Applied Data Science Capstone

Name :Ahmed

8/7/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion
- Summary

Executive Summary

Project Overview: SpaceX Falcon 9 Landing Prediction

Objective: Predict the success of the SpaceX Falcon 9 first stage landing using various machine learning classification algorithms.

Methodologies:

1. Data Collection & Wrangling
   - Gather and preprocess data for analysis.
2. Exploratory Data Analysis (EDA)
   - Visualization:
     Interactive graphs highlighting correlations between features and launch outcomes.
   - SQL Analysis:
     Querying and analyzing data for insights.
3. Data Visualization:
   - Interactive Map:
     Using Folium to display launch locations and outcomes.
   - Dashboard:
     Building an interactive dashboard with Plotly Dash to explore data and results.
4. Predictive Analysis:
   - Classification Algorithms:
     Implementing and comparing various algorithms to predict landing success.
     Conclusion: Decision Tree algorithm shows promising results for prediction.

Executive Summary

Results Summary:

1. EDA Insights:
   - Key features correlated with launch outcomes identified.
2. Interactive Analytics:
   - Screenshots of interactive visualizations and dashboard.
3. Predictive Analysis:
   - Results of classification models and effectiveness of decision tree.

Introduction

- Capstone Project Overview: Predicting Falcon 9 First Stage Landing Success
- **Project Background:**
  - **SpaceX's Cost Efficiency:**
    - Falcon 9 launches cost $62 million, significantly less than competitors' $165 million.
    - Cost efficiency largely due to the reusability of the Falcon 9 first stage.
  - **Objective:**
    - Predict the success of Falcon 9 first stage landings to estimate launch costs and provide competitive insights.
- **Main Research Questions:**
  - **Feature Impact Analysis:**
    - How do payload mass, launch site, orbit type, and other variables affect landing success?
  - **Trend Analysis:**
    - Has the rate of successful landings improved over the years?
  - **Algorithm Evaluation:**
    - What is the most effective classification algorithm for predicting landing success?

Methodology

- **1. Data Collection:**
  - **Sources:**
    - SpaceX REST API
    - Web scraping from Wikipedia
- **2. Data Wrangling:**
  - **Processes:**
    - Filtering and cleaning data
    - Handling missing values
    - One Hot Encoding for binary classification
- **3. Exploratory Data Analysis (EDA):**
  - **Tools:**
    - Pandas and NumPy for data manipulation
    - SQL for advanced querying
  - **Techniques:**
    - Visualization with Matplotlib and Seaborn
    - Correlation analysis and feature exploration

Methodology

- **4. Data Visualization:**
  - **Tools:**
    - Folium for interactive maps
    - Plotly Dash for interactive dashboards
- **5. Machine Learning Prediction:**
  - **Algorithms:**
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree
    - K-Nearest Neighbors (KNN)
  - **Tasks:**
    - Building and tuning models
    - Evaluating performance to ensure optimal results

Methodology

Methodology: Data Collection via API

1. Accessing Data:

- **API Used:**
  - SpaceX API

2. Data Extraction Process:

- **Requesting Data:**
  - Fetching rocket launch data via the SpaceX API.
- **Decoding:**
  - Using .json() method to parse response content.
  - Converting JSON data into a dataframe with .json_normalize().
- **Filtering:**
  - Data is filtered to include only Falcon 9 launches.

- **3. Data Processing:**
  - **Handling Missing Values:**
    - Replacing missing values with the mean of the respective column.
  - **Data Structure:**
    - Resulting dataset: 90 rows (instances) and 17 columns (features).

Methodology: Data Collection via API

4. Data Transformation:

- **Custom Functions:**
  - Applying custom functions to extract relevant information.
- **Dictionary Construction:**
  - Organizing the data into a dictionary format.
- **Dataframe Creation:**
- Creating a dataframe from the constructed dictionary.

5. Data Snapshot:

- **Example Data:**
  - Displaying the first few rows of the filtered dataset.

Methodology: Data Collection via Web Scraping

1. Data Source:

- **Website:**
  - [Wikipedia Falcon 9 Launches]()

2. Data Extraction Process:

- **Requesting Data:**
  - Scraping launch data from the specified Wikipedia page.
- **BeautifulSoup:**
  - Creating a BeautifulSoup object from the HTML response to parse and extract data.
- **Column Extraction:**
  - Extracting column names from the HTML table header.

3. Data Collection:

HTML Parsing:

Parsing HTML tables to collect data.

Dictionary Construction:

Organizing the scraped data into a dictionary format.

Dataframe Creation:

Creating a dataframe from the constructed dictionary.

Methodology: Data Wrangling

1. Data Processing:

- **Handling Missing Entries:**
  - Ensured there are no missing values.
- **Encoding:**
  - Applied one-hot encoding to categorical features.
- **Additional Column:**
  - Added ‘Class’ column:
    - 0 for failed launches
    - 1 for successful launches
- **Final Dataset:**
  - Resulting in 90 rows (instances) and 83 columns (features).

2. Wrangling Details:

- **Outcome Conversion:**
  - Converted launch outcomes into training labels:
    - 1 for successful landings
    - 0 for unsuccessful landings
- **Launch Outcome Types:**
  - True/False labels for landing regions:
  - True/False Ocean
  - True/False RTLS (ground pad)
  - True/False ASDS (drone ship)

Methodology

Methodology: Data Visualization

1. **Matplotlib and Seaborn:**
    - **Visualization Techniques:**
        - Scatterplots, bar charts, and line charts.
    - **Key Insights:**
        - **Flight Number vs. Launch Site:**
            - Analyzing the distribution of flight numbers across different launch sites.
        - **Payload Mass vs. Launch Site:**
            - Examining the correlation between payload mass and launch sites.
        - **Success Rate vs. Orbit Type:**
            - Understanding success rates in relation to different orbit types.
2. **Folium:**
    - **Interactive Maps:**
        - **Markers for Launch Sites:**
            - Added markers for all launch sites with details like latitude, longitude, and labels.
        - **Launch Outcomes:**
            - Colored markers for successful (Green) and failed (Red) launches, using Marker Cluster to highlight high success rates.
        - **Distances to Proximities:**
            - Colored lines showing distances from KSC LC-39A to nearby features such as railways, highways, coastlines, and closest cities.

