Using Data Science to Predict Falcon 9 Landing

Luciana Rios

26 – 09 – 2023

IBM Developer

SKILLS NETWORK

OUTLINE

- Summary
- Introduction
- Methodology
  - Data Collection and Data Wrangling
  - Exploratory Data Analysis (EDA) and Data Wrangling
  - EDA and Interactive Visual Analytics
  - Predictive Analysis
- Results
  - EDA - Visualization and SQL Results
  - Folium Maps
  - Dashboard
  - Predictive Analysis (Classification)
- Conclusion
- Appendix

SUMMARY

SpaceX rocket launches are relatively inexpensive, as the first stage of their rockets can be reused. It is possible to predict the success of the next landing using rocket science, but is it possible to make such a prediction using data science? Next, we will see that, with the right data and a careful methodology, the answer to this question is positive.

Figure 1: Falcon Heavy launch (figure on the side) and its reusable side boosters landing (cover figure) - SpaceX

INTRODUCTION

- SpaceX was the first private company to send humans into space.
- Their rocket launches are relatively inexpensive, as the rocket's first stage can usually be reused.
- SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars, while other providers cost upwards of 165 million dollars.
- If it is possible to determine whether the first stage will land successfully, it is also possible to determine the cost of the next launch.
- For a rocket company that would like to compete with SpaceX, pricing each competitor's launch involves determining whether the first stage will be reused.
- Instead of using rocket science to determine whether the first stage will land successfully, public data will be used to train a machine learning model to predict if SpaceX will reuse the first stage.

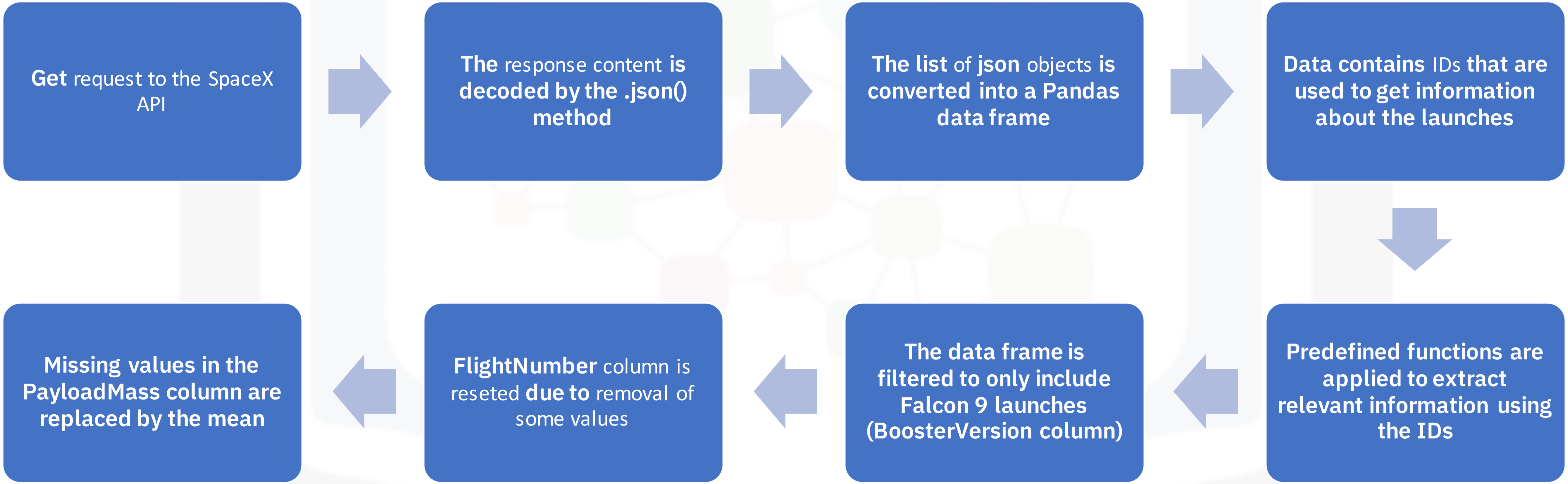
METHODOLOGY

- Data Collection and Data Wrangling
- Exploratory Data Analysis (EDA) and Data Wrangling
- EDA and Interactive Visual Analytics
- Predictive Analysis

DATA COLLECTION AND DATA WRANGLING

- Launch data was gathered from the SpaceX REST API.
- The API provides data that includes information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Functions were defined to help the API to extract relevant information using identification numbers in the launch data.
- The goal is to use the collected data to predict whether the Falcon 9 first stage will land successfully (landing outcome); therefore, data associated with the Falcon 1 launch has been removed.
- We also had to deal with some missing values for the payload mass (PayloadMass attribute), which were replaced by the mean.
- Null values found in the LandingPad column of the final data table were retained as they represent when a landing pad was not used.

DATA COLLECTION AND DATA WRANGLING



Get request to the SpaceX API

The response content is decoded by the .json() method

The list of json objects is converted into a Pandas data frame

Data contains IDs that are used to get information about the launches

Predefined functions are applied to extract relevant information using the IDs

The data frame is filtered to only include Falcon 9 launches (BoosterVersion column)

FlightNumber column is reseted due to removal of some values

Missing values in the PayloadMass column are replaced by the mean

DATA COLLECTION AND DATA WRANGLING

- Falcon 9 historical launch records were also obtained from Wikipedia through web scraping.
- Both sources of data were useful to generate a clean dataset which provided meaningful information on the situation we are trying to address.
- After using Python to web scrape some HTML tables with valuable launch records, the data is parsed and the tables are converted into a data frame for further visualization and analysis.
- Some functions were defined to help to process the web scraped HTML table.

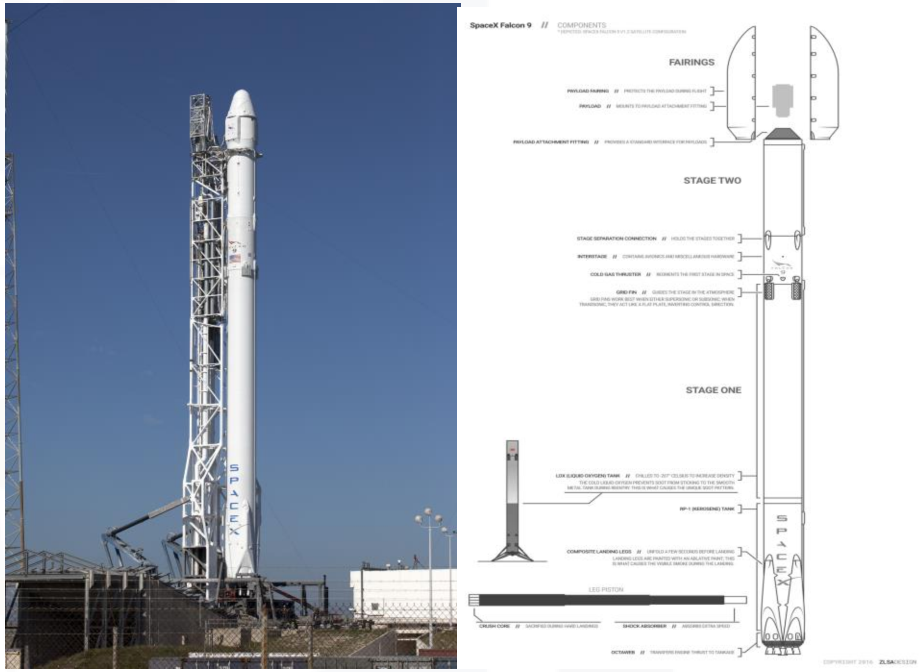
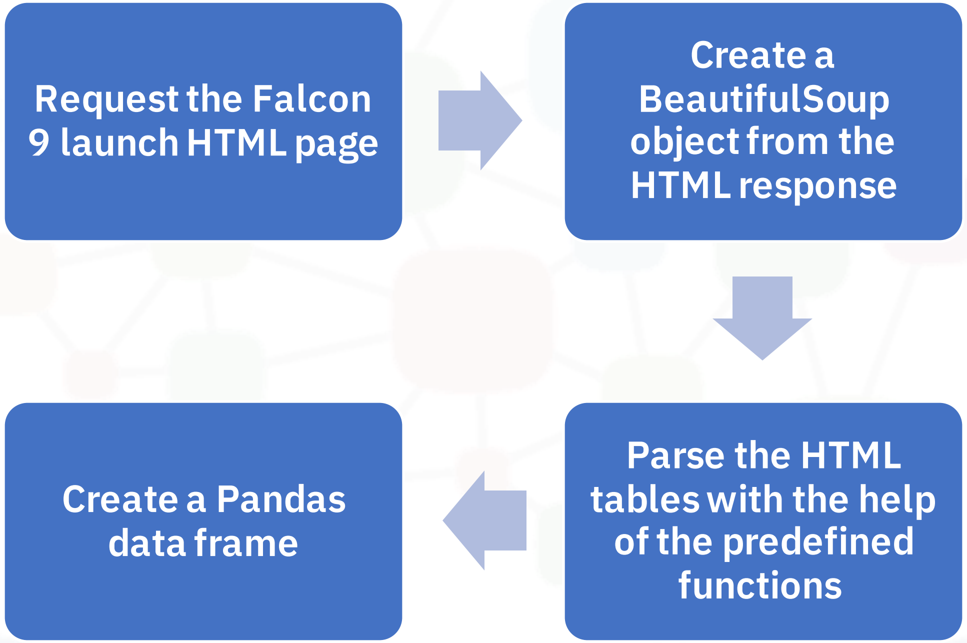

SpaceX Falcon 9 // COMPONENTS
FAIRINGS
PAYLOAD FAIRING // PROTECTS THE PAYLOAD DURING FLIGHT
PAYLOAD // MOUNTS TO PAYLOAD ATTACHMENT FITTING
PAYLOAD ATTACHMENT FITTING // PROVIDES A STANDARD INTERFACE FOR PAYLOADS
STAGE TWO
STAGE SEPARATION CONNECTION // HOLDS THE STAGES TOGETHER
INTERSTAGE // CONTAINS AVIONICS AND MISCELLANEOUS HARDWARE
COLD GAS THRUSTER // REORIENTS THE FIRST STAGE IN SPACE
GRID FIN // GUIDES THE STAGE IN THE ATMOSPHERE
STAGE ONE
LOX (LIQUID OXYGEN) TANK // CHILLED TO -207° CELSIUS TO INCREASE DENSITY
RP-1 (KEROSENE) TANK
COMPOSITE LANDING LEGS // UNFOLD A FEW SECONDS BEFORE LANDING
LEG PISTON
CRUSH CORE // SACRIFICED DURING HARD LANDINGS
SHOCK ABSORBER // ABSORBS EXTRA SPEED
OCTAWEB // TRANSFERS ENGINE THRUST TO TANKAGE
SPACEX
COPYRIGHT 2016 ZLSADESIGN

Figure 2: Falcon 9 Full Thrust (also known as Falcon 9 v1.2) and a infographic displaying its components (Kim Shiflett - NASA/zlsadesign.com)

DATA COLLECTION AND DATA WRANGLING



Request the Falcon 9 launch HTML page

Create a BeautifulSoup object from the HTML response

Parse the HTML tables with the help of the predefined functions

Create a Pandas data frame

EDA AND DATA WRANGLING

- The data collection process resulted in the creation of a data frame.
- It is important to perform some exploratory analysis on the resulting data frame to find patterns and determine the label and relevant attributes for training supervised models.
- Some relevant attributes are:

Flight Number (increases over time)	Booster Version	Payload Mass	Orbit (orbit of the payload)
Launch Site	Grid Fins (help with landing)	Legs (used in landing)	Landing Pad

