

IBM Applied Data Science Capstone Project

Presented By:

Soumesh khuntia

10th Nov, 2023











Results



Conclusion



Appendix

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly & Dash
- Predictive Analysis (Classification)

Summary of all results

- Exploratory Data Analysis
- Interactive Analytics Demo in screenshots
- Predictive Analysis results



Introduction

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Methodology



Data Collection Methodology

- Using SpaceX Rest API
- Using Web Scraping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL Performed interactive visual analytics using Folium and Plotly Dash Performed predictive analysis using classification models - Building, tuning and evaluation of classification models to ensure the best results



Data Collection

This process involved a combination of API requests from SpaceX **REST API** and **Web Scrapping** data from a table in SpaceX's Wikipedia entry.

We had to use both data collection methods to get complete information about the launches for a more detailed analysis.

Data Columns obtained from **SpaceX REST API**:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns obtained by using Wikipedia Web Scrapping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection SpaceX API



Requesting rocket launch date from SpaceX API



Decoding the response content using .json() and turning it into a dataframe using .json_normalize()



Requesting needed information about the launches from SpaceX API by applying custom functions



Constructing data we have obtained into a dictionary



Creating a dataframe from the dictionary



Filtering the dataframe to only include Falcon-9 launches



Replacing missing values of payload mass columns with calculated .mean()



Exporting the data to CSV

Data Collection – Web Scrapping

1

Requesting Falcon-9 launch from Wikipedia 2

Creating a **Beautiful Soup** object from the HTML response

3

Extracting all column names from the HTML table header



Collecting the data by parsing HTML tables

5

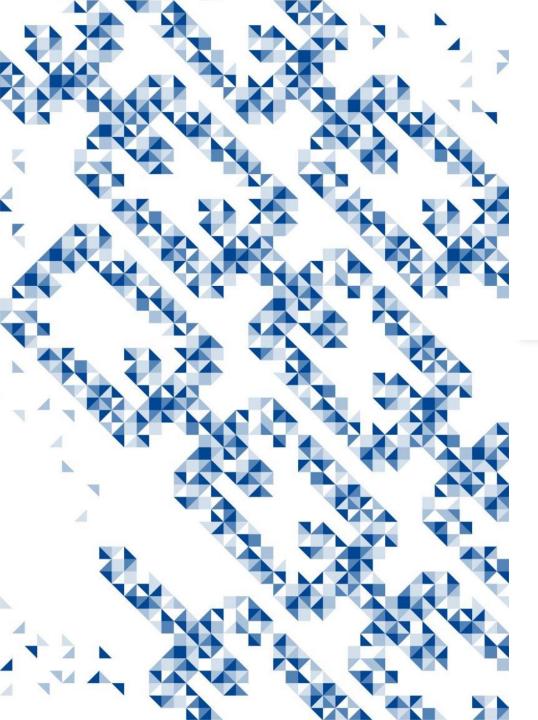
Constructing data we've obtained into a dictionary

6

Creating a dataframe from the dictionary



Exporting the data into CSV



Data Wrangling

- Perform Exploratory Data Analysis and determine training labels
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Creating a landing outcome label from **Outcome column**
- Exporting the data to CSV

EDA with Data Visualization



Charts were plotted: Flight Number vs.
Payload Mass, Flight Number vs. Launch
Site, Payload Mass vs. Launch Site, Orbit
Type vs. Success Rate, Flight Number vs.
Orbit Type, Payload Mass vs Orbit Type and
Success Rate Yearly Trend Scatter plots show
the relationship between variables.