Methodology: Data Visulaization

3. Dash:

- **Interactive Dashboard:**
  - **Launch Sites Dropdown List:**
    - Dropdown menu to select different launch sites.
  - **Pie Chart:**
    - Displays total successful launches for all sites or a specific site.
  - **Payload Mass Range Slider:**
    - Slider to select a range of payload masses.
  - **Scatter Chart:**
    - Shows correlation between payload mass and success rate for different booster versions.

Methodology: Machine Learning Models

1. Data Preparation:

- **Standardization:**
    - Using StandardScaler to standardize the data for better model performance.
- **Data Splitting:**
    - Splitting data into training and test sets using train_test_split.

2. Model Creation and Training:

- **Models Implemented:**
    - **Logistic Regression**
    - **Support Vector Machine (SVM)**
    - **Decision Tree**
    - **K-Nearest Neighbors (KNN)**
- **Hyperparameter Tuning:**
    - Using GridSearchCV with cv=10 to find the best hyperparameters for each model.
- **Model Fitting:**
    - Training the models on the training set.

Methodology

Methodology: Machine Learning Models

3. Model Evaluation:

- **Accuracy Scores:**
  - Calculating the accuracy of each model on the test data using .score().
- **Confusion Matrix:**
  - Examining confusion matrices to understand classification performance.
- **Performance Metrics:**
  - Evaluating models using accuracy.

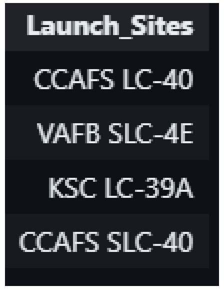
4. Summary:

- **Best Model Selection:**
  - Identifying the most effective model based on accuracy.

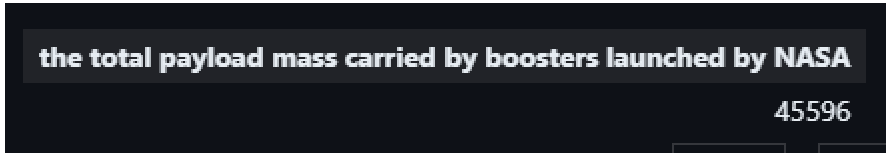
Results

1. SQL (EDA with SQL):

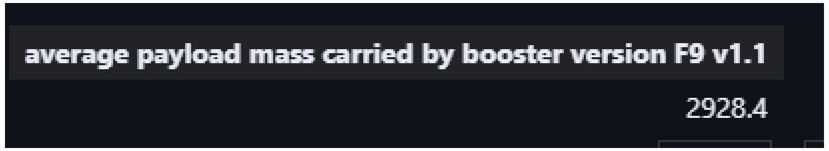
- **Key Findings:**
  - Unique launch sites
  - Total payload mass carried by NASA (CRS)
  - Average payload mass for booster version F9 v1.1