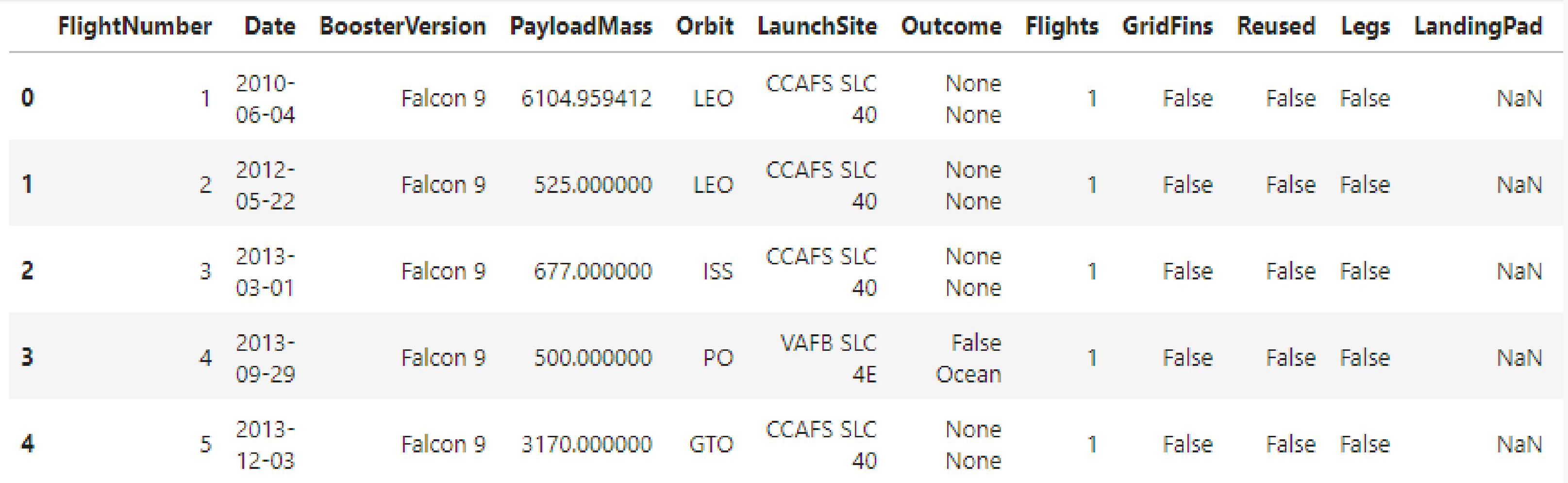
EDA AND DATA WRANGLING

- The column Outcome indicates if the first stage successfully landed (some unsuccessful landings are planned). There are 8 options:

First Stage Landed (Success)	First Stage Failed to Land
True Ocean (landed in the ocean)	False Ocean
True RTLS (landed to a ground pad)	False RTLS
True ASDS (landed on a drone ship)	False ASDS
	None ASDS
	NoneNone

EDA AND DATA WRANGLING

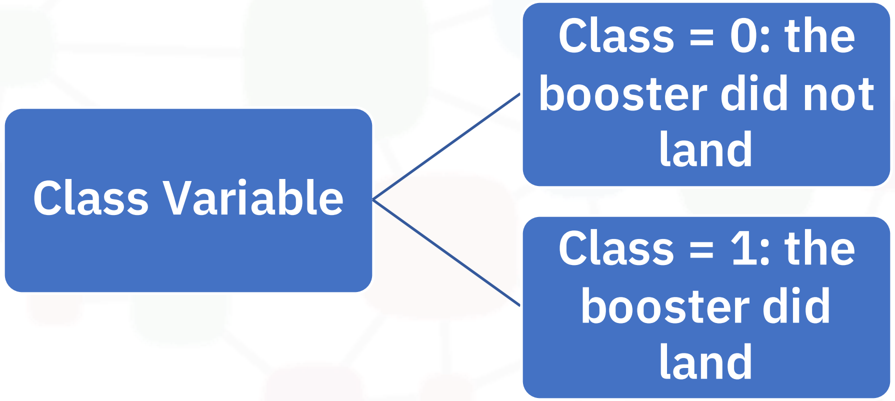
- Figure 3 shows some rows and columns of the data frame, including column Outcome.
- A new variable, Class, was created and associated to the landing outcome. It works as a classification variable that represents the outcome of each launch.



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN

Figure 3: First 5 rows and 12 columns of the data frame.

EDA AND DATA WRANGLING



Class Variable

Class = 0: the booster did not land

Class = 1: the booster did land

- The outcomes were converted to training labels 0 and 1 because our analysis will be based on classification models and the outcome of the launch is the variable that will be predicted.

EDA AND INTERACTIVE VISUAL ANALYTICS

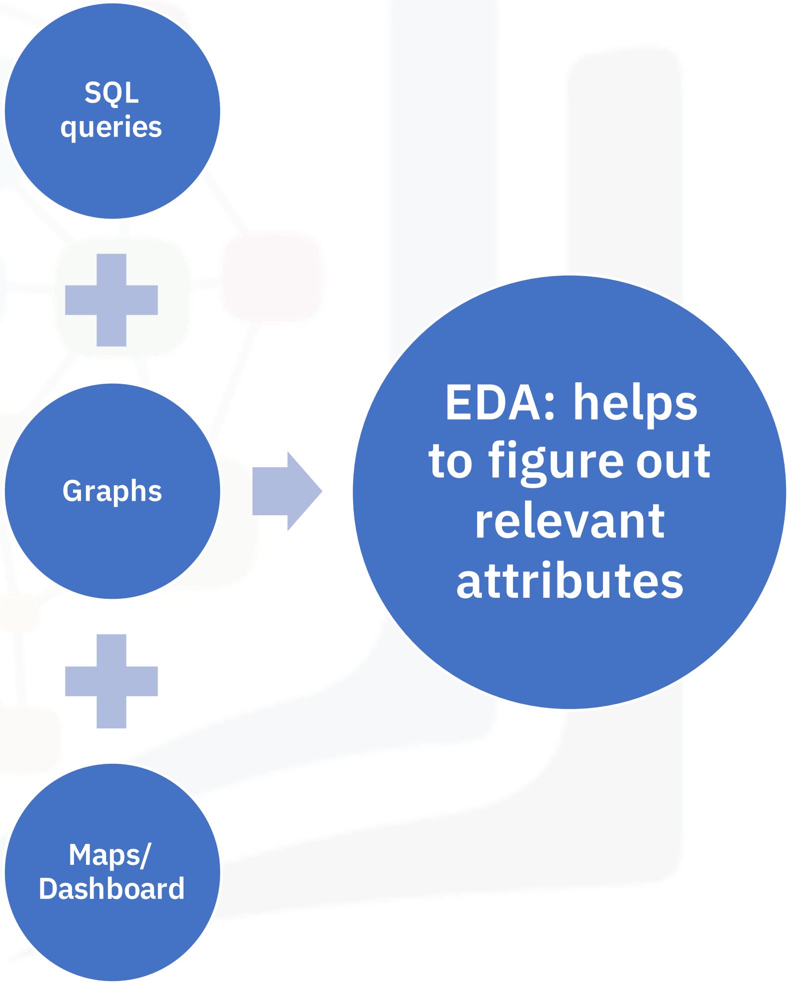
- The dataset we work with includes a record for each payload carried during a Falcon 9 mission into outer space.
- Since a version of the dataset was loaded into a database, a connection was stablished through a SQL extension.
- Using SQL, blank rows were removed from the dataset.
- We then executed SQL queries to perform some exploratory data analysis.
- The results obtained from the exploratory analysis are discussed in the Results section.
- We also performed EDA through graphs built with Matplotlib and Seaborn libraries. The main results are also presented in the Results section.
- Exploratory analysis allowed us to determine and/or confirm which attributes affect the success rate of landings.

EDA AND INTERACTIVE VISUAL ANALYTICS

- After figuring out which attributes are relevant, some of them needed to be prepared and/or transformed to be used in the machine learning models (Features Engineering).
- The categorical variables, like Orbit, were converted using one hot encoding.
- With the dataset containing only numbers, it is important to cast the entire data frame to type float64.
- After performing exploratory analysis using SQL queries and Matplotlib/Seaborn graphs, the process was complemented with interactive visual analytics and the construction of a dashboard.
- Interactive visual analytics enables users to explore and manipulate data in an interactive way, allowing them to find visual patterns faster and more effectively.
- Interactive data visualization, or dashboarding, can always tell a more appealing story.

EDA AND INTERACTIVE VISUAL ANALYTICS

- We used Folium and Plotly Dash to build an interactive map and dashboard to perform interactive visual analytics.
- The Folium maps were used to get more information about the launch sites, since launch success rate may also depend on the location of the initial position of rocket trajectories.
- In the dashboard application, the success rate for the landings carried out at each site and the effect of the payload mass are analyzed.
- The results obtained from the analysis of the maps will be presented in the section Results, together with the dashboard.

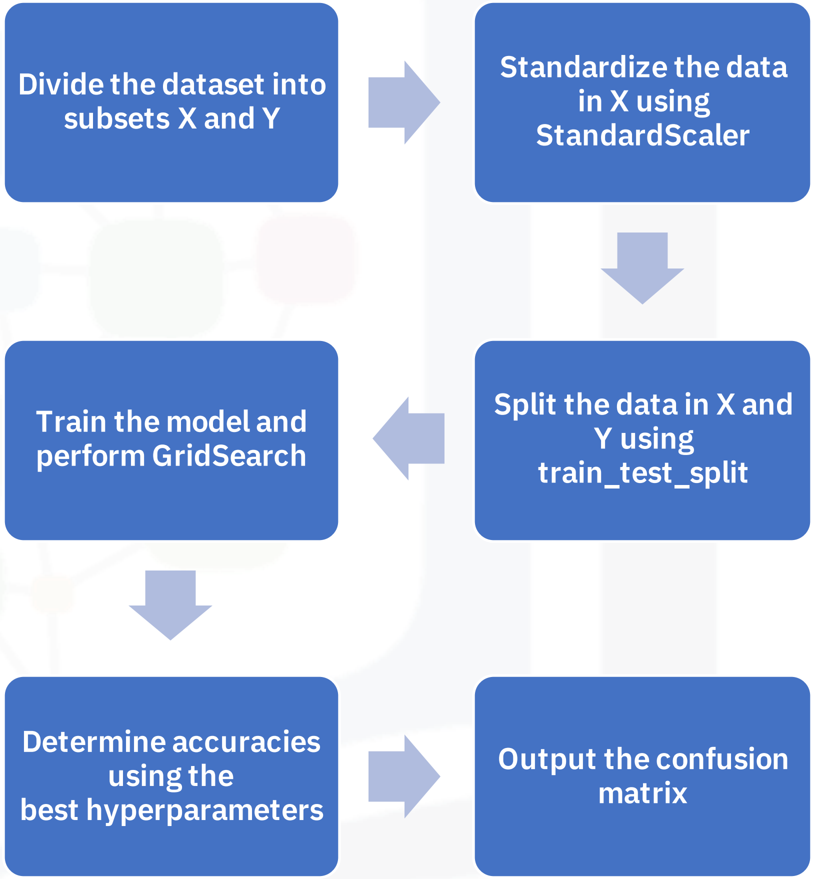


PREDICTIVE ANALYSIS

- The last part consisted of building a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully.
- The pipeline includes:
  - Preprocessing, where we standardize our data;
  - Train_test_split, which allows us to split the data into training and testing data.
- Before preprocessing, the dataset was divided into subsets Y and X, which represent the variables that we would like to predict (Class) and the independent variables (attributes), respectively.
- Both X and Y should be splitted into train and test sets (test_size=0.2).
- Four classification models were tested: Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree Classifier (DTC) and K-Nearest Neighbors (KNN).
- To train the models, we performed Grid Search to find the best hyperparameters.