If a relationship exists, they could be used in machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value. Line charts show trends in data over time (time series)

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- 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 when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship 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 failed landing outcomes in drone ship, their booster versions and launch site names for the months in year
- Ranking 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





Interactive Visual Analytics using Folium, Plotly and Dash

Markers of all Launch Sites: - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location. - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts. Coloured Markers of the launch outcomes for each Launch Site: - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates. Distances between a Launch Site to its proximities: - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

GHL THE SEE TO 576 151.18 061 128 CV 0,60 060,000 270 Ale 100 17 10.01 204.01 53.10. 19.69 80.11 Nº0 CAL 168 .02 T'SE SE ST. ST. ST. SE . Co 3. .45 0000 605 22.42 973 620 10. 20.6 057.98 3. 20 260 150 40°.24 585.01 864 114 13°. 200 10. 25 170 1.12 161.06 81.69 -co.cs B. 0. 30.932 Sel 85 O STAN 20 cho 18/ 196.68 076.46 ANS, c639.11 19 75% 188 1. 101.00 100 21.010 324 15,500 06.12 0,12 75.21 ·c> 100,00 63.841 A15.08 CC 178°69 60°, YY 00. की. पीने ch 969 121.08 0. 27.27 A CELEBERT OF COLOR OF THE OFFICE OFF .00 00.00 540 40.00 01% 13.601 0. .80 c30.0c3 16° 25' 111.00 00.00 250.06 064 Col 000 10.00 to 10. 5.10.16 10.01 3. 40°-24 c69.01 2007 7.c50 OCT. ट्टिंट ट्टिं . 16,000 86° SA 81.69 50 OG . 81. ANCS V 92.908 32 · (2) 861.17 Olyter ter at 16.198 .00 3. 139.cg/ 15 .00 CS 026 1.57 1.69 ° COL 200 80.08 . 60°. . 50 32.029 VV. 75°,01 196 10.102 250 .18 500.65 38.70. AS cc0. V6 012.68 100 11000 MY 59.3A 121.08 10.01 1. 46.96 29,131 15.00 80 ट्युर वृद् ्रति । 12.1201 6 2005 c/c.18 ASSOCIAL SOCIAL 8. 123.72.036 160.01 W11.08 10.00 812.07 16.01 C16.V1 0 3. 012.21 0,40 1997 120° 6 180° 60° 57.345 2. वद्धः वि 0,7% So. 19.548 ٧٥٠. 0.088 851.14 14. 888 710 .00 OLYCE VEST CES 11000 Dr. 125 .05 000 26.29.111 7.65 1. 21.664 698.66 100 V68 78 .1 19.136 200 550 AQ 736 1. 25. CO. 10.01 NI .0 112.05 20.32.11 34.609 46.95 .08 VV. 188 66. 10 5. 10 1. 10 20° 5 986 000 920 1.00 38.13 10. 16.372 217 150.07 200 N 901 60.15 6.129.72 2.90.191 844 3. NOW. 30.18 69.662 186 0,7% So. 84.859 2504 988 66.69 1689 81.91 16.181 203 000 83.101 45.794 NO.989 77.916 1 10. 505 97.891 80.500 NY 10,01 5 80.11 81.382 69.306 10. 64.045 Dr. 3.55 All 68.971 36.732 34. 3 1754 965 10. 500 251 13.013 15.158 10.0% 28.1341 2.920 150 8.38.71 18.19.188 BB 181 15.278 22.211 S. 140 201 dr. 52.025 6h. 65.634 18.988 1.283 160 13.991 10. 13. 16. 20 V.205 84. 12.20 \$. 84. NA NA. 978 14.95 94.833 450 66.578 15.991 D. 99.683 39.889 37.40 7.474 N. 62. J. 8.2.88 150.878 15.85. VOV 77.355 68.971 66.201 51.336 25.500 1,16.13 31.27 19:39:48 19:39:48 94.86 18.0 V.003 94.883 62 12.920 10.00 81.61 21.00.139 , 25° C 92.902 68,44 15.278 10.00 16.693 18.5TA 86.11 52.025 10. 10.086 18.98 200 8.20.12 11 .ch / 12. of 51.919 06.921 80 . AS 82.623 63.65 82.89A N.502 3. 29. 1. 20°. 53.026 3. 45.00 63.81 0.108 81.293 83. 43. 43. 184 83. 43. 43. 184 30. 67.307 12.014 क्षि. रुवेन 16.911 39.88g 501 54.52.46 S. 11.819 VY. 30 38.161 63.02.179 5 14.601 13.601 or. 17.653 OV. 10.010 94.550 %. VO. 140 81.01 2.10.65 13.03 12. 13.97 23. M. MA ST 81.315 NO.38 57.322 18.003 31.18 F61.5% 65 65 57 A. OV. 16.00 19.198 31,138 10.316 10.021 B. 38. 0 92.908 432.098 16.000 15.08.07 17.062 B. 51.919 85°. 10.532 63.575 63.81 0.109 10.105 60. 1.7.7. 1/2. 37.63.61° 61.75.65.76° 61.60° 23.180.130 30. 1717 60.37A 143 36.561 1.03 VV 63.032 22. 150.3A 00 24.584 21.90 38.161 63.03.13.03.23.15 13.601 15.00 VV. WJ-054 91.5TA 050.050 2.37,389 2. 84.555 1.31.03 85.00 35h 13.994 932 13 3. 2. 9. 23.121 83.579 1. 2 9.012 99.131 9.83 NOT 1.00 NA.ATA 13. 80 BOY 57.315 10.814 92.908 B. 39. 19.518 15.01. 19 nag 17.078 17. 8. 88. 88. 16.198 84.57 85. 85. 151.784 96.049 26.039 21.661 30.05 32.029 92.02 10. 10.02 97.397 19,136 15.111 98.913 38.113 25 250 1.6000 731.135 %. 160.80 12.00.37.98.888 12.00.11.12 34.00 131 36. 88. 88° 9.228 20.282 3 VV. V80 60. 81. 63.29.13.12.030.15 100 180.7 AND 8 20. V38 12.101 67.217 18. 16.372 D. 85. 85B 63.746 13.651 91:01 20. 6.73.572 2.591,91 %. 1/4 51.381 60.103 11.138 051 7.000 60. B. 2.31.01 16.0% 84.859 30.18 19.518 2033 81. 181 do. 74.08 20.16.49.18481 83.104 12.957 20.12.11 13.533 1. 98.58 16. 89. 89. 24.000 000 0. 1.01.00 77.916 88.123 P. L 10.130 0 01.891 82.3831 1. 121 80,411 00.13 262 21.382 00. 306 5 Mg 3002 2010 SYS %. 1. 918

