SpaceX Falcon 9 Landing Prediction Capstone Project – Data Science

Sabrina Monay
Applied Data Science
16/12/25

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results & Visualization
- Dashboard
- Discussion
- Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Project Objective

To determine if the first stage of the Falcon 9 rocket will land successfully, allowing for an estimation of the actual launch cost and an evaluation of SpaceX's competitiveness against other providers.

Approach

Data collection via API and web scraping

Data cleaning and preparation

Exploratory Data Analysis (EDA)

Interactive visualization (Folium and Dash)

Supervised classification models

Key Result

The SVM model achieved the best predictive performance on test data.

INTRODUCTION

Context

SpaceX offers launches for $62 million, significantly cheaper than competitors (over $165 million).

The savings come from the reuse of the first stage.

Problem

The cost depends on whether the first stage lands successfully.

Key Question

Can we predict whether the first stage will land using historical data?

METHODOLOGY

Data Collection ✅
Sources
SpaceX
REST API Web Scraping (Wikipedia)
CSV files provided by IBM Skills Network

Collected
Data Launch site
Payload mass
Orbit type
Booster type
Landing result

Data Cleaning and Preparation ✅

Tasks performed

Removal of null values
Normalization of numeric variables
One-hot coding of categorical variables
Creation of the target variable **Class**
1 = successful landing
0 = failed landing

METHODOLOGY

EDA Methodology Visual and Interative Analysis ✅

Tools

Pandas, Matplotlib, Seaborn
Plotly Express
Folium
Dash

Objective

To identify patterns between operational variables and landing success.

RESULTS

- A clear relationship was identified between payload mass and the probability of a successful landing.

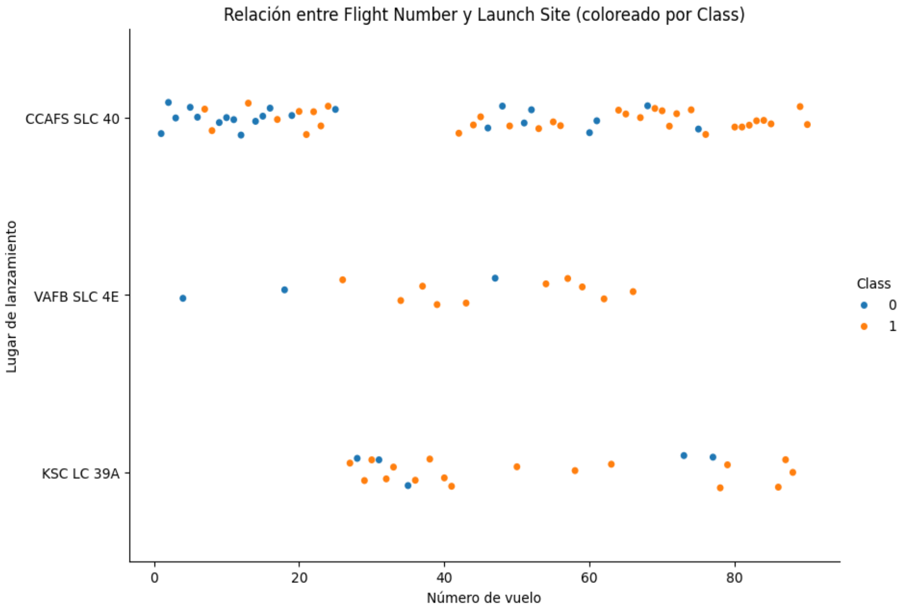
- Launches with extremely high payloads exhibit a lower success rate.

- Some launch sites consistently show better results than others.