PREDICTIVE ANALYSIS

- The best hyperparameters allow a given algorithm to perform best.
- Using the best hyperparameter values, the accuracies of the models are determined for the validation data (part of the training data) and test data.
- A discussion about the best model based on the accuracies and confusion matrices is presented in the next section.



Divide the dataset into subsets X and Y

Standardize the data in X using StandardScaler

Split the data in X and Y using train_test_split

Train the model and perform GridSearch

Determine accuracies using the best hyperparameters

Output the confusion matrix

RESULTS

- EDA – Visualization and SQL Results
- Folium Maps
- Dashboard
- Predictive Analysis (Classification)

EDA – VISUALIZATION AND SQL RESULTS

- Exploratory data analysis is the first step of any data science project.
- To better understand the data we work with, SQL queries were executed to answer some questions.
- We started by researching which launch sites were used in the space missions:

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

EDA – VISUALIZATION AND SQL RESULTS

- By investigating the dataset it is possible to discover which SpaceX customers are. We can, for example, display the total payload mass carried by boosters launched by NASA (CRS):

Booster_Version	PAYLOAD_MASS__KG	Customer	Booster_Version	PAYLOAD_MASS__KG	Customer
F9 v1.0 B0006	500	NASA (CRS)	F9 B4 B1039.1	3310	NASA (CRS)
F9 v1.0 B0007	677	NASA (CRS)	F9 FT B1035.2	2205	NASA (CRS)
F9 v1.1	2296	NASA (CRS)	F9 B4 B1039.2	2647	NASA (CRS)
F9 v1.1 B1010	2216	NASA (CRS)	F9 B4 B1045.2	2697	NASA (CRS)
F9 v1.1 B1012	2395	NASA (CRS)	F9 B5B1050	2500	NASA (CRS)
F9 v1.1 B1015	1898	NASA (CRS)	F9 B5B1056.1	2495	NASA (CRS)
F9 v1.1 B1018	1952	NASA (CRS)	F9 B5B1056.2	2268	NASA (CRS)
F9 FT B1021.1	3136	NASA (CRS)	F9 B5B1059.1	2617	NASA (CRS), Kacific 1
F9 FT B1025.1	2257	NASA (CRS)	F9 B5B1059.2	1977	NASA (CRS)
F9 FT B1031.1	2490	NASA (CRS)	F9 B5B1058.4	2972	NASA (CRS)
F9 FT B1035.1	2708	NASA (CRS)			

EDA – VISUALIZATION AND SQL RESULTS

- As seen in the previous table, there are several booster versions. We can check, for example, what is the average payload mass carried by booster version F9 v1.1:

AVG_PAYLOAD_MASS_KG
2534.67

- As mentioned earlier, rockets can land in different ways. Checking now the date of the first successful landing outcome in ground pad:

Date	Landing_Outcome
2015-12-22	Success (ground pad)

EDA – VISUALIZATION AND SQL RESULTS

- We can also investigate which booster had success landing in drone ship for a given payload mass range (mass greater than 4000 kg, but less than 6000 kg):
- Now we list the total number of successful and failure mission outcomes:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT 1031.2

Mission_Outcome	
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

EDA – VISUALIZATION AND SQL RESULTS

- Note that Mission_Outcome is a different attribute from Landing_Outcome.
- Checking the names of the booster versions which have carried the maximum payload mass:

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

EDA – VISUALIZATION AND SQL RESULTS

- By analyzing the dataset, we can also get information about different events in a specific year.
- Below, we list the records for the months of 2015 containing the failed landings in drone ships and their respective boosters versions and launch sites:

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

- In another example of collecting information for a specific period of time, we rank the count for each type of landing outcomes between 2010-06-04 and 2017-03-20 in descending order:

EDA – VISUALIZATION AND SQL RESULTS

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

- Next, we present the results of the graph analysis using Matplotlib and Seaborn.

EDA – VISUALIZATION AND SQL RESULTS

- First, we analyze how the FlightNumber, which indicates the continuous launch attempts, and Payload variables affect the launch outcome. The plot shows FlightNumber vs PayloadMass, and the colors represent the landing outcomes (Class 0 and Class 1). We see that, as FlightNumber increases, the first stage is more likely to land successfully (Class 1). Observing the first attempts, it seems the more massive the payload, the less likely the first stage will return.

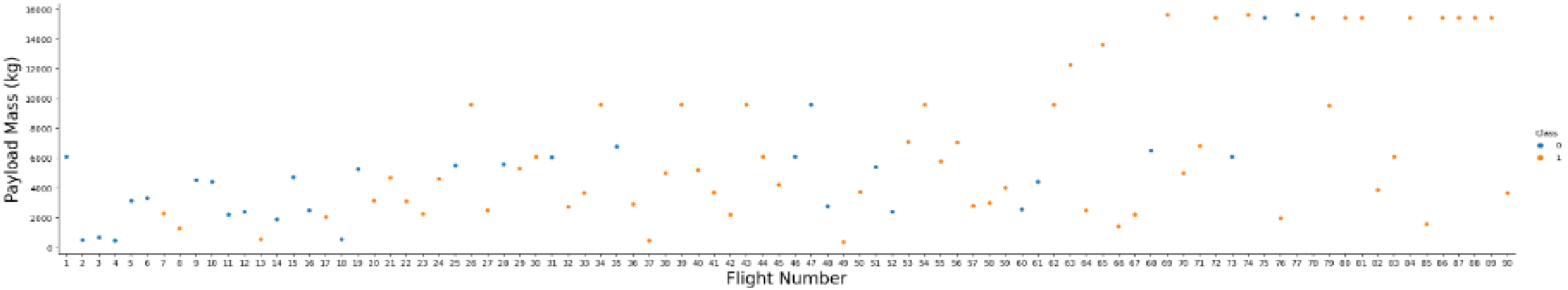


Figure 4: FlightNumber vs PayloadMass.

EDA – VISUALIZATION AND SQL RESULTS

- As we will see, the different launch sites have different success rates. Figure 5 shows a plot of FlightNumber vs LaunchSite for three different sites, with the colors once more representing the landing outcomes. We observe that, over time, releases began to be more widely distributed across different launch sites. It is also possible to observe the different success rates for different sites.

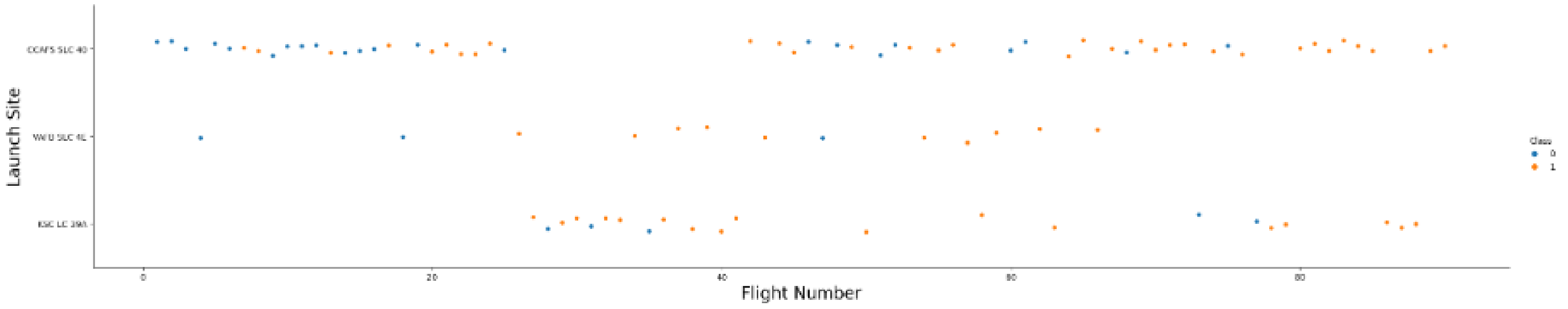


Figure 5: FlightNumber vs LaunchSite.

EDA – VISUALIZATION AND SQL RESULTS

- Now we want to know if there is any relationship between launch sites and the payload mass. Observing the PayloadMass vs LaunchSite scatter plot, we can see that most rockets carrying low payload masses (less than 8000 kg) were launched from the CCAFS SLC-40 site. We can also observe that there are no rockets launched for heavy payload mass (greater than 10000 kg) from the VAFB SLC launch site.

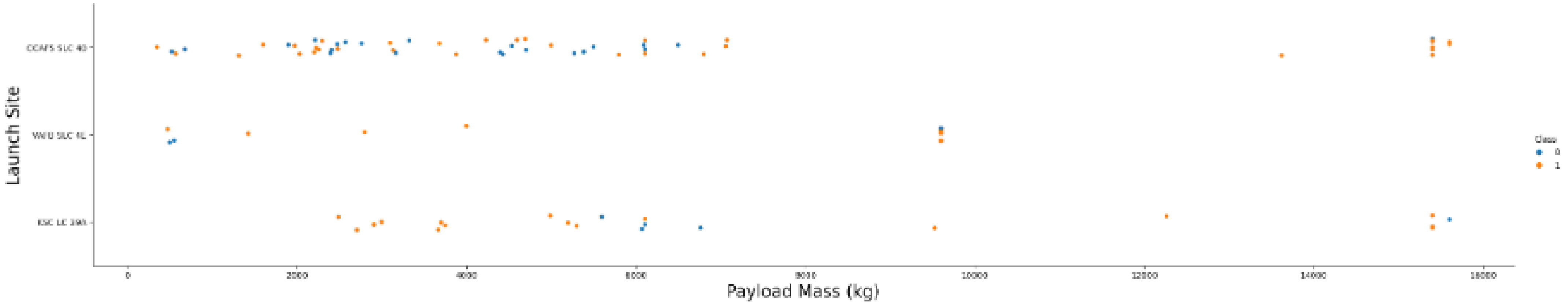


Figure 6: PayloadMass vs LaunchSite.

EDA – VISUALIZATION AND SQL RESULTS

- Next, we check if there are any relationship between success rate and orbit type. To search for this relationship, we create a bar chart for the success rate of each orbit. Four orbits have a hundred percent success rate: ES-L1, GEO, HEO and SSO.
- Curiosity: observe that the success rate for the International Space Station (ISS) orbit is around 60%.

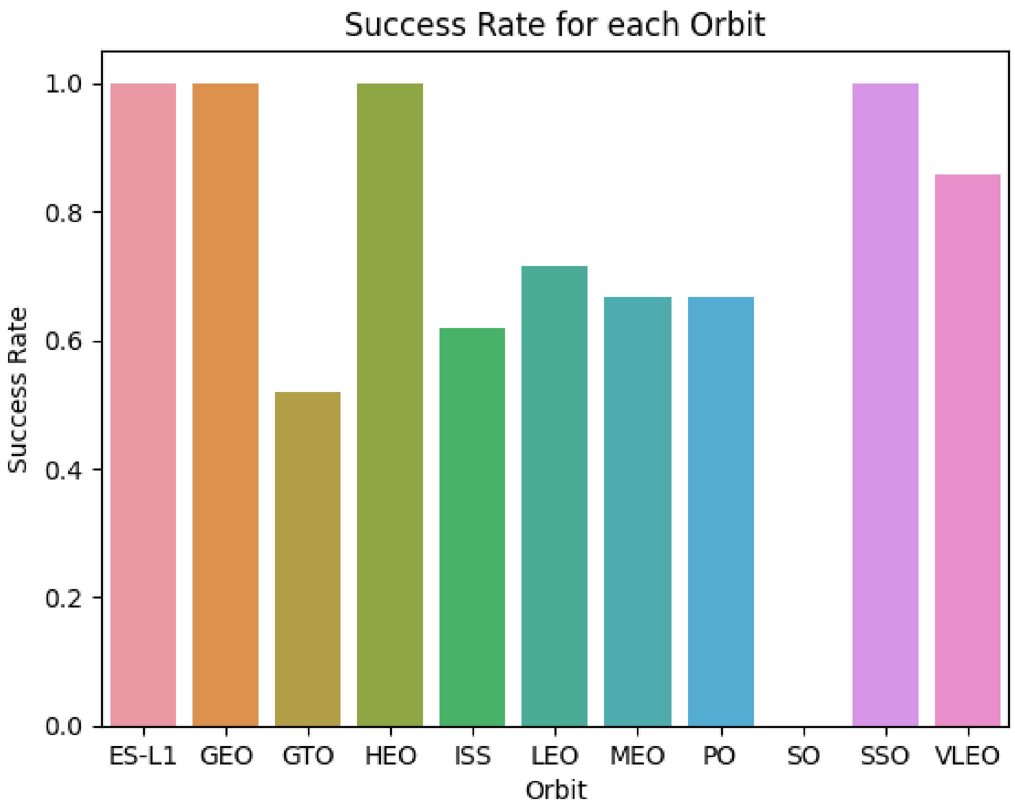


Figure 7: Bar chart for the success rate of each orbit.