3

c'b

%

218

60

18, .0

42.

10.

10,

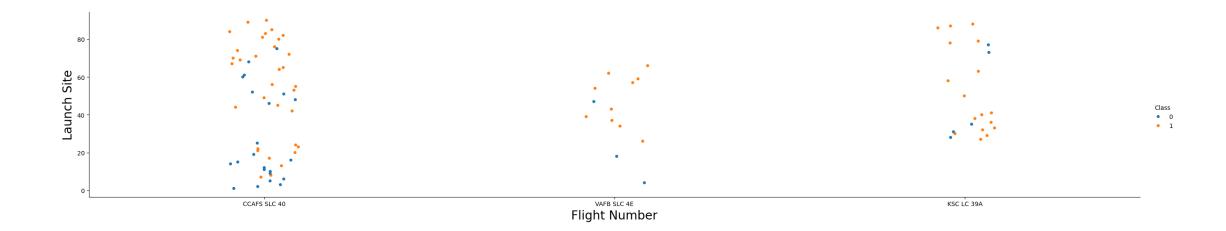
2

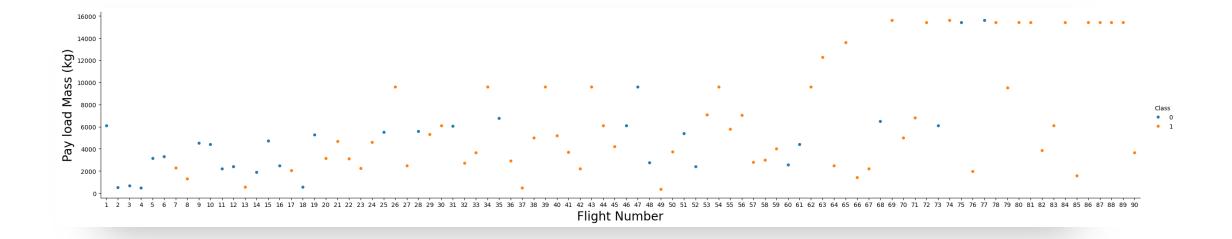
0.