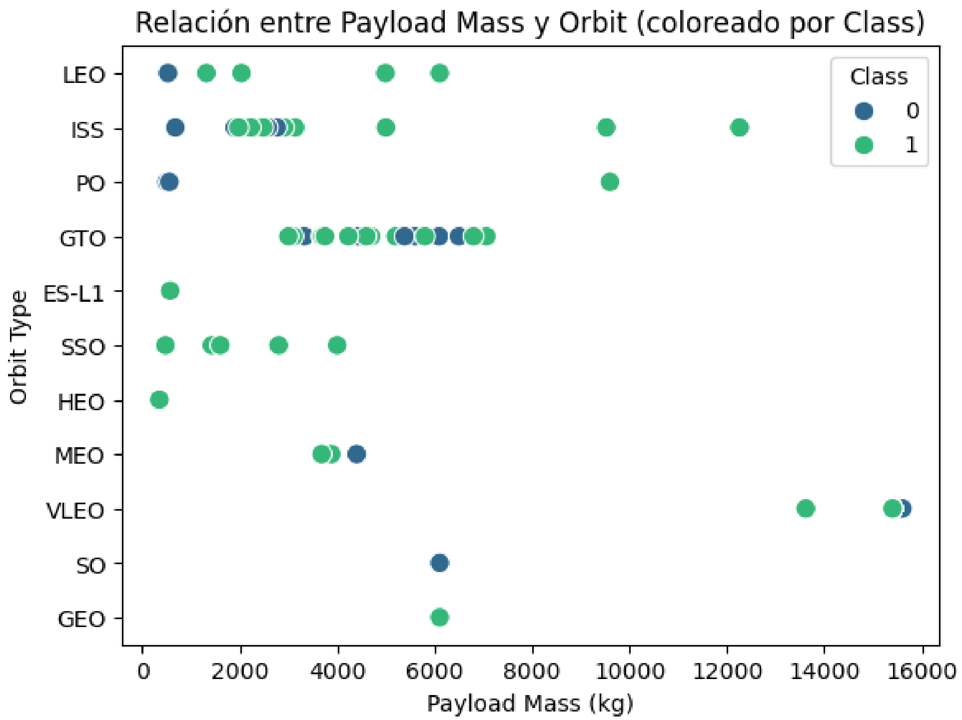
- Reused boosters tend to have higher successful landing rates.

- Some orbits have a higher success rate.

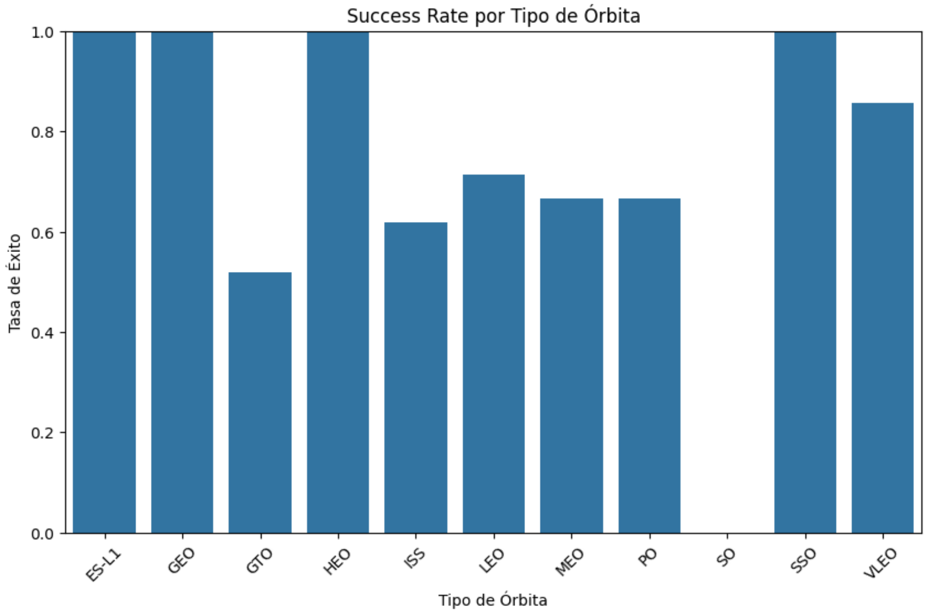
RESULTS



Relación entre Flight Number y Launch Site (coloreado por Class)
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Lugar de lanzamiento
0
20
40
60
80
Número de vuelo
Class
0
1