EDA – VISUALIZATION AND SQL RESULTS

- Now we visualize the relationship between flight number and orbit type. Observing the FlightNumber vs Orbit scatter plot, we see that some orbits were only recently accessed, like SO and GEO (both have only one landing, which were unsuccessful and successful, respectively). Besides, while in the LEO orbit success appears related to the number of flights, for the GTO orbit such a relationship does not seem to exist.

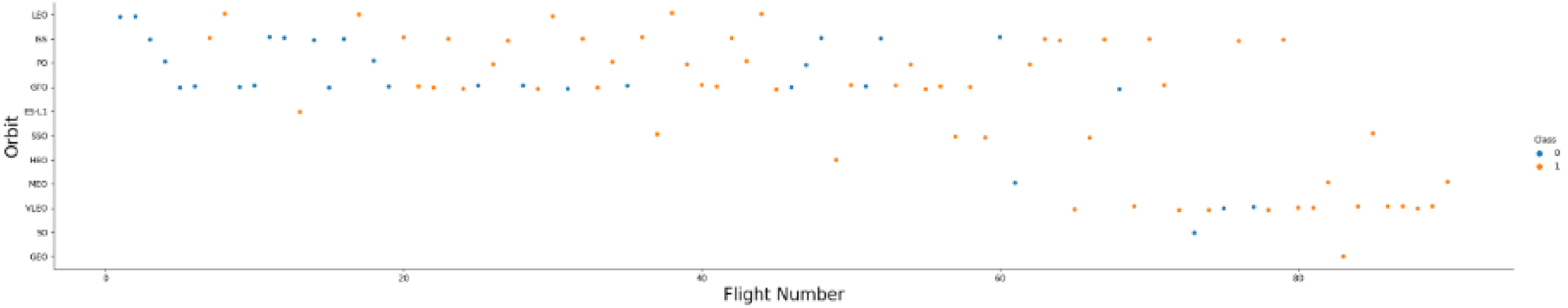


Figure 8: FlightNumber vs Orbit.

EDA – VISUALIZATION AND SQL RESULTS

- We can also investigate the relationship between payload and orbit type. Figure 9 shows the PayloadMass vs Orbit scatter plot. It is easy to notice that only a few orbits are used for heavy payload masses (greater than 8000 kg), like PO and VLEO. Besides, we can see that, from intermediate to heavy payloads, the rate of success for GTO orbit seems to be smaller than for other orbits, like LEO, PO and ISS.

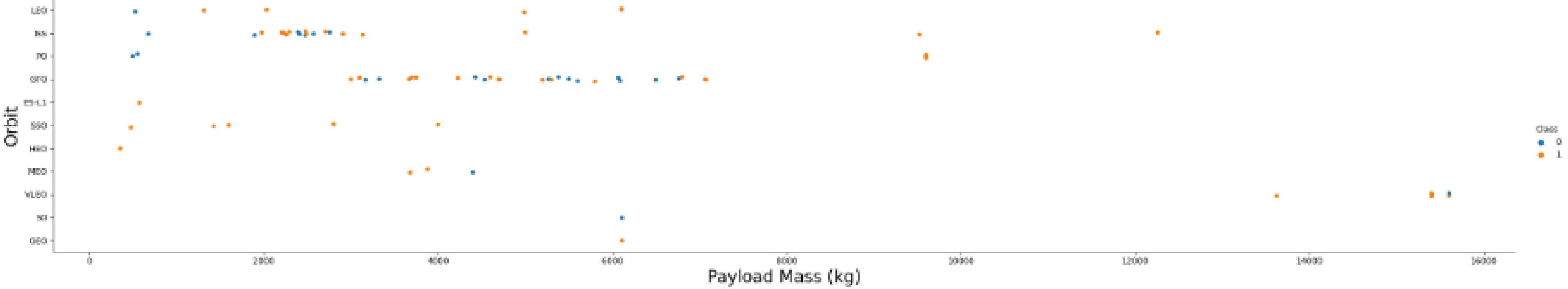


Figure 9: PayloadMass vs Orbit.

EDA – VISUALIZATION AND SQL RESULTS

- To observe the launch success yearly trend, we can plot a line chart with x axis to be Year and y axis to be average success rate. It is easy to observe that the success rate kept increasing from 2013 until 2020.
- From previous analysis, we obtain some preliminary insights about how each important variable affect the success rate. Now, we can start the selection of the features that will be used in success prediction.

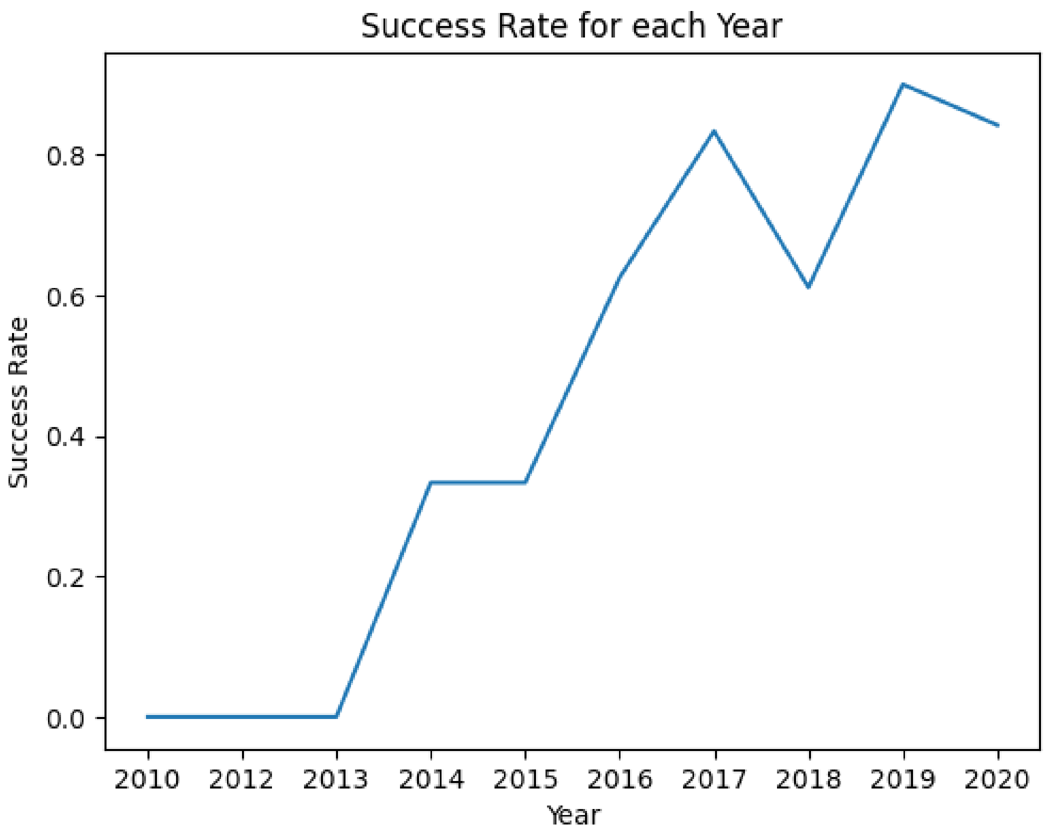


Figure 10: Year vs Success Rate.

FOLIUM MAPS

- It is already clear that the launch success rate may depend on many factors such as payload mass and orbit type. It may also depend on the location of a launch site, i.e., the initial position of rocket trajectories.
- Finding an optimal location for building a launch site certainly involves many factors.
- We now discuss the Folium maps. Here we focus on analyzing launch sites proximities.
- First, the locations of the launch sites are marked and their close proximities are investigated. Then, we can explore the map with those markers and try to discover any patterns. Finally, we should be able to explain the parameters to be considered when choosing an optimal launch site.
- Each site location was added on a map using the site's latitude and longitude coordinates.

FOLIUM MAPS

Launch Site	Latitude	Longitude
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

- As we can see, all the sites are located on the East and West coast.

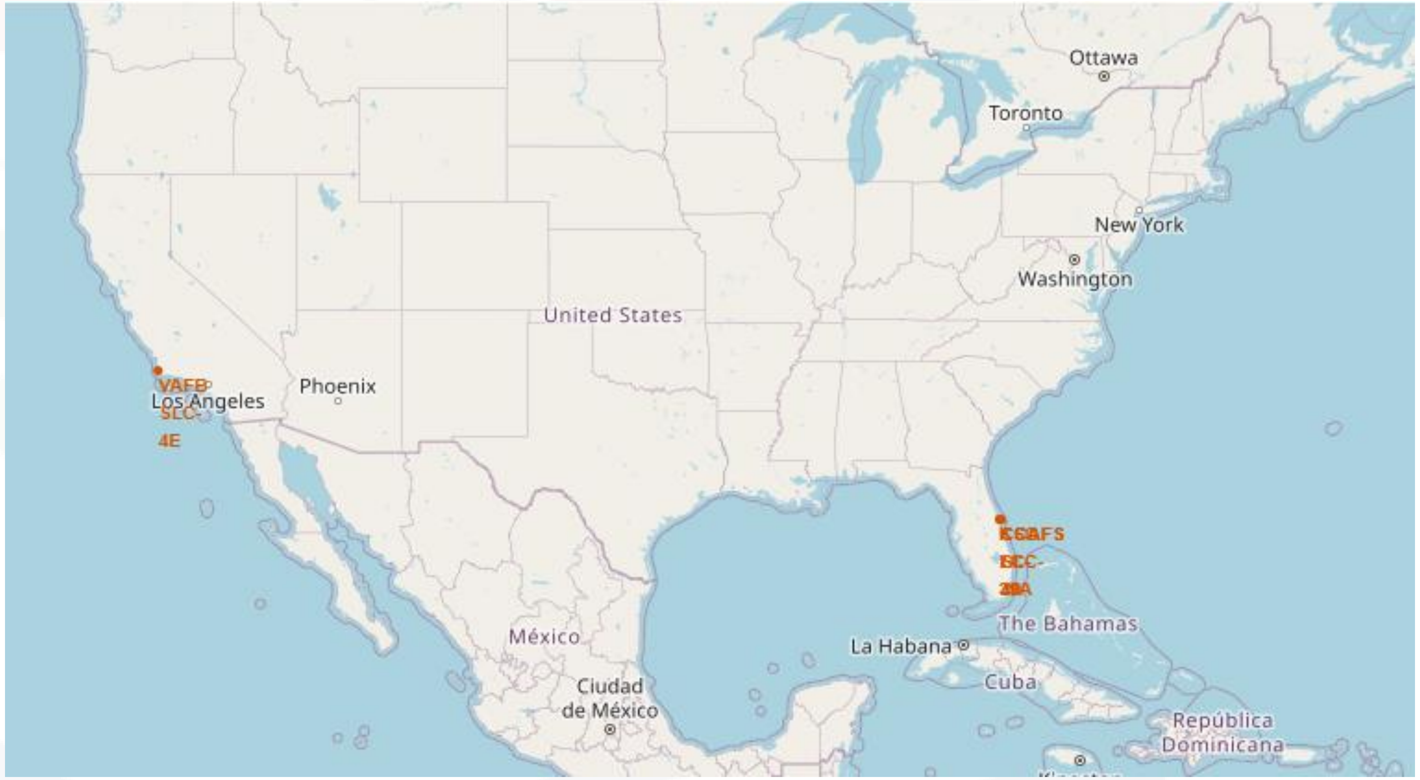


Figure 11: Distribution of the launch sites on the map of the United States.