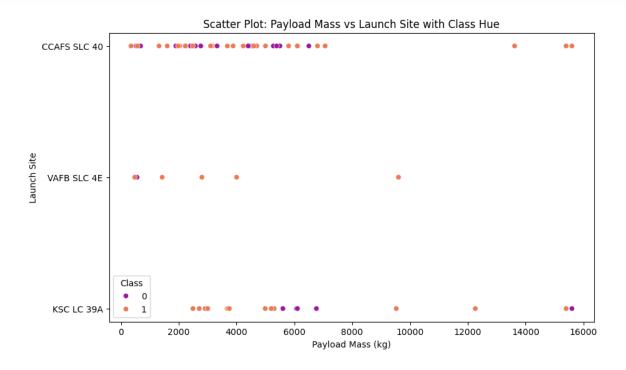
0

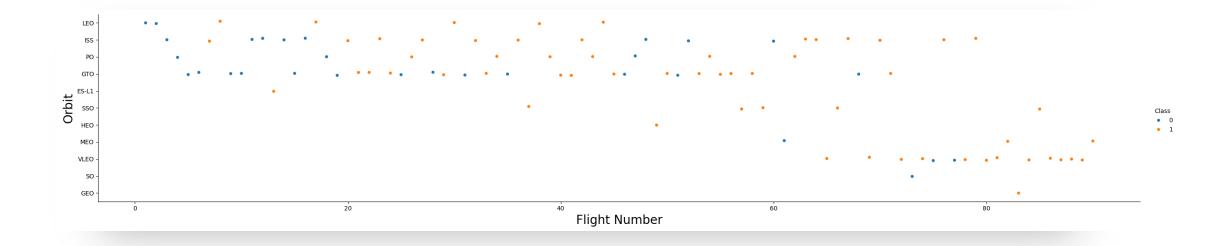
0.