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



the total payload mass carried by boosters launched by NASA
45596

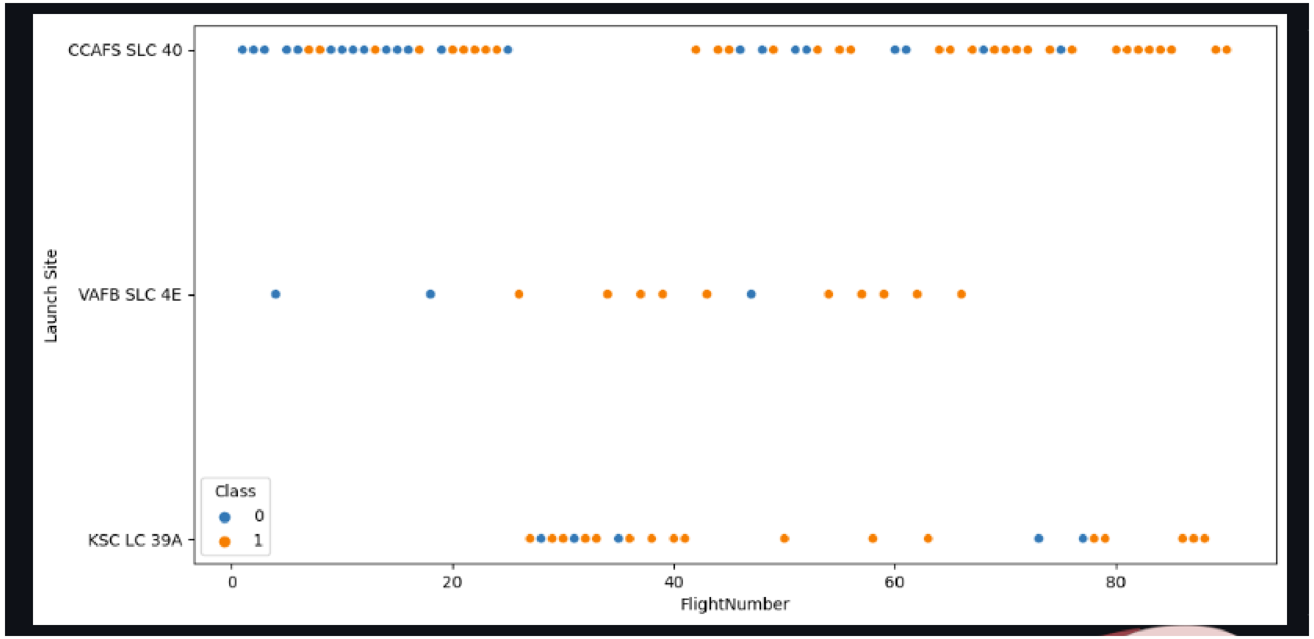


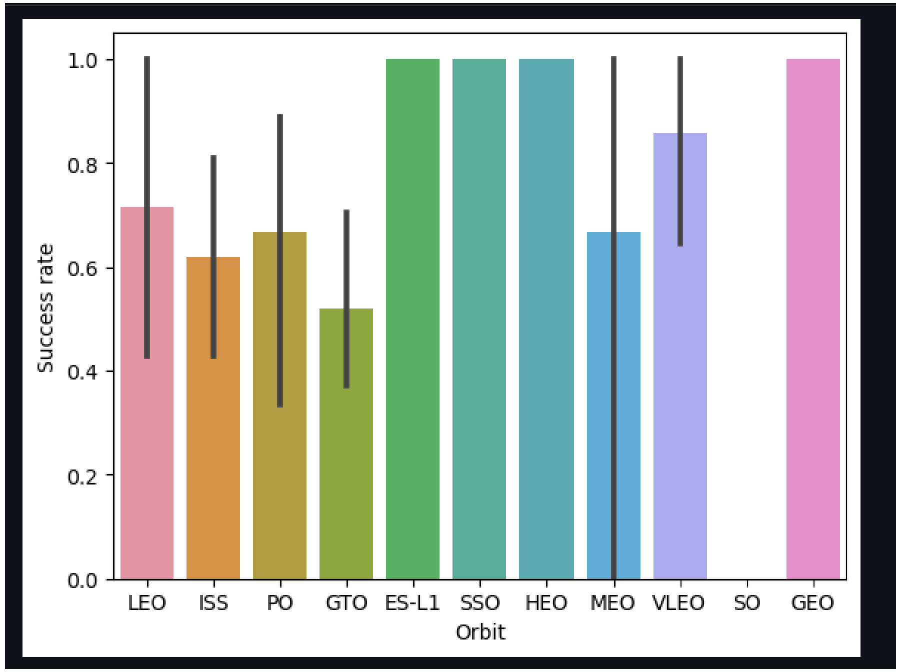
average payload mass carried by booster version F9 v1.1
2928.4

Results

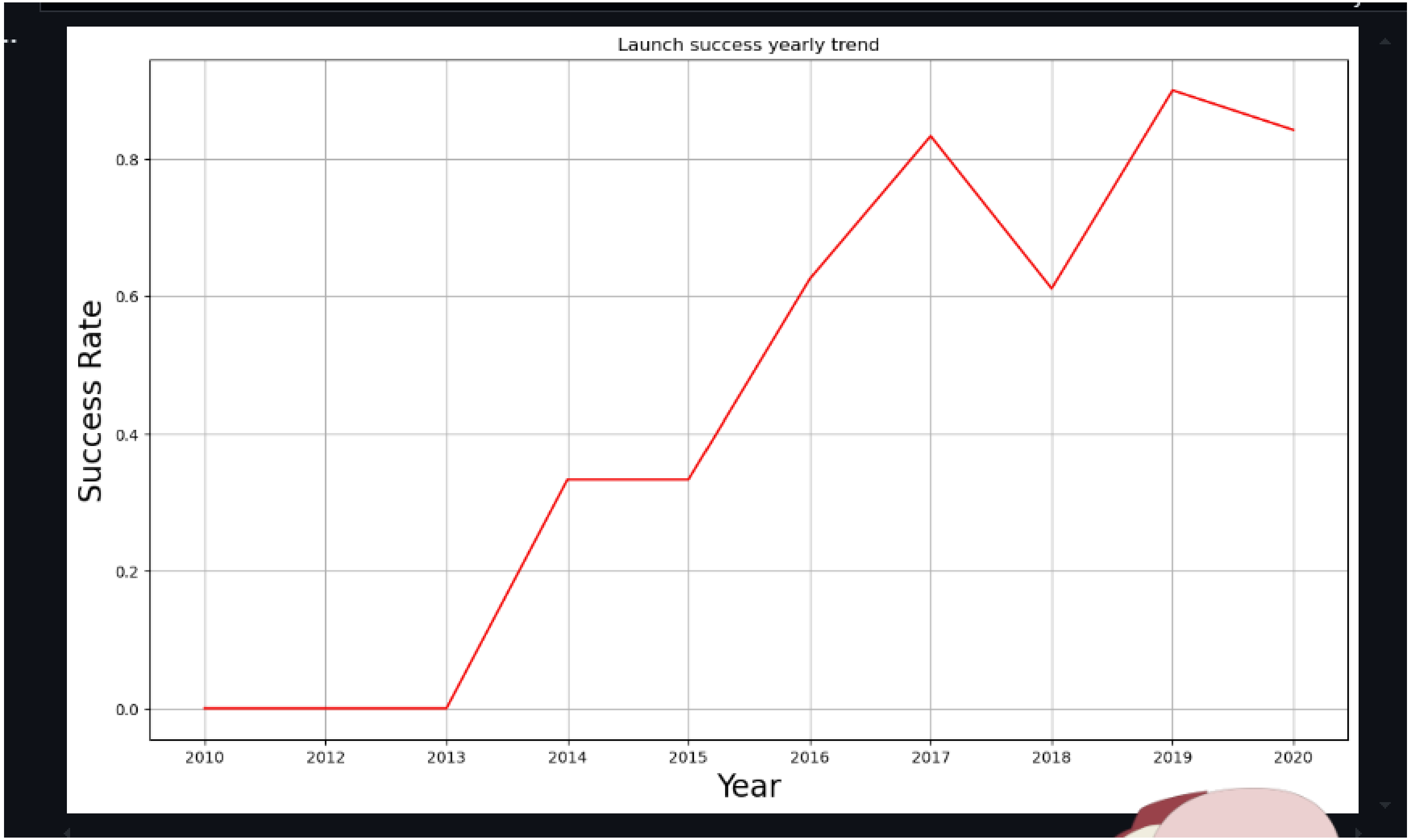
2. Matplotlib and Seaborn (EDA with Visualization):

