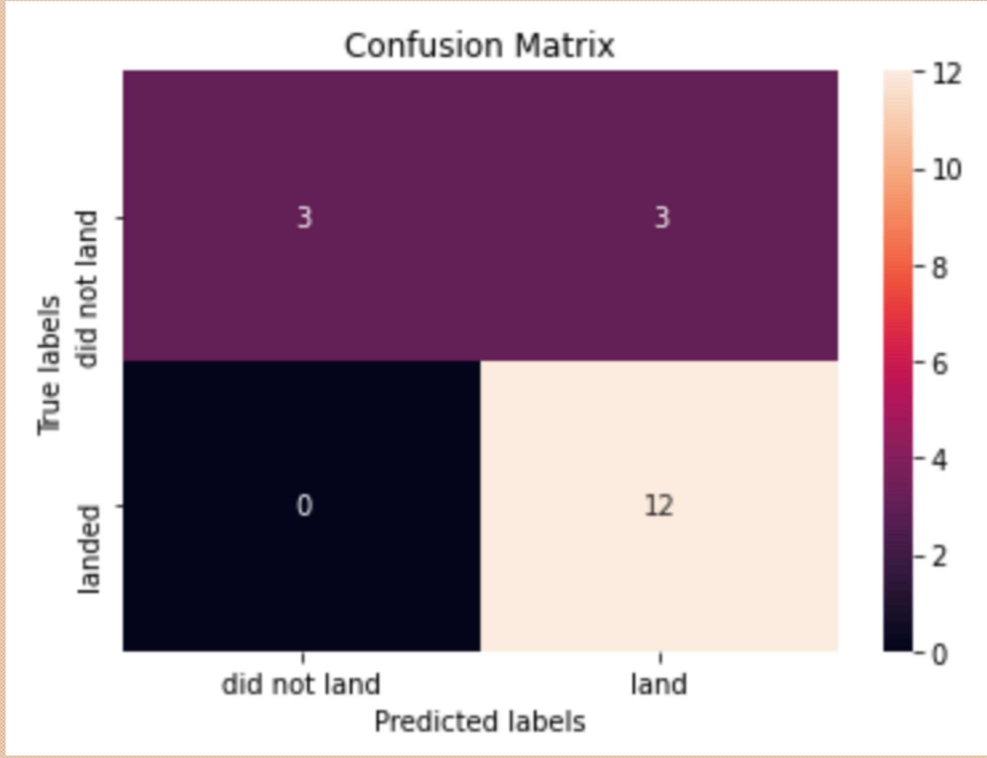
➢ PREDICTIVE ANALYSIS

Support Vector Machine (SVM) Results
Best Score from GridSearchCV: 0.848
Accuracy on Test Set: 0.833
Confusion Matrix:



➢ PREDICTIVE ANALYSIS

K-Nearest Neighbors (KNN) Results
Best Score from GridSearchCV: 0.848
Accuracy on Test Set: 0.833
Confusion Matrix:



➢ PREDICTIVE ANALYSIS

Model Comparison & Ranking

All four machine learning models Logistic Regression, SVM, Decision Tree, and KNN produced identical accuracy scores (83.3%) and confusion matrices when evaluated on the test set.

Since the accuracy was the same across models, we used their GridSearchCV best scores to determine overall performance. Based on those scores, the models are ranked from best to worst as follows:

Decision Tree: *Best Score: 0.889*

K-Nearest Neighbors (KNN): *Best Score: 0.848*

Support Vector Machine (SVM): *Best Score: 0.848*

Logistic Regression : *Best Score: 0.846*

This ranking highlights the Decision Tree model as the top performer when hyperparameters are optimized.

➢ PREDICTIVE ANALYSIS

- The core objective of this project was to predict the success or failure of the Falcon 9 rocket's first-stage landing. This task holds significant value, as a successful landing greatly reduces SpaceX's launch expenses by enabling reusability of rocket components.
- To tackle this, we explored various launch-related features such as payload mass, orbit type, and others to understand their impact on mission outcomes.
- We applied multiple machine learning algorithms on historical launch data to identify meaningful patterns that could improve future predictions.
- Out of all the models, the Decision Tree classifier emerged as the top performer, showing the highest predictive accuracy among the approaches tested.

Thank You