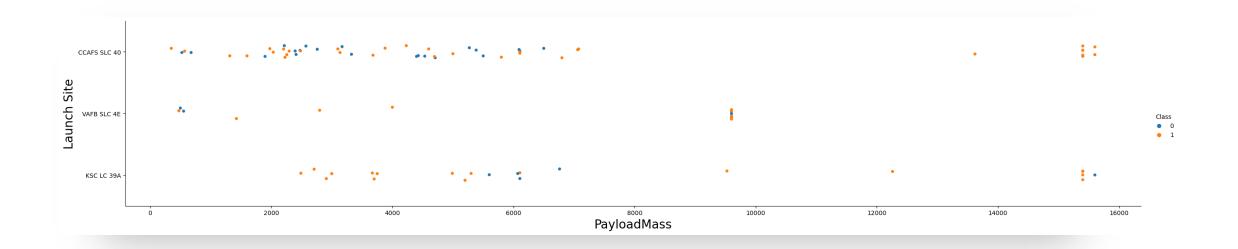
12

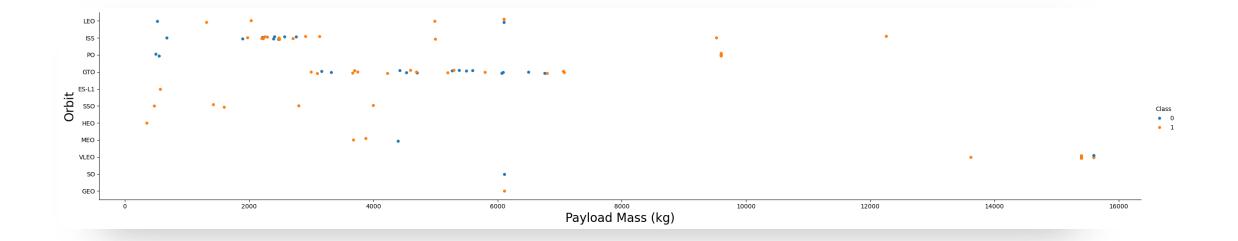




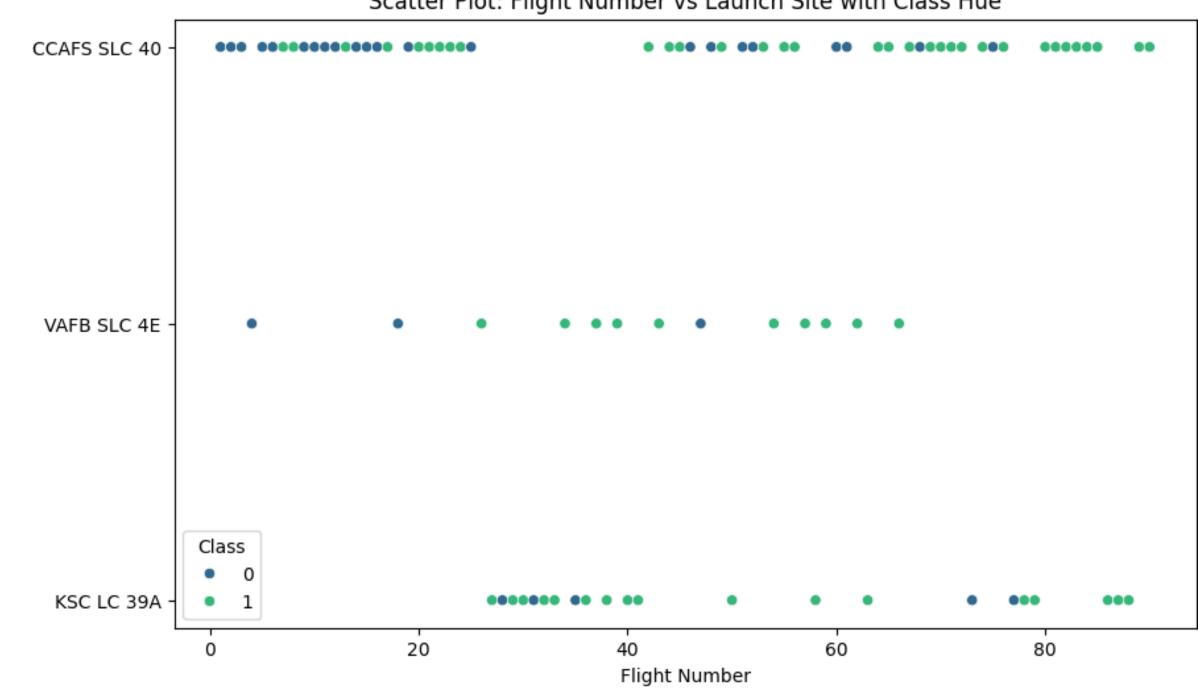


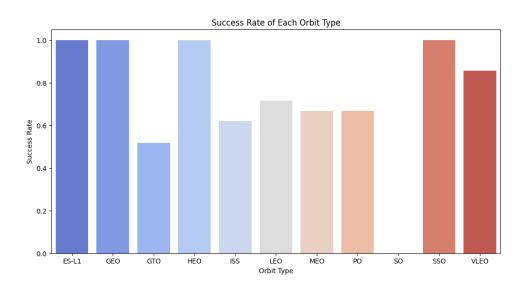




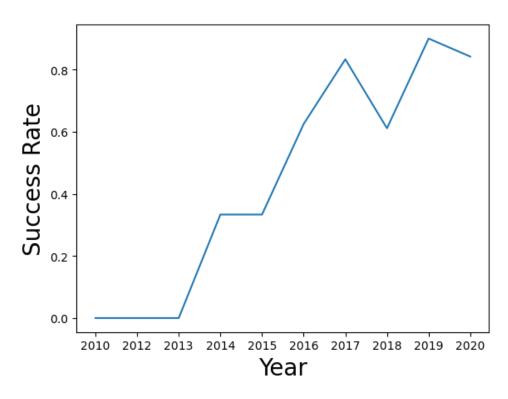






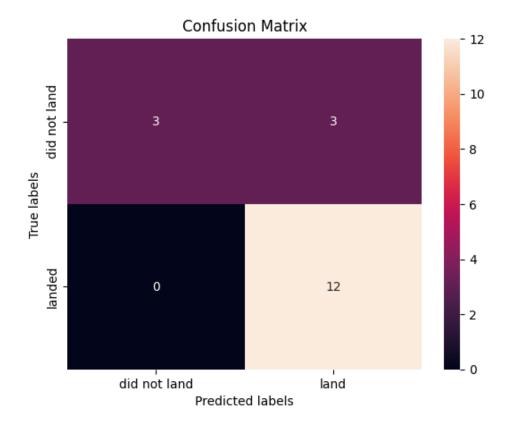


Success Rate