- **Visualizations:**
  - Scatterplots, bar charts, and line charts





Results

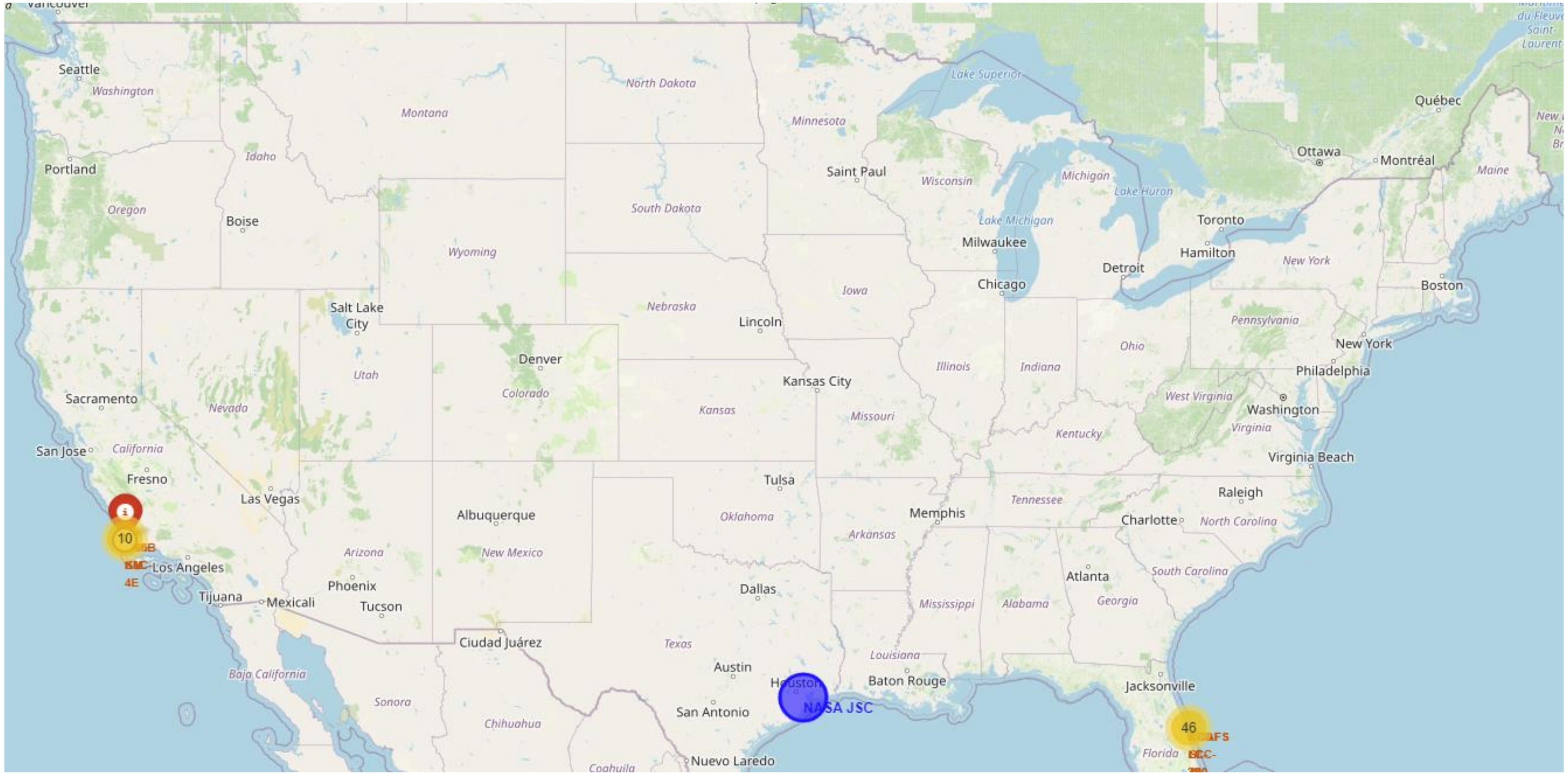


- **Insights:**
  - Relationship between flight number and launch site
  - Payload mass distribution by launch site
  - Success rate by orbit type

Results

3. Folium

- **Interactive Maps:**
  - Markers for launch sites and outcomes
  - Distances between launch sites and key proximities (e.g., railways, highways)