FOLIUM MAPS

- The launch sites usually are close to the coast because, if anything goes wrong during the rocket's ascent, the debris would fall into ocean's waters, far from populated areas. For this same reason, launch sites are usually built far away from densely populated areas and busy highways, for example, as we will see.
- We can also observe that all launch sites are relatively close to the Earth's Equator. When a spacecraft is launched from a site near the Equator line, it can take advantage of the high rotational speed of the Earth (areas that are closer to the Equator spin around Earth's axis faster than areas closer to the poles).
- In the next slides, the maps are enhanced and the launch outcomes are added for each site, so we can see which sites have high success rate.
- Remember that Class=1 represents a successful launch (green marker), while Class=0 means the launch was failed (red marker).

FOLIUM MAPS

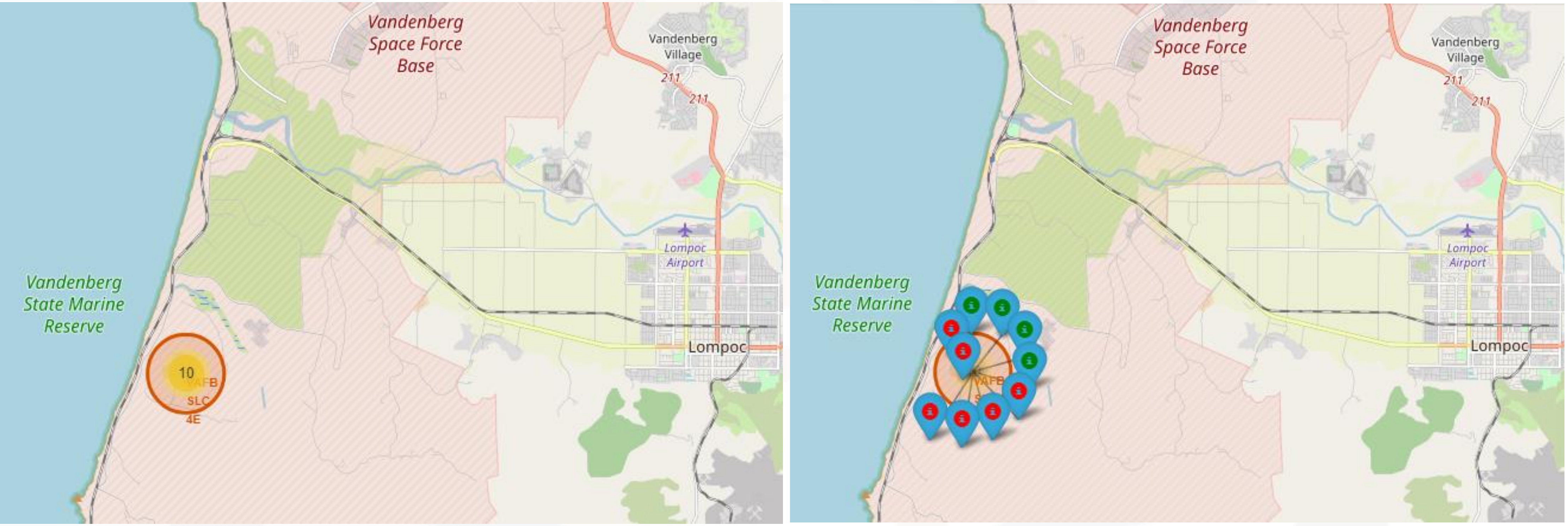

Vandenberg Space Force Base
Vandenberg Village
211
211
Lompoc Airport
Vandenberg State Marine Reserve
Lompoc
10
VAFB SLC 4E
Vandenberg Space Force Base
Vandenberg Village
211
211
Lompoc Airport
Vandenberg State Marine Reserve
Lompoc
VAFB

Figure 12: The first figure shows the total number of launches for the VAFB SLC-4E site (10), while the second shows the markers representing the successful landings (green) and the failed ones (red).

FOLIUM MAPS

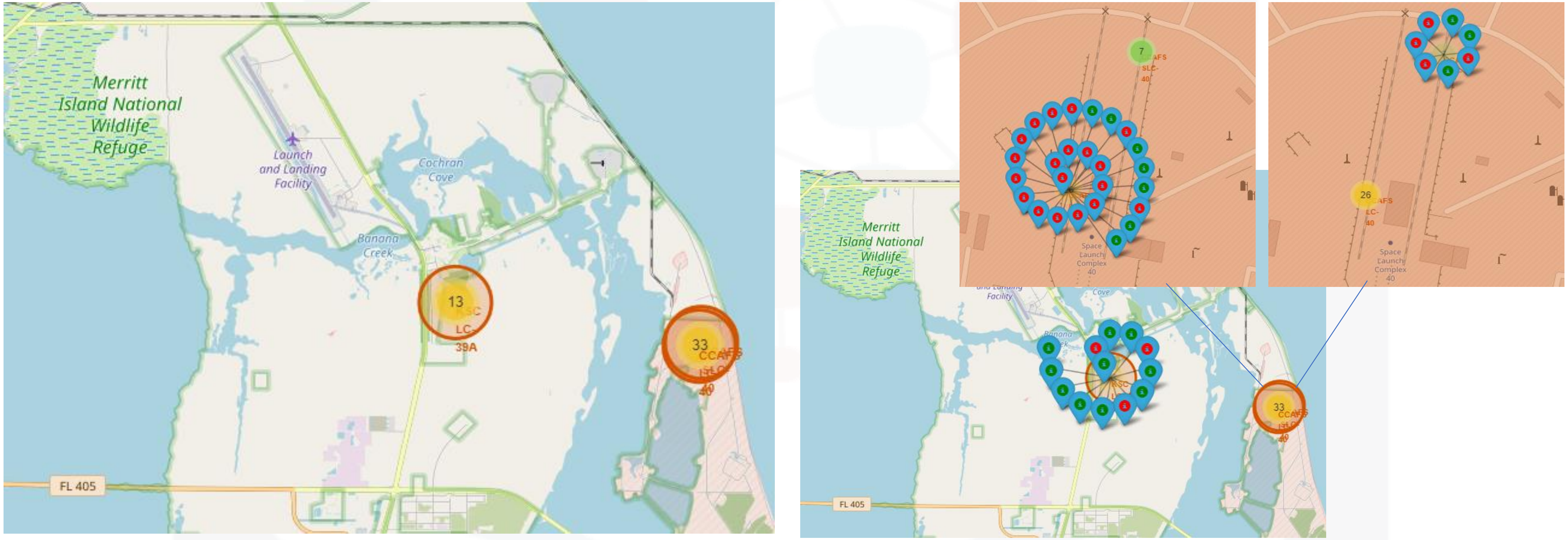


Figure 13: The first figure shows the total number of launches for the KSC LC-39A (13) and CCAFS LC-40/CCAFS SLC-40 sites (33, where 26 launches occurred in the first site, and 7 in the second), while the second shows the markers representing the successful landings (green) and the failed ones (red) for each site (CCAFS LC-40 and CCAFS SLC-40 sites are highlighted).

FOLIUM MAPS

- As mentioned before, launch sites are usually built far away from densely populated areas for security measures. It is possible to see, in Figures 12 and 13, that all the sites considered in this study are built in isolated locations, but close to the coast.
- We now explore the proximities of one of the launch sites, the VAFB SLC-4E site (Figure 14). Analyzing the map, it is observed that the closest city, Lompoc, is more than 14 km away. We also did not identify any highways close to the site.
- Looking closer, we observe that the coastline is at a distance of more or less 1 km, as well as a small road.
- Although launch sites are mostly built in isolated areas, we also observe a section of the Coast Line railroad approximately 1 km away.

FOLIUM MAPS

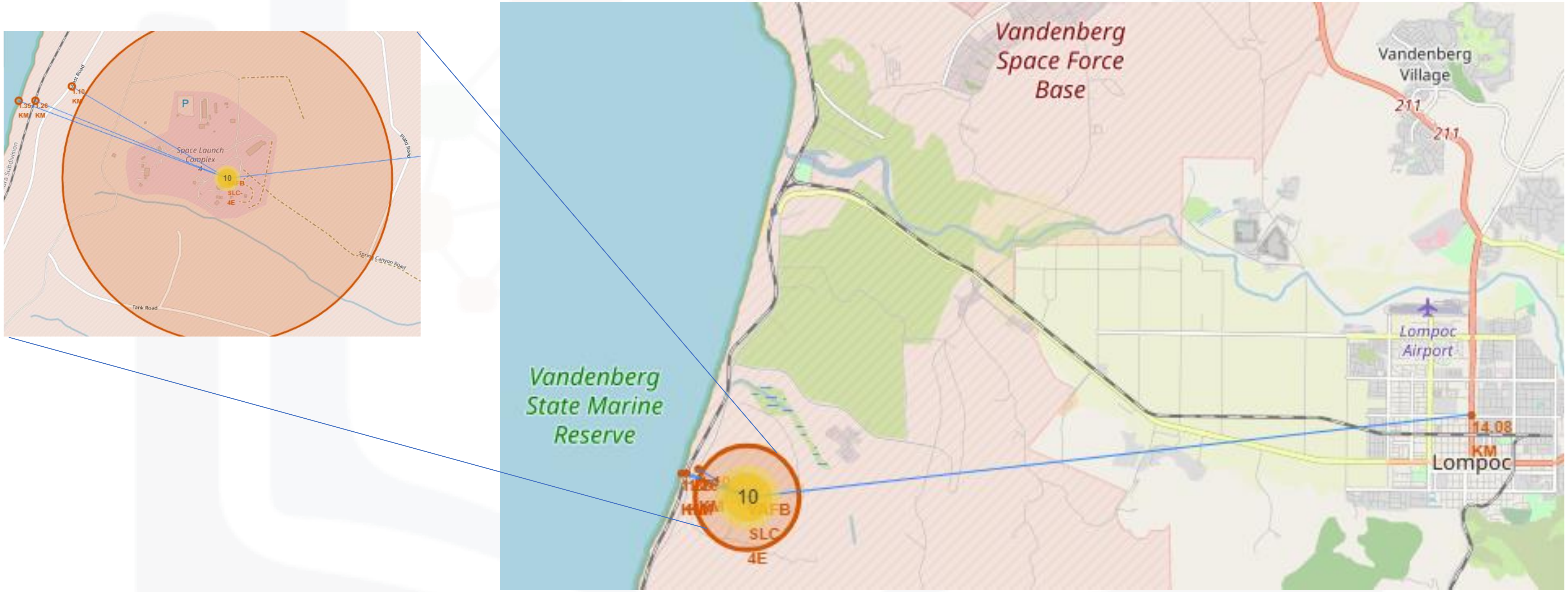


Figure 14: The figure shows the distance from the VAFB SLC-4E site to its closest city, Lompoc. The featured image shows the vicinity of the launch site.