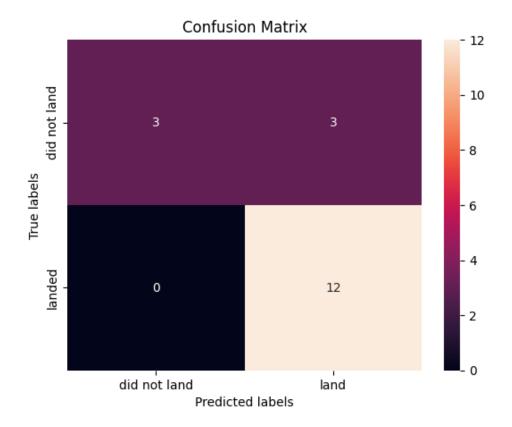
Confusion Matrix

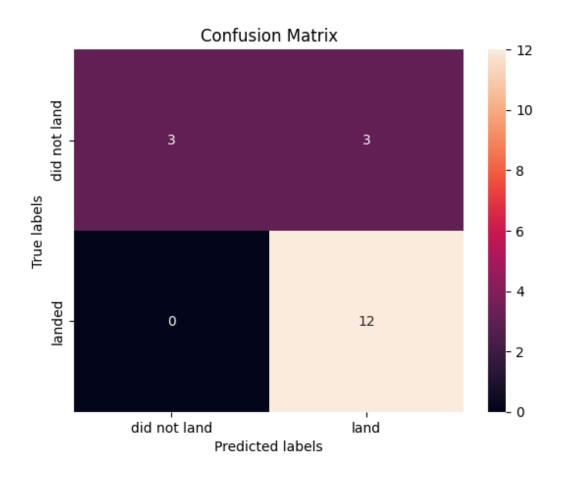
SVM



Confusion Matrix

Decision Tree



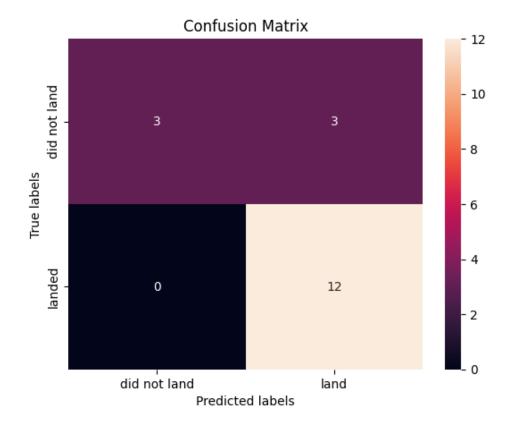


Confusion Matrix

KNN

Confusion Matrix

Logistic Regression



Conclusion

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate