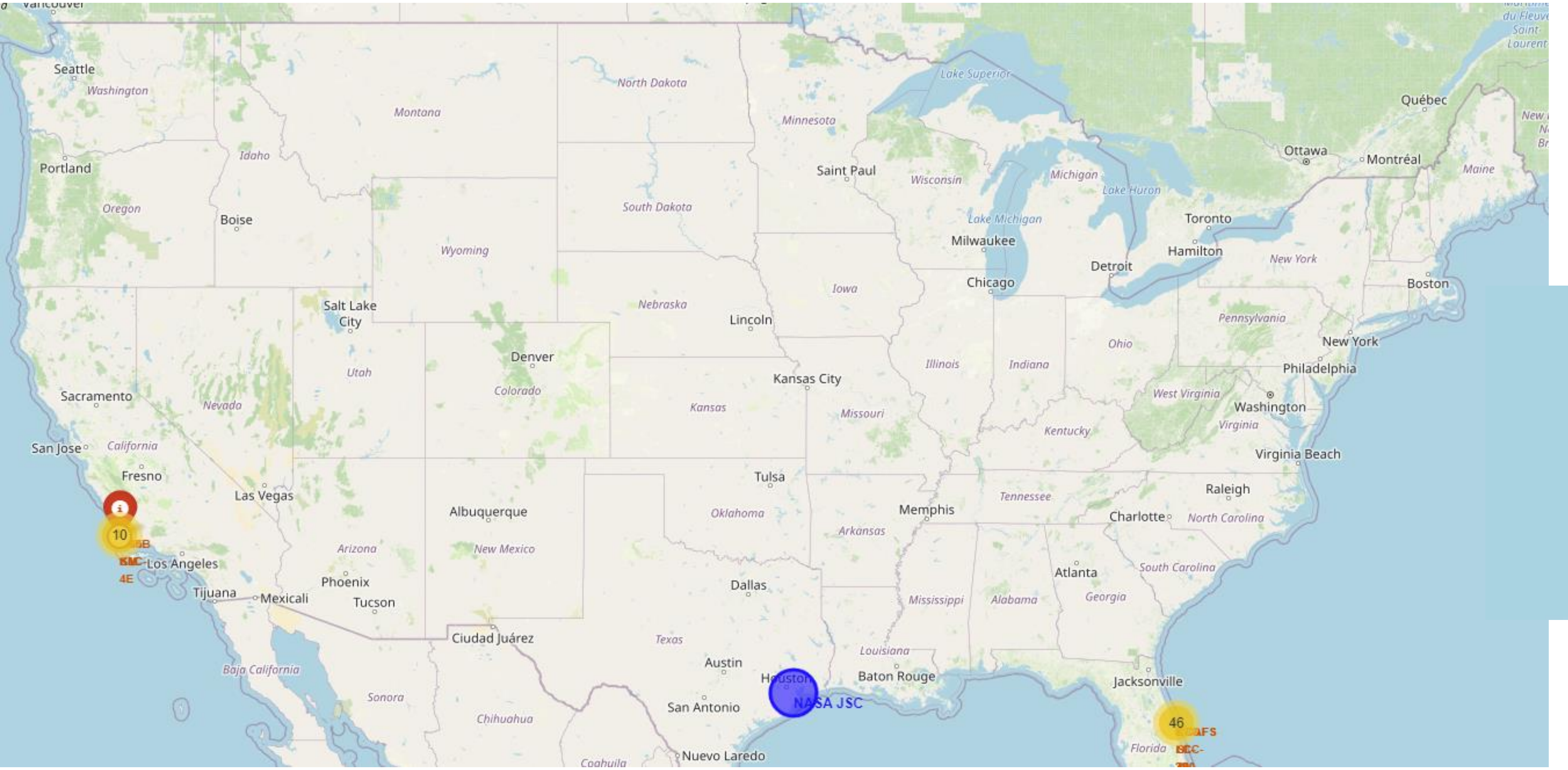


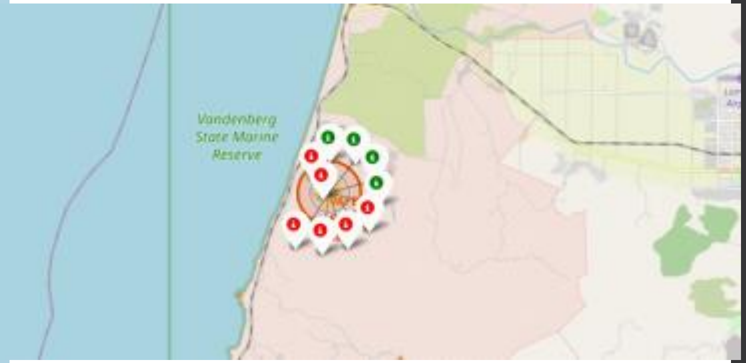
Results

Results

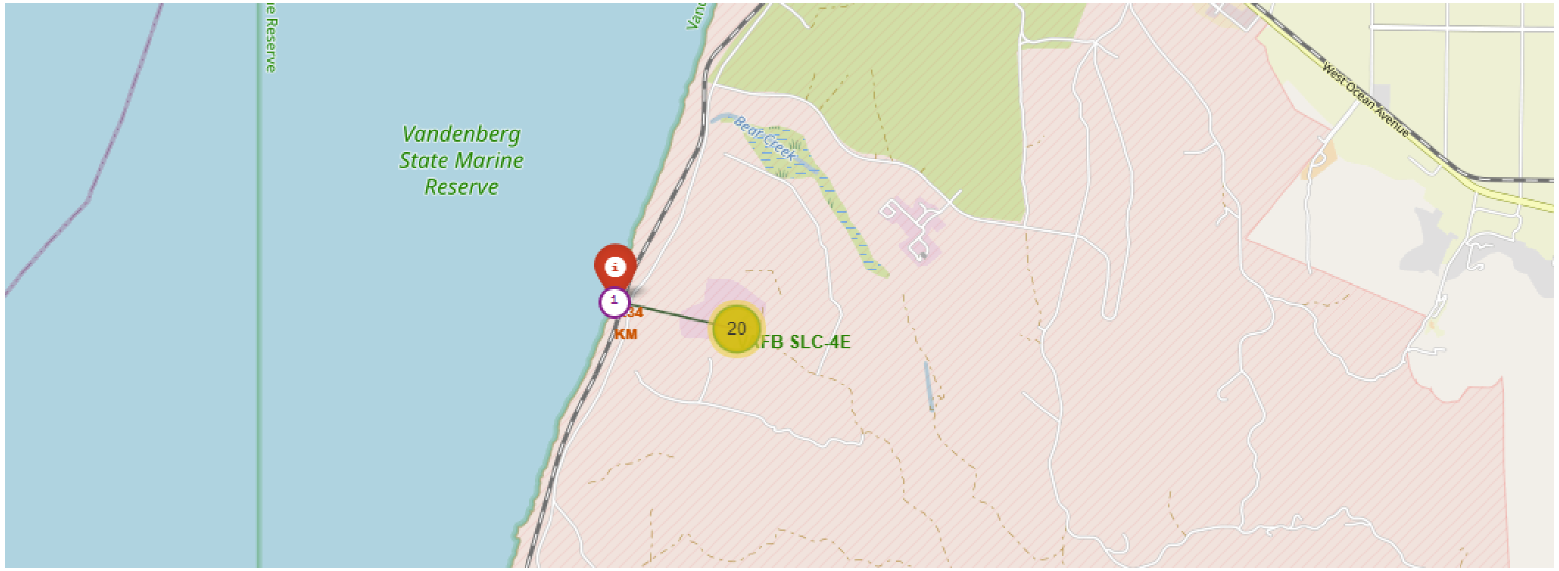
3. Folium

- **Interactive Maps:**
  - Markers for launch sites and outcomes
  - Distances between launch sites and key proximities (e.g., railways, highways)





Results



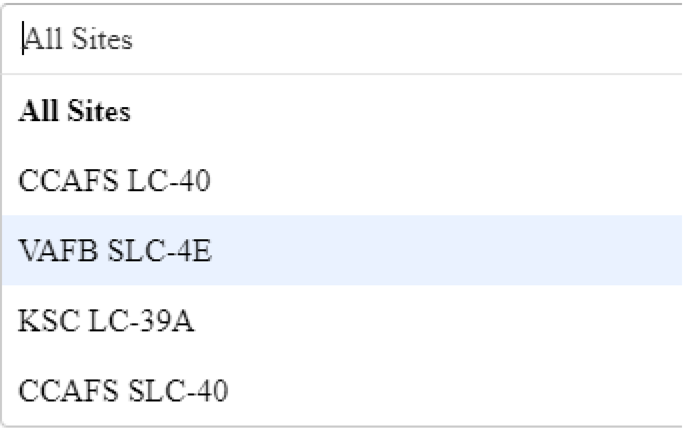
- The distances between a launch site to its proximities such as the nearest city.
    - The picture above shows the distance between the VAFB SLC-4E launch site and the nearest coastline

Results

Results

4. dash

- **Interactive Dashboard:**
  - Dropdown for selecting launch sites
  - Pie chart showing successful launches
  - Payload mass range slider
  - Scatter chart for payload mass vs. success rate



All Sites

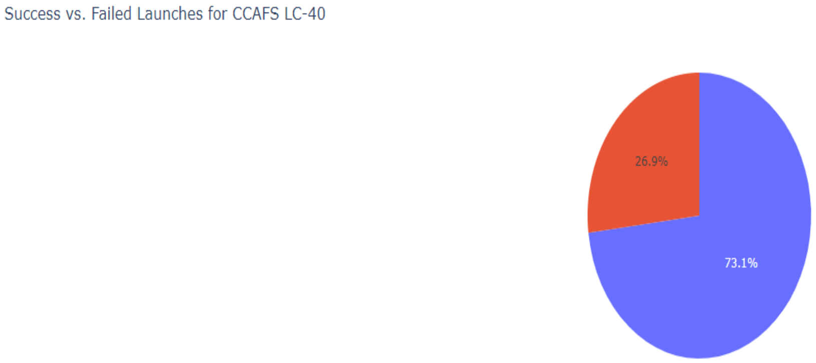
All Sites

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

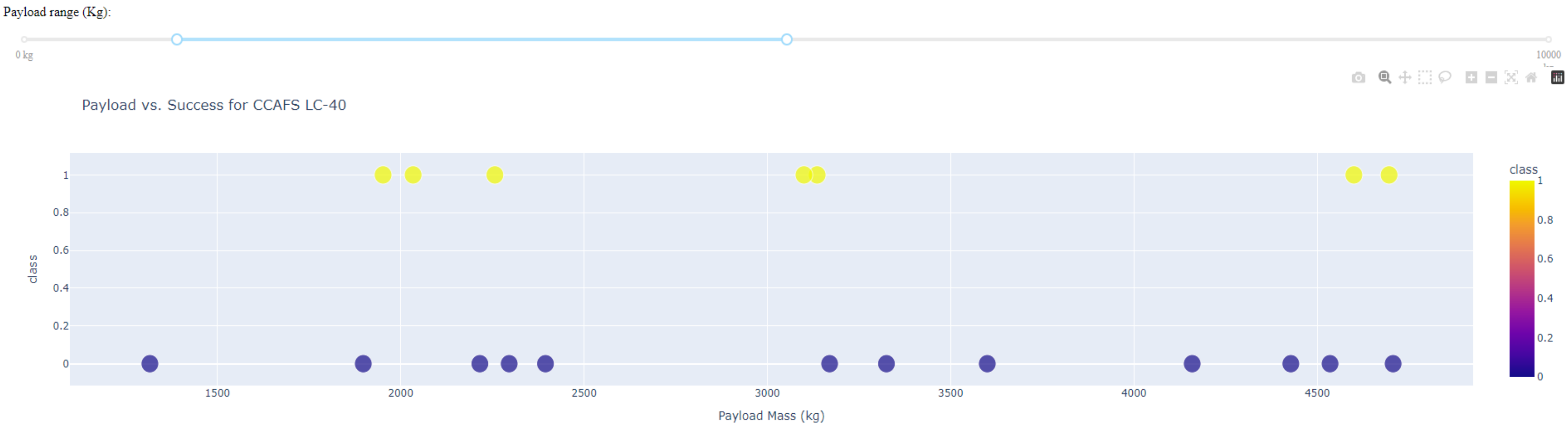


Results

Results

4. dash

- **Interactive Dashboard:**
  - Payload mass range slider
  - Scatter chart for payload mass vs. success rate

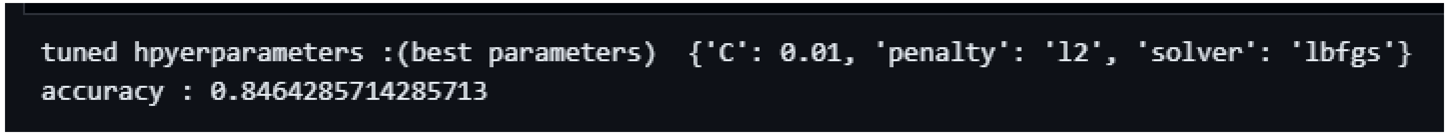


Results

5. Predictive Analysis

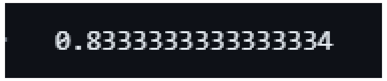
- **Logistic regression**

Best parameters:
Accuracy:



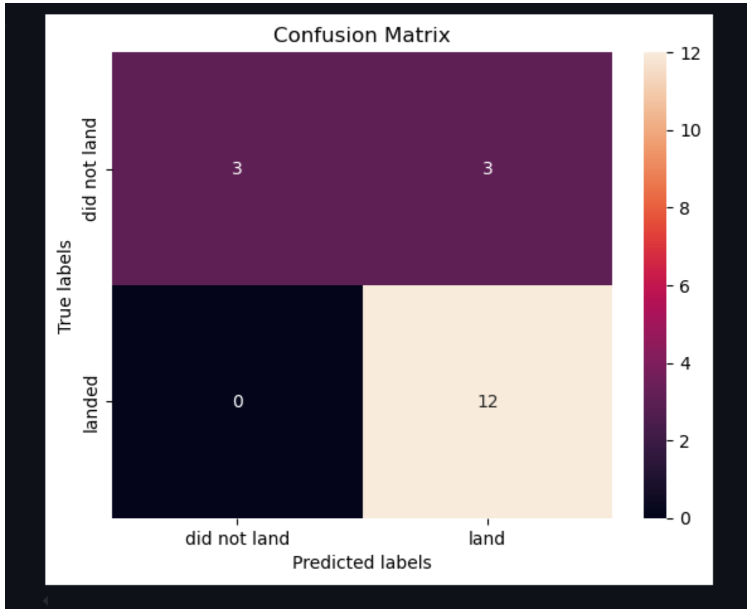
```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

Test accuracy:



```
0.8333333333333334
```

Confusion matrix:



Results

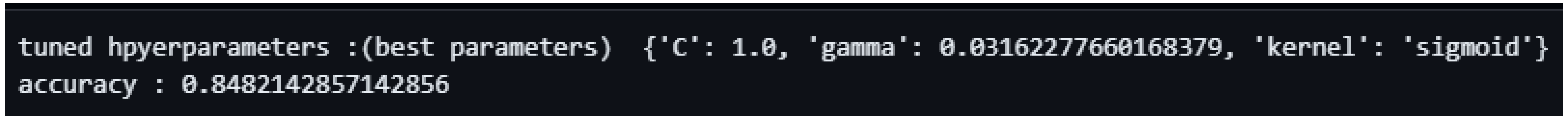
Results

Results

5. Predictive Analysis

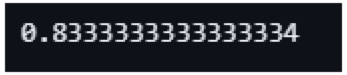
- **Support vector machine**

Best parameters:
Accuracy:



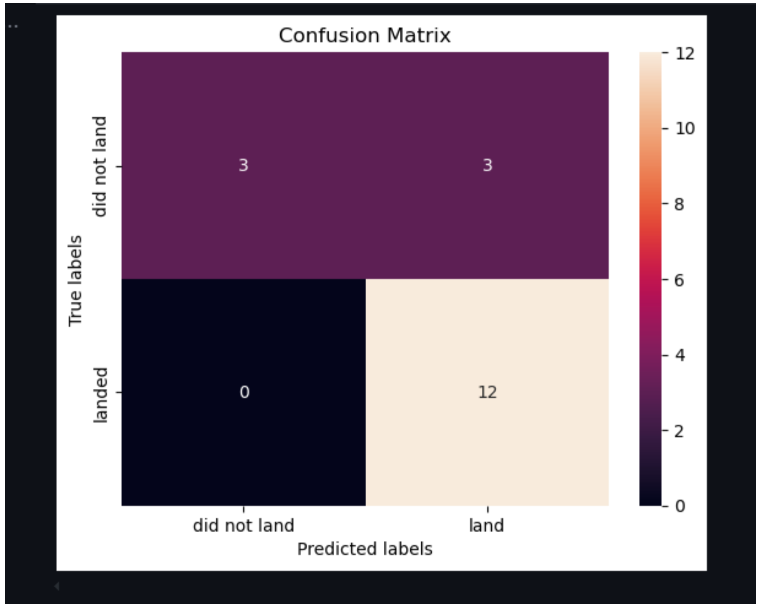
```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

Test accuracy:



```
0.8333333333333334
```

Confusion matrix:



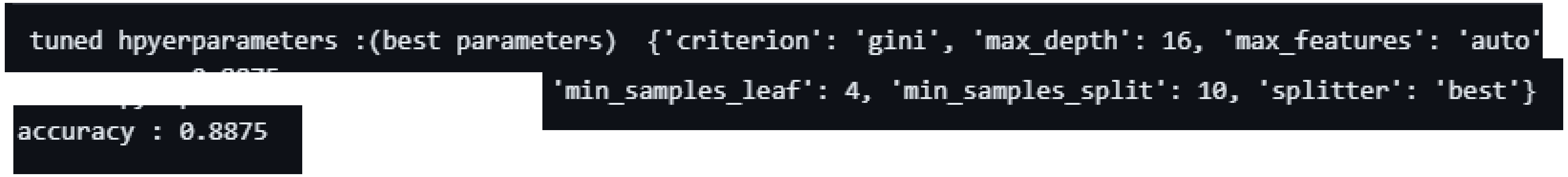
Results

5. Predictive Analysis

- **Decision tree**

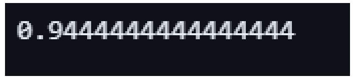
Best parameters:

Accuracy:



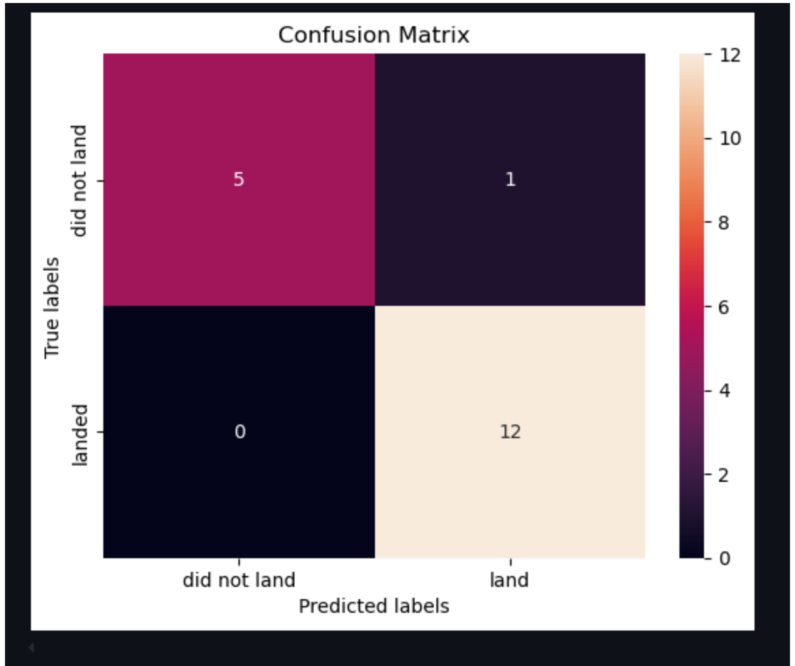
```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto'
'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}
accuracy : 0.8875
```