Relación entre Payload Mass y Orbit (coloreado por Class)
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
Orbit Type
0
2000
4000
6000
8000
10000
12000
14000
16000
Payload Mass (kg)
Class
0
1



Success Rate por Tipo de Órbita
1.0
0.8
0.6
0.4
0.2
0.0
Tasa de Éxito
ES-L1
GEO
GTO
HEO
ISS
LEO
MEO
PO
SO
SSO
VLEO
Tipo de Órbita

RESULTS

SQL Analysis ✅

An uneven distribution of launches was observed across the different sites.

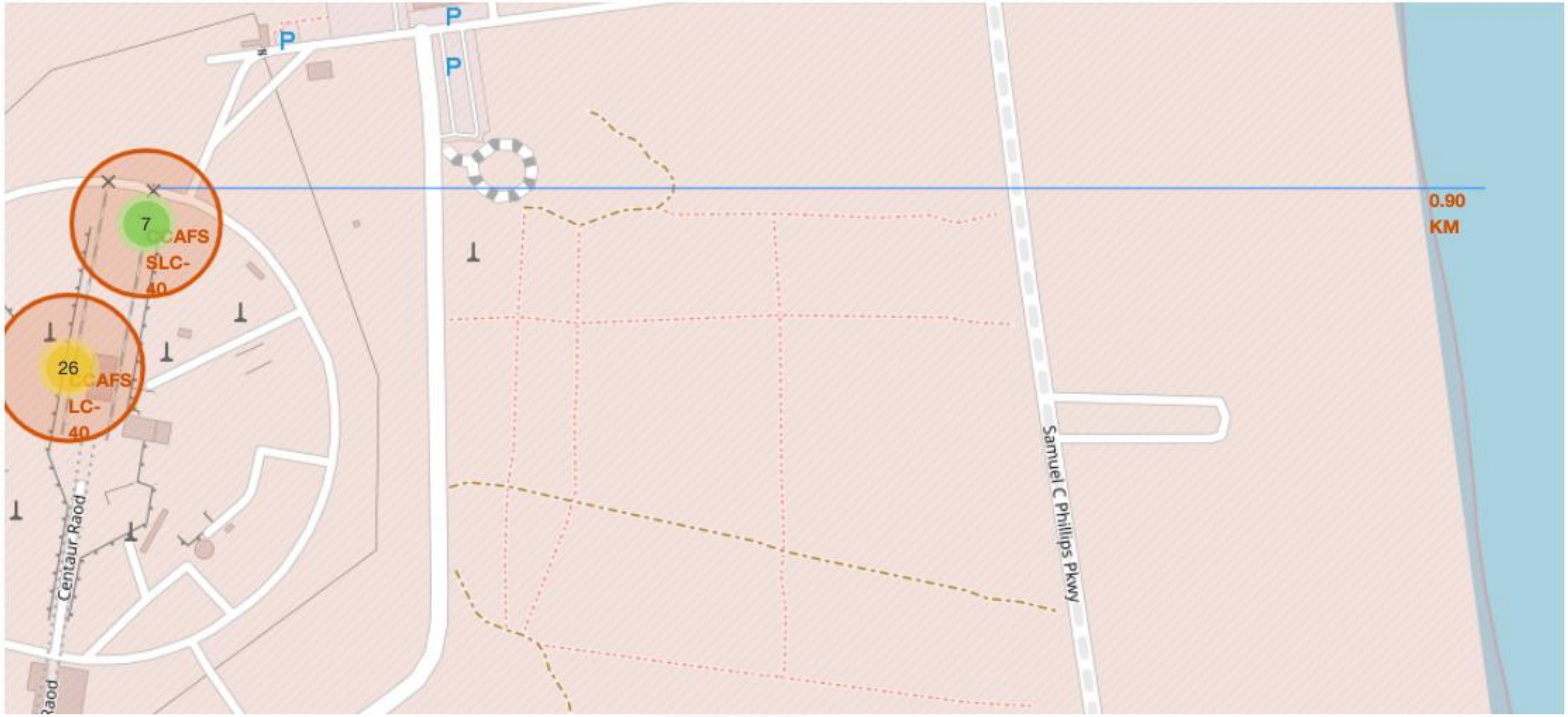
Certain orbits exhibited higher success rates.