DASHBOARD

- The next step is discuss the results evidenced by the dashboard. We can choose to analyze all launch sites at once, or one site at a time.
- We start by analyzing the total success launches by site (Figure 15). It is observed that the most successful landings are related to the site KSC LC-39A.
- Observing the scatter chart, we see that there are very few successful landings for heavy payloads (mass greater than 6000 kg) and that the range with the highest launch success rate is 3000-4000 kg.
- Comparing the boosters, the version with the highest launch success rate is the FT (we can also include B5, but it has only one landing related to it).
- Figures 16-19 show the statistics for each site. It is easy to notice that the success rate depends on the launch site, with the KSC LC-39A site being the most successful facility. In fact, it is the only one with more successful landings than failed ones.
- We can also notice that the VAFB SLC-4E is the only one with a successful landing of a very heavy payload (mass around 10000 kg).

DASHBOARD

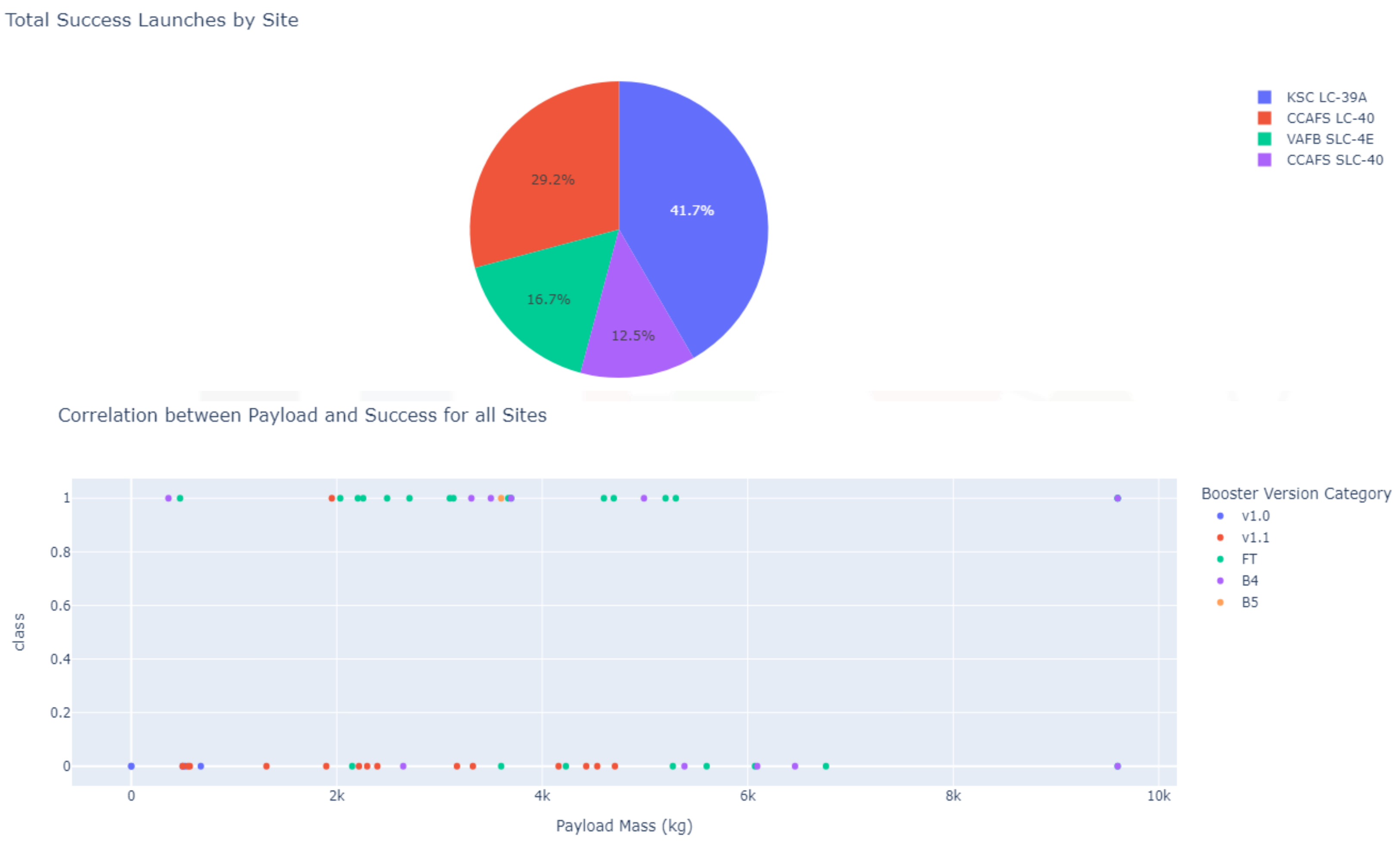


Figure 15: Pie chart for the total success launches by site and scatter chart showing the correlation between Payload Mass and successful landing for all sites.

DASHBOARD

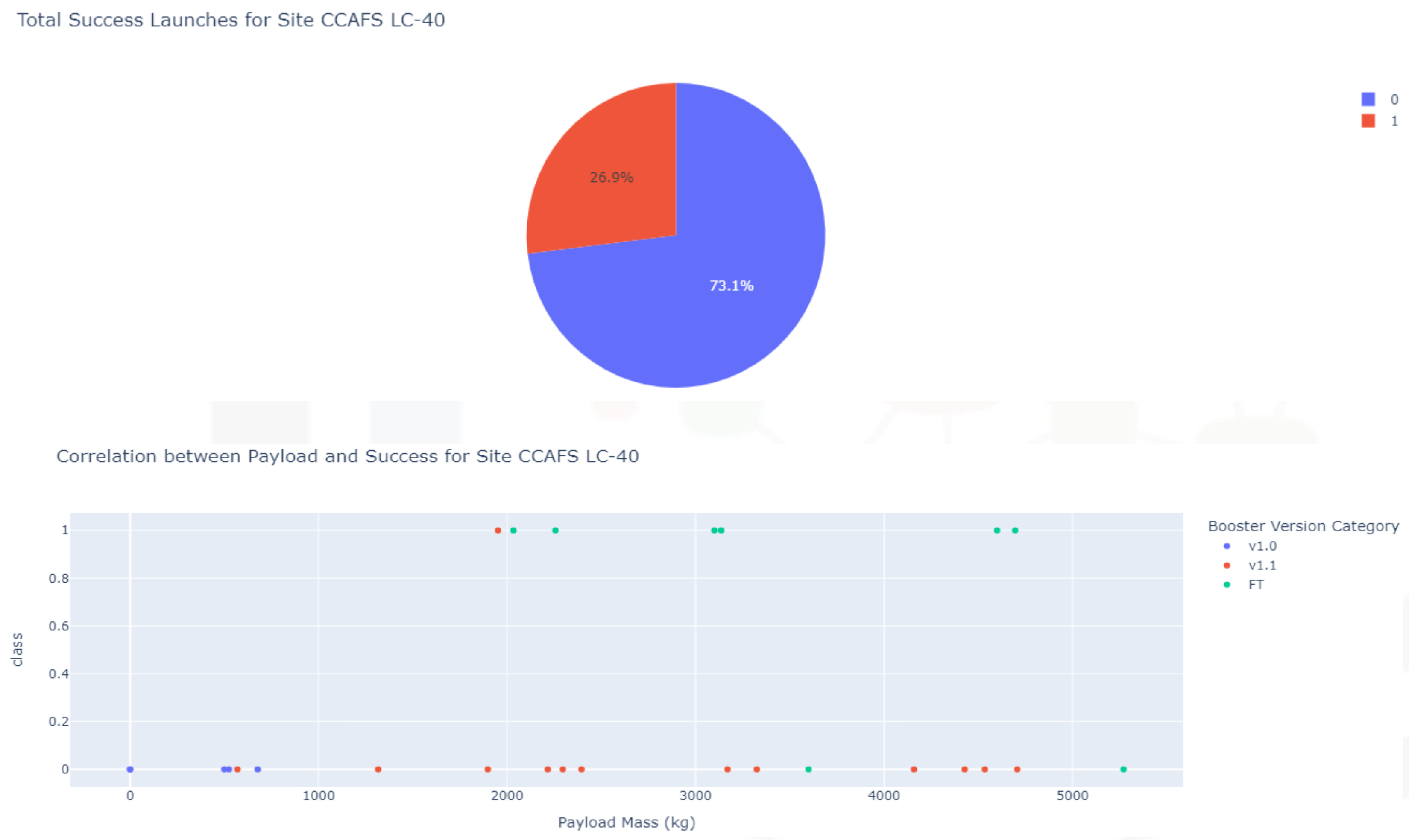
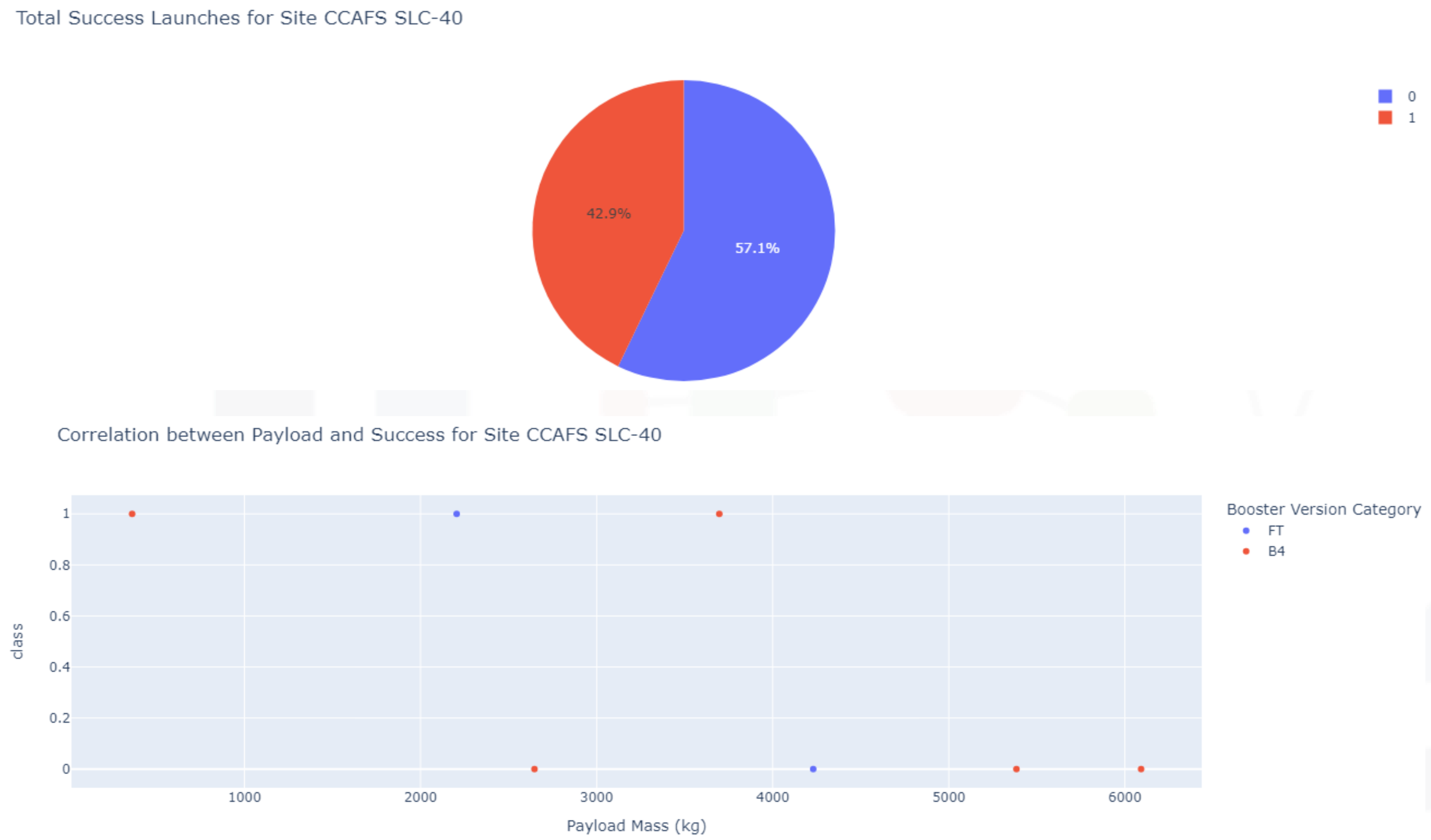


Figure 16: Pie chart for the total success launches and scatter plot showing Payload vs Class for site CCFAS LC-40.