Test accuracy:



```
0.9444444444444444
```

Confusion matrix:

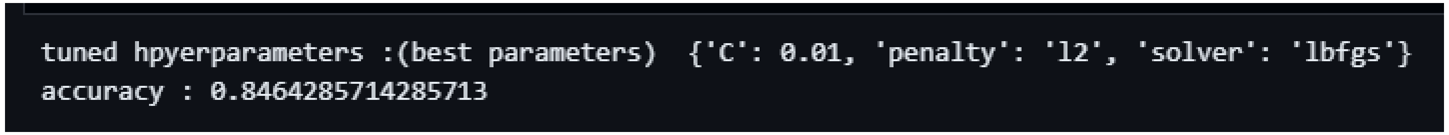


Results

5. Predictive Analysis

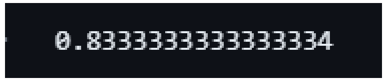
- **KNN**

Best parameters:
Accuracy:



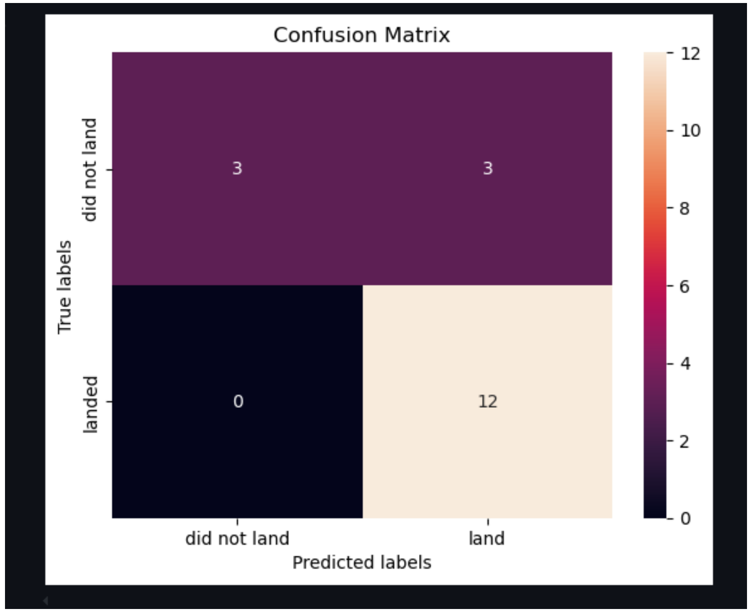
```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

Test accuracy:



```
0.8333333333333334
```

Confusion matrix:



Results

- 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 traingin set.
- But the best one is decision tree on the test set with accuracy 94%

Discussion

1. Feature Correlations:

- **Payload and Orbit Types:**
  - **Polar, LEO, and ISS Orbits:**
    - Higher success rates observed with heavy payloads.
  - **GTO Orbit:**
    - Mixed success and failure rates, making it harder to distinguish outcomes based on payload mass alone.

2. Feature Impact:

- **Feature Influence:**
  - Each feature impacts the final mission outcome in different ways.
  - Identifying precise impacts can be complex due to the multifaceted nature of features and outcomes.

3. Machine Learning Approach:

- **Pattern Recognition:**
  - Machine learning algorithms can be employed to uncover patterns in historical data.
  - Predictive models can learn from past mission data to forecast future mission success based on the given features.

4. Model Utilization:

- **Predictive Analysis:**
  - Using models like Logistic Regression, SVM, Decision Tree, and KNN to predict mission outcomes.
  - Leveraging these models to handle complex interactions between features and improve prediction accuracy.

Conclusions

- In our analysis of Falcon 9 launches, we aimed to predict the success of the first stage landing to aid in determining launch costs. Through rigorous data analysis and machine learning, several key insights emerged.
- **Effectiveness of Machine Learning Algorithms**
- Among the various machine learning algorithms tested, the Decision Tree model demonstrated the most robust performance. This model excelled in predicting the outcome of Falcon 9 launches, outperforming others such as Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). The Decision Tree’s ability to handle complex feature interactions and provide clear decision boundaries made it the best fit for this dataset.

Conclusions

- **Impact of Payload Mass**
- Our analysis revealed a notable trend regarding payload mass. Launches with lower payload masses generally achieved better success rates compared to those with heavier payloads. This finding suggests that payload weight significantly influences landing success, with lighter payloads being associated with higher chances of a successful landing.
- **Geographical Insights**
- The geographical distribution of launch sites also plays a crucial role in mission outcomes. Most Falcon 9 launch sites are situated close to the Equator and near coastlines. This proximity to the Equator is advantageous for maximizing the efficiency of launches, while coastal locations facilitate safe and efficient launches by providing access to open water for landing attempts.

Conclusions

- **Trends Over Time**
- Another important observation is the increasing success rate of launches over the years. This upward trend highlights advancements in technology and improvements in launch strategies, contributing to more reliable and successful missions.
- **Site-Specific Success Rates**
- Among the various launch sites, Kennedy Space Center Launch Complex 39A (KSC LC-39A) stands out with the highest success rate. This site’s superior performance underscores its effectiveness in achieving successful launches.

Conclusions

- **Orbit-Specific Success Rates**
- In terms of orbit types, certain orbits such as ES-L1, GEO, HEO, and SSO have achieved a 100% success rate. These orbits exhibit a perfect track record for successful landings, indicating their favorable conditions for Falcon 9 missions.
- **Site-Specific Success Rates**
- Among the various launch sites, Kennedy Space Center Launch Complex 39A (KSC LC-39A) stands out with the highest success rate. This site’s superior performance underscores its effectiveness in achieving successful launches.

Summary

- In summary, our project demonstrated that the Decision Tree algorithm is the most effective for predicting Falcon 9 landing outcomes. Factors such as payload mass, launch site geography, and orbit type all influence mission success. The insights gained from this analysis not only improve our understanding of Falcon 9 launches but also provide valuable guidance for future missions and cost estimations.

Thank You