The SQL queries confirmed the patterns detected in the visual EDA.

The structured analysis allowed for reproducible validation of the results.

Folium Results

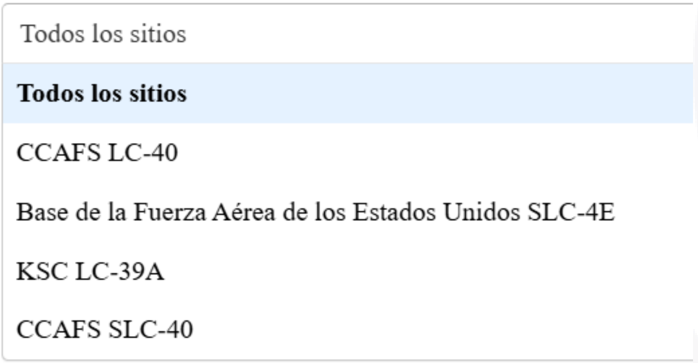
Interactive maps showed that coastal launch sites facilitate the recovery of the first stage.

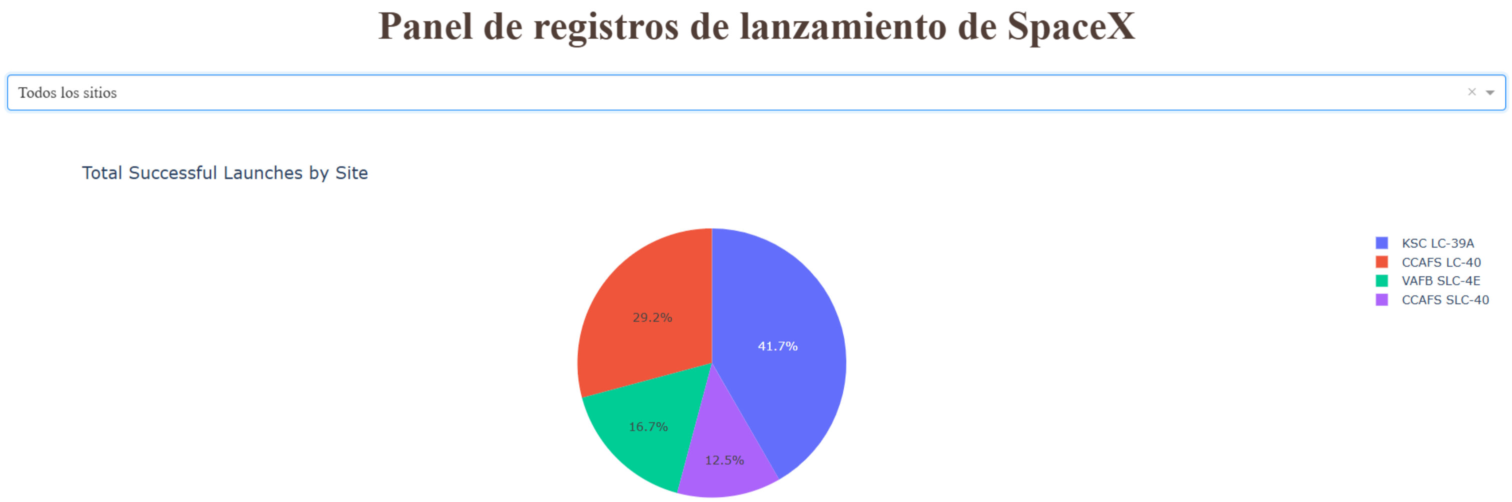


DASHBOARD Results

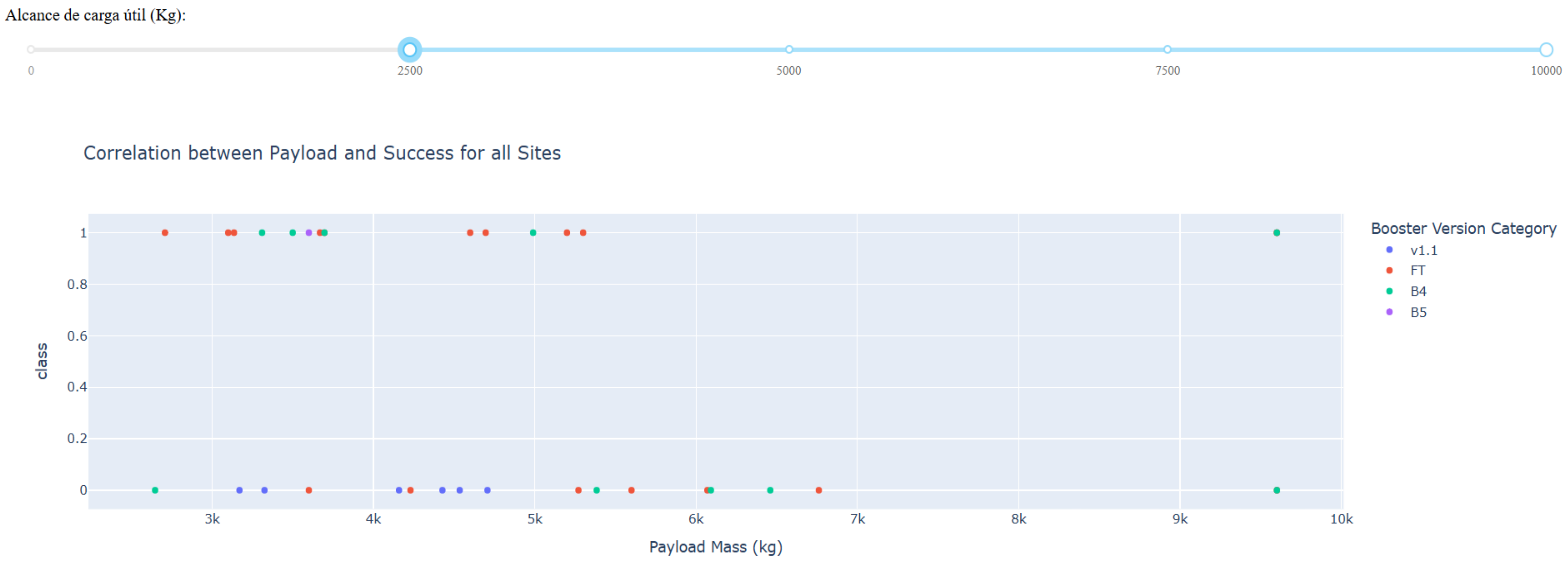
The Plotly Dashboard allowed for dynamic exploration of:

- Hits by site
- Impact of payload range
- Differences by booster version





DASHBOARD



Alcance de carga útil (Kg):
0
2500
5000
7500
10000
Correlation between Payload and Success for all Sites
class
1
0.8
0.6
0.4
0.2
0
3k
4k
5k
6k
7k
8k
9k
10k
Payload Mass (kg)
Booster Version Category
v1.1
FT
B4
B5