DASHBOARD

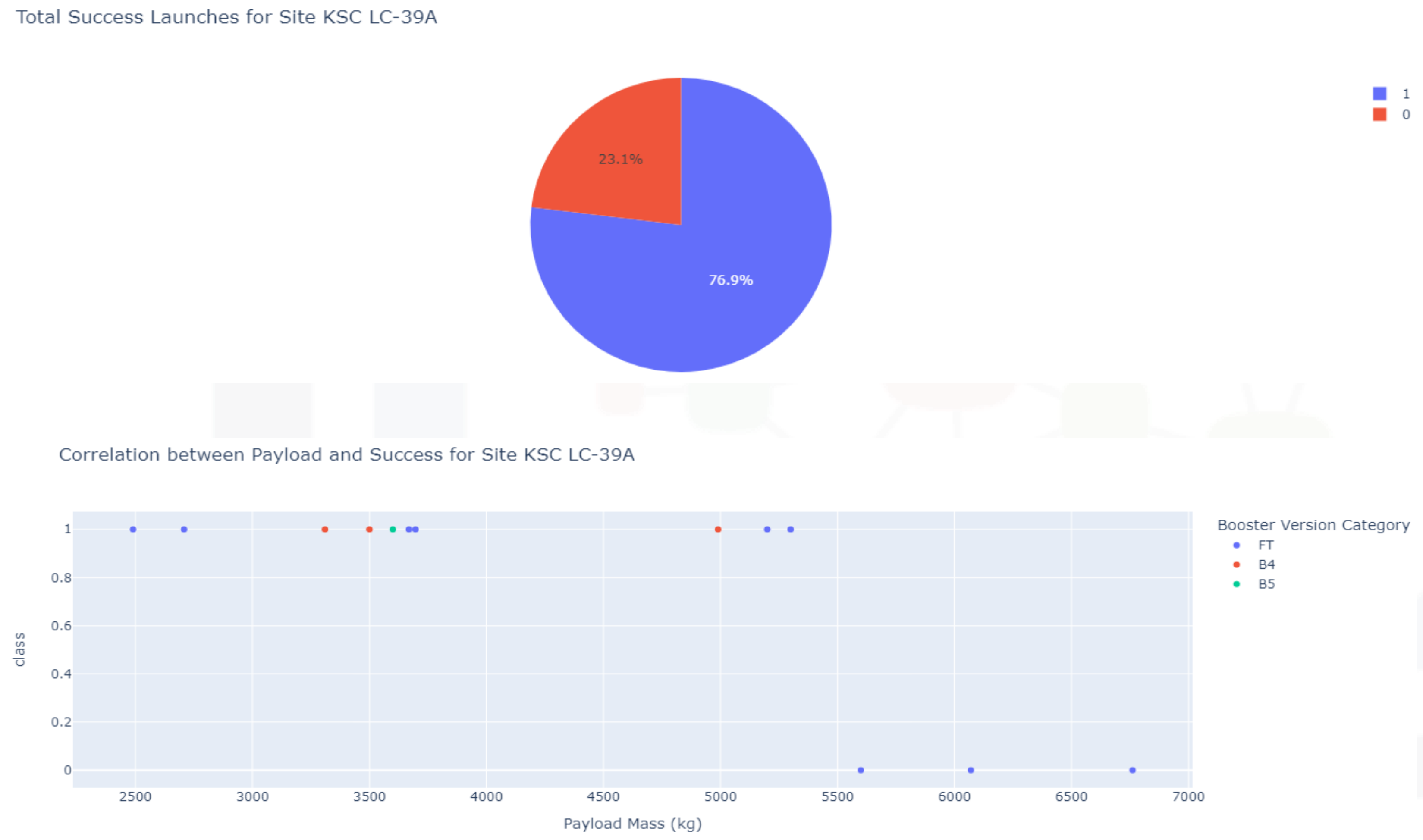


Total Success Launches for Site CCAFS SLC-40

Correlation between Payload and Success for Site CCAFS SLC-40

Figure 17: Pie chart for the total success launches and scatter plot showing Payload vs Class for site CCFAS SLC-40.

DASHBOARD

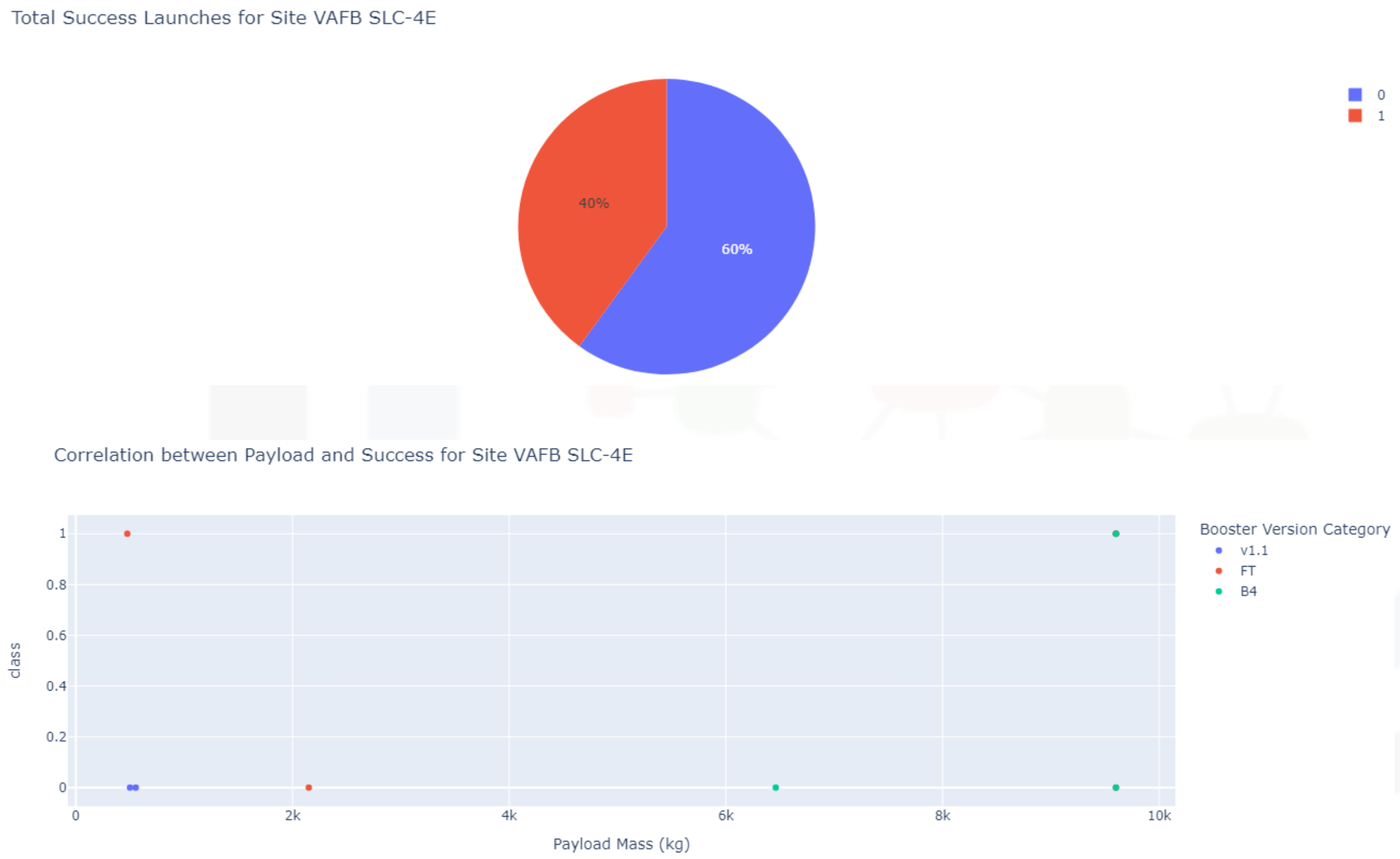


Total Success Launches for Site KSC LC-39A

Correlation between Payload and Success for Site KSC LC-39A

Figure 18: Pie chart for the total success launches and scatter plot showing Payload vs Class for site KSC LC-39A.

DASHBOARD



Total Success Launches for Site VAFB SLC-4E

Correlation between Payload and Success for Site VAFB SLC-4E

Figure 19: Pie chart for the total success launches and scatter plot showing Payload vs Class for site VAFB SLC-4E.

PREDICTIVE ANALYSIS (CLASSIFICATION)

- From the results and analyzes presented so far, it became possible to determine which attributes are most strongly correlated with a successful landing.
- These attributes were used as the independent variables in our machine learning models, whose objective is to predict if the first stage will successfully land.
- We used 83 variables, where some of them were generated due to the transformation of categorical variables (one hot encoding).
- After standardize the data and split it into training and testing sets, we trained the models and performed Grid Search, which allowed us to find the best hyperparameters.
- Four models were trained: Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree Classifier (DTC), and K-Nearest Neighbors (KNN). Next we show the confusion matrices and the accuracies for the models.

PREDICTIVE ANALYSIS (CLASSIFICATION)

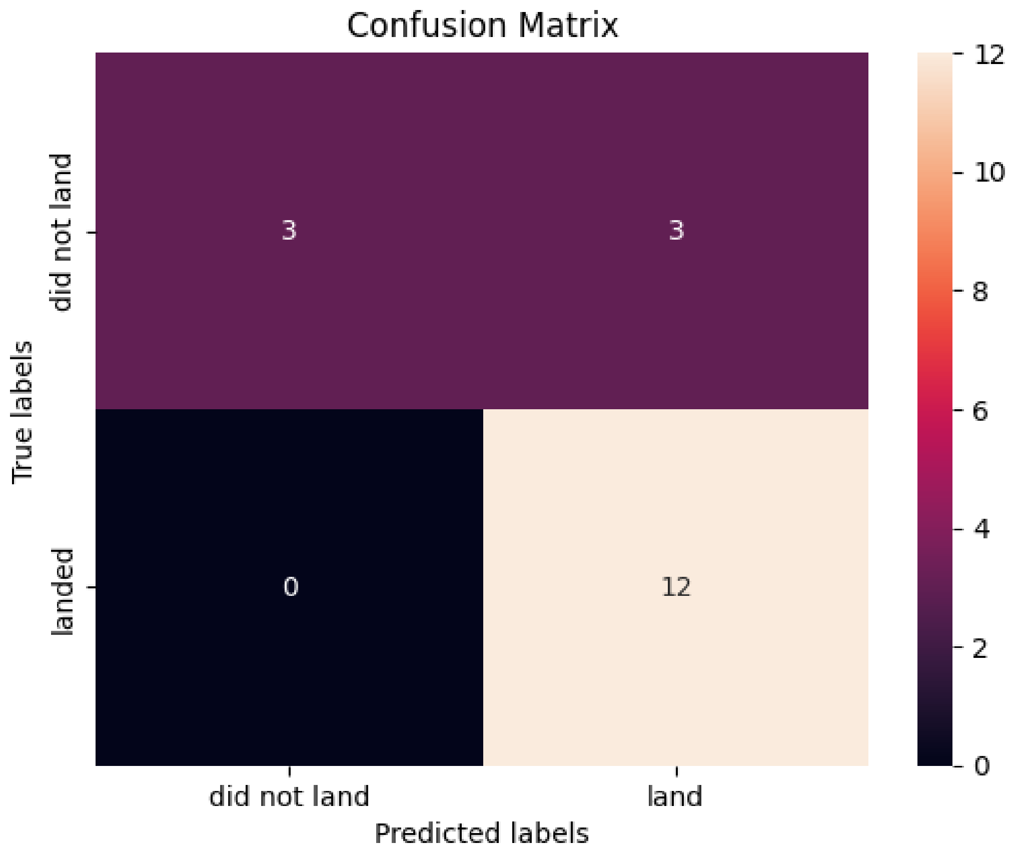


Figure 20: Confusion matrix for LR model.