Results of the Classification Models

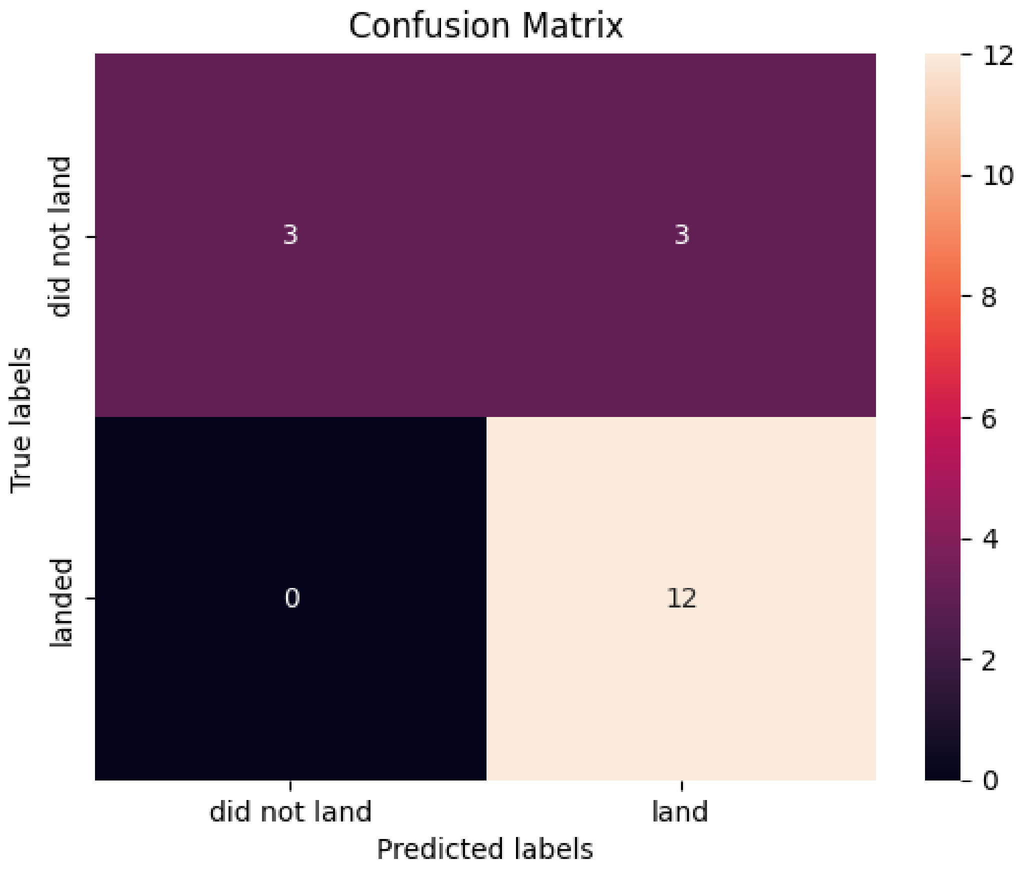
Four supervised models were trained and optimized:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K-Nearest Neighbors (KNN)

The SVM model achieved the highest accuracy on the test data.

The results indicate that it is possible to predict landings with good accuracy using historical data.

Classification Models



Modelo	Accuracy
Regresión Logística	~0.83
SVM	**~0.86**
Árbol de Decisión	~0.78
KNN	~0.81

CONCLUSION

- It is possible to predict the Falcon 9 landing with good accuracy.
- The payload and launch site are key variables.
- SVM offers the best predictive performance.
- This approach can help estimate costs and compete with SpaceX.

CREATIVITY AND INNOVATIVE IDEAS ✅

- Incorporate meteorological data
- Use deep learning models
- Real-time cost prediction
- Public dashboard for simulations

APPENDIX

```
spacex_df.tail(10)
```

	Launch Site	Lat	Long	class
46	KSC LC-39A	28.573255	-80.646895	1
47	KSC LC-39A	28.573255	-80.646895	1
48	KSC LC-39A	28.573255	-80.646895	1
49	CCAFS SLC-40	28.563197	-80.576820	1
50	CCAFS SLC-40	28.563197	-80.576820	1
51	CCAFS SLC-40	28.563197	-80.576820	0
52	CCAFS SLC-40	28.563197	-80.576820	0
53	CCAFS SLC-40	28.563197	-80.576820	0
54	CCAFS SLC-40	28.563197	-80.576820	1
55	CCAFS SLC-40	28.563197	-80.576820	0