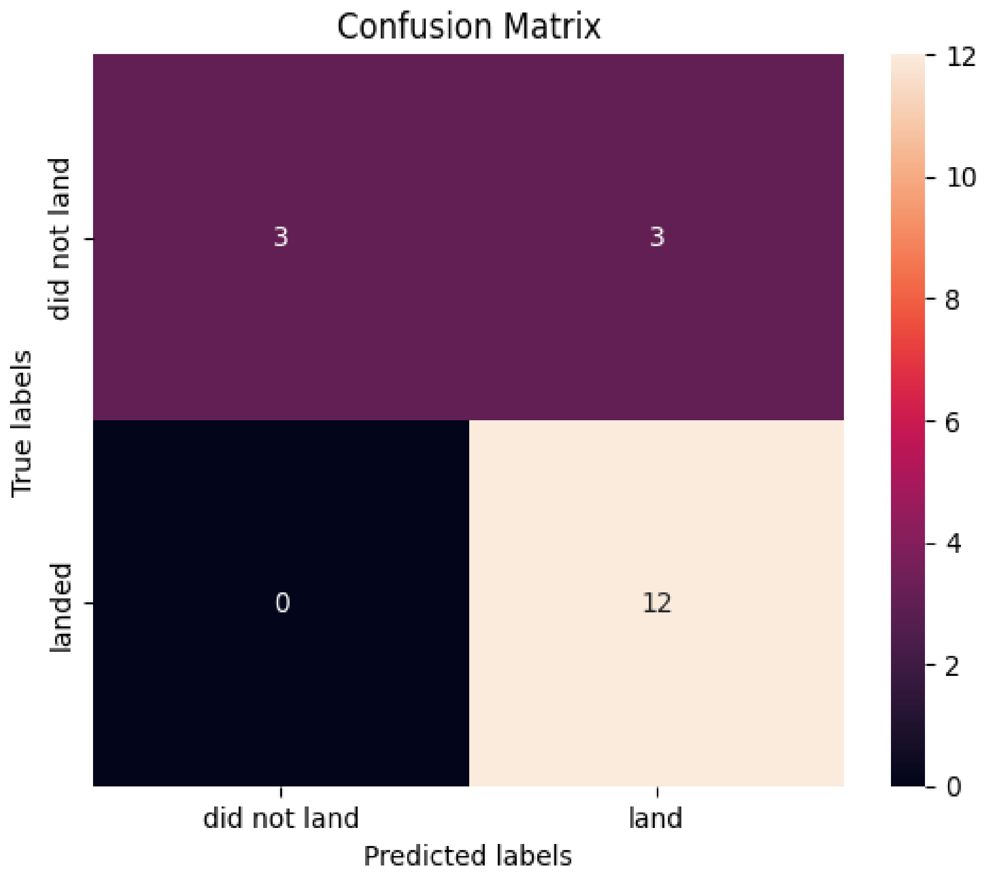


Figure 21: Confusion matrix for SVM model.

PREDICTIVE ANALYSIS (CLASSIFICATION)

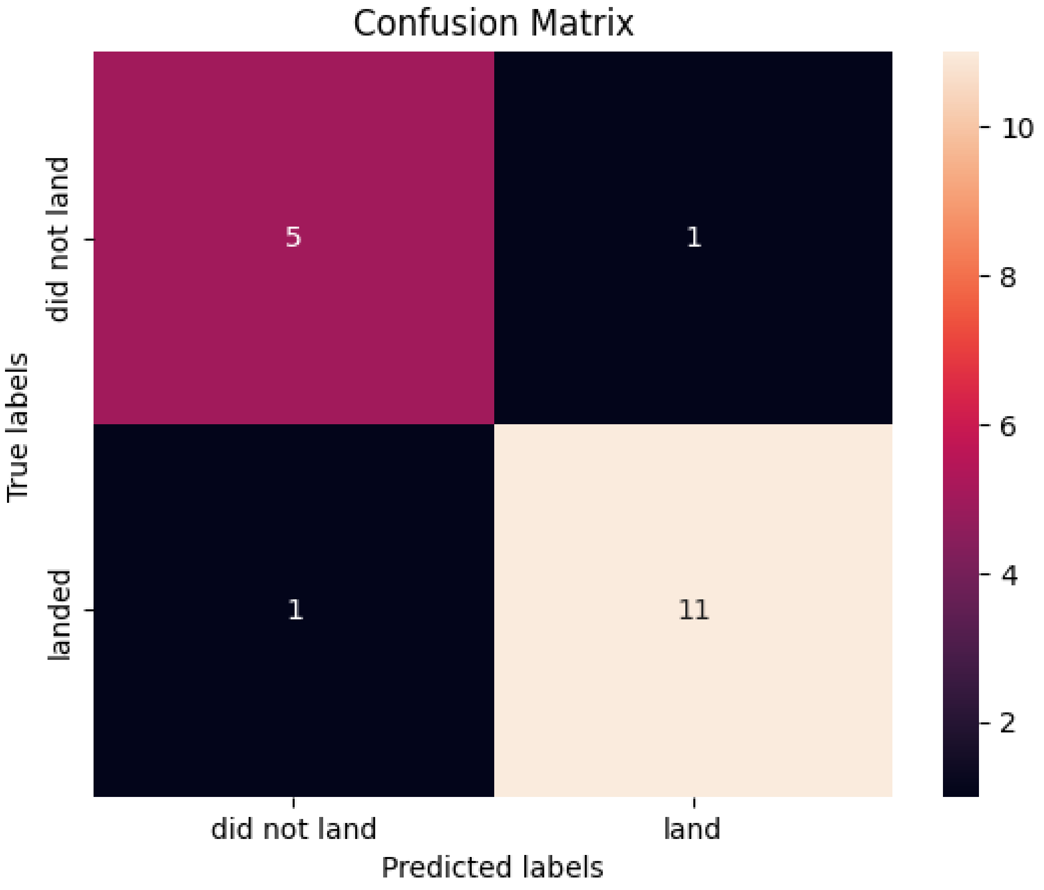


Figure 22: Confusion matrix for DTC model.

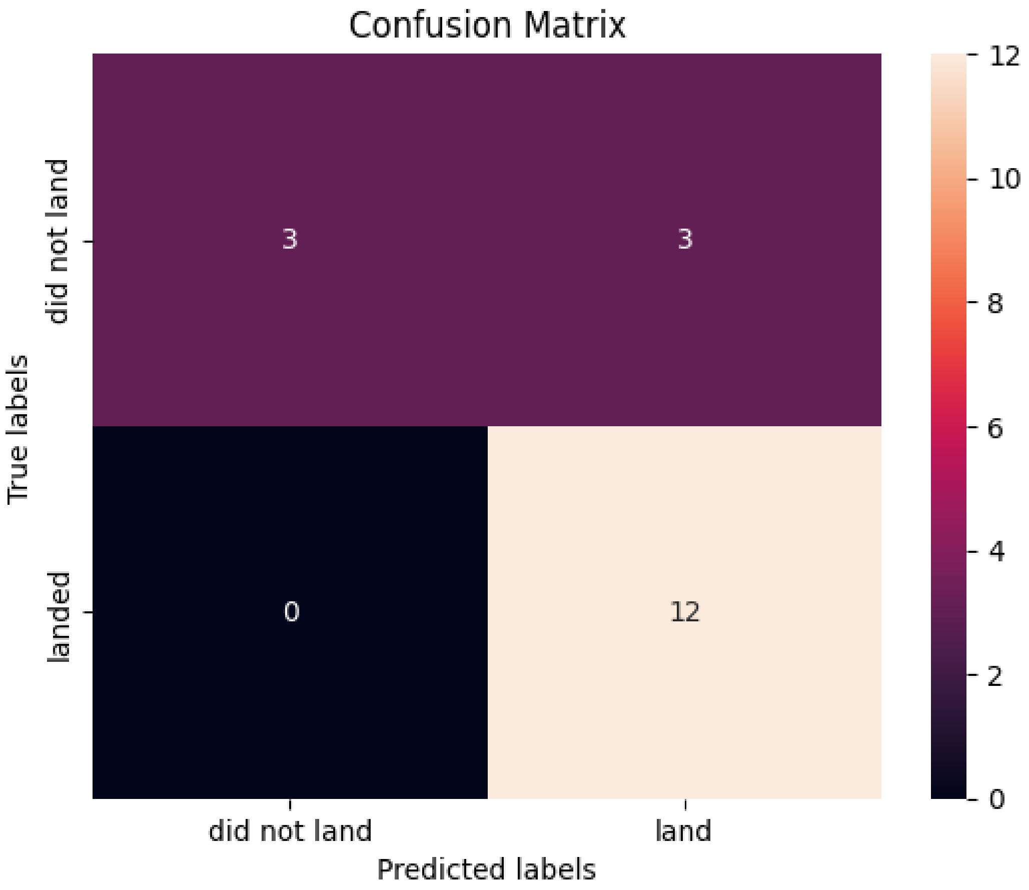


Figure 23: Confusion matrix for KNN model.

PREDICTIVE ANALYSIS (CLASSIFICATION)

- We observe that LR, SVM and KNN models have the same confusion matrix. The models can distinguish between the different classes, and the major problem is false positives. However, this is a minor problem for the DTC's confusion matrix. As we can see, the DTC model is also very efficient in distinguishing between the different classes.

Model	CV Score	Test Score
LR	0.846	0.833
SVM	0.848	0.833
DTC	0.877	0.889
KNN	0.848	0.833

- The table shows the accuracies of the models determined with the validation data – Cross Validation (CV) Score – and test data – Test Score. The values of the scores are very close for the four models, with the DTC model being the best one by a small difference. At the same time, the confusion matrix for the DTC model confirms that it appears to be the best for predicting the landing outcome.

CONCLUSION

- In the present report, we discussed how data science can be used to predict whether the first stage of the Falcon 9 rocket will land successfully.
- We have seen how important it is to obtain data from different sources so that the final dataset is as complete and informative as possible.
- The first step in the data science process is exploratory data analysis. This phase is extremely important in preparing the data that will be used in machine learning models, both to adapt them to the use of the models (one hot encoding, for example), and to know which attributes are relevant for the prediction.
- There are several ways to carry out exploratory analysis. In the present work we used SQL queries, graphics and maps to analyze the data. We also built a dashboard which enabled interactive visual analysis.
- All classification models tested have high accuracy, indicating that they are all efficient in predicting the landing result. However, the DTC model proved to be slightly better at distinguishing between classes (success or failure).

APPENDIX

- All the notebooks and codes used in this work, as well as a copy of the present report, can be found at https://github.com/LRios22/Final_Project