Space X Falcon 9 First Stage Landing Prediction

Assignment: Machine Learning Prediction

Estimated time needed: **60** minutes

Space X 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 Space X 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 space X for a rocket launch. In this lab, you will create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.

Several examples of an unsuccessful landing are shown here:

Most unsuccessful landings are planed. Space X; performs a controlled landing in the oceans.

Objectives

Perform exploratory Data Analysis and determine Training Labels

- create a column for the class
- Standardize the data
- Split into training data and test data

-Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

- Find the method performs best using test data

Import Libraries and Define Auxiliary Functions

We will import the following libraries for the lab

In [404…

```
# Pandas is a software library written for the Python programming language for d
import pandas as pd
# NumPy is a library for the Python programming language, adding support for lar
import numpy as np
# Matplotlib is a plotting library for python and pyplot gives us a MatLab like 
import matplotlib.pyplot as plt
#Seaborn is a Python data visualization library based on matplotlib. It provides
import seaborn as sns
# Preprocessing allows us to standarsize our data
from sklearn import preprocessing
# Allows us to split our data into training and testing data
from sklearn.model_selection import train_test_split
# Allows us to test parameters of classification algorithms and find the best on
from sklearn.model_selection import GridSearchCV
# Logistic Regression classification algorithm
from sklearn.linear_model import LogisticRegression
# Support Vector Machine classification algorithm
from sklearn.svm import SVC
# Decision Tree classification algorithm
from sklearn.tree import DecisionTreeClassifier
# K Nearest Neighbors classification algorithm
from sklearn.neighbors import KNeighborsClassifier
```

This function is to plot the confusion matrix.

In [405…

```
def plot_confusion_matrix(y,y_predict):
    "this function plots the confusion matrix"
    from sklearn.metrics import confusion_matrix

    cm = confusion_matrix(y, y_predict)
    ax= plt.subplot()
    sns.heatmap(cm, annot=True, ax = ax); #annot=True to annotate cells
    ax.set_xlabel('Predicted labels')
    ax.set_ylabel('True labels')
    ax.set_title('Confusion Matrix');
    ax.xaxis.set_ticklabels(['did not land', 'land']); ax.yaxis.set_ticklabels([
```

Load the dataframe

Load the data

In [406…

```
data = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain

# If you were unable to complete the previous lab correctly you can uncomment an

# data = pd.read_csv('https://cf-courses-data.s3.us.cloud-object-storage.appdoma

data.head()
```

Out[406]:

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

In [407…

```
X = pd.read_csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cl

# If you were unable to complete the previous lab correctly you can uncomment ar

#X = pd.read_csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.c

X.head(100)
```

Out[407]:

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GT
0	1.0	6104.959412	1.0	1.0	0.0	0.0	0.0	0.
1	2.0	525.000000	1.0	1.0	0.0	0.0	0.0	0.
2	3.0	677.000000	1.0	1.0	0.0	0.0	0.0	0.
3	4.0	500.000000	1.0	1.0	0.0	0.0	0.0	0.
4	5.0	3170.000000	1.0	1.0	0.0	0.0	0.0	1.
...	...	...	...	...	...	...	...	.
85	86.0	15400.000000	2.0	5.0	2.0	0.0	0.0	0.
86	87.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.
87	88.0	15400.000000	6.0	5.0	5.0	0.0	0.0	0.
88	89.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.
89	90.0	3681.000000	1.0	5.0	0.0	0.0	0.0	0.

90 rows × 83 columns

TASK 1

Create a NumPy array from the column `Class` in `data`, by applying the method `to_numpy()` then assign it to the variable `Y`,make sure the output is a Pandas series (only one bracket df['name of column']).

In [408…

```
Y = data['Class'].to_numpy()
```

TASK 2

Standardize the data in `X` then reassign it to the variable `X` using the transform provided below.

In [409…

```
# students get this
transform = preprocessing.StandardScaler()
```

In [410…

```
X = transform.fit_transform(X)
X
```

Out[410]:

```
array([[-1.71291154e+00, -1.94814463e-16, -6.53912840e-01, ...,
         -8.35531692e-01,  1.93309133e+00, -1.93309133e+00],
       [-1.67441914e+00, -1.19523159e+00, -6.53912840e-01, ...,
         -8.35531692e-01,  1.93309133e+00, -1.93309133e+00],
       [-1.63592675e+00, -1.16267307e+00, -6.53912840e-01, ...,
         -8.35531692e-01,  1.93309133e+00, -1.93309133e+00],
       ...,
       [ 1.63592675e+00,  1.99100483e+00,  3.49060516e+00, ...,
          1.19684269e+00, -5.17306132e-01,  5.17306132e-01],
       [ 1.67441914e+00,  1.99100483e+00,  1.00389436e+00, ...,
          1.19684269e+00, -5.17306132e-01,  5.17306132e-01],
       [ 1.71291154e+00, -5.19213966e-01, -6.53912840e-01, ...,
         -8.35531692e-01, -5.17306132e-01,  5.17306132e-01]])
```

We split the data into training and testing data using the function `train_test_split`. The training data is divided into validation data, a second set used for training data; then the models are trained and hyperparameters are selected using the function `GridSearchCV`.

TASK 3

Use the function train_test_split to split the data X and Y into training and test data. Set the parameter test_size to 0.2 and random_state to 2. The training data and test data should be assigned to the following labels.

`X_train, X_test, Y_train, Y_test`

In [411…

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_
```

we can see we only have 18 test samples.

In [412…

```
Y_test.shape
```

Out[412]:

```
(18,)
```

In [413…

```
X_train.shape, Y_train.shape
```

Out[413]:

```
((72, 83), (72,))
```

In [414…

```
X_test.shape, Y_test.shape
```

Out[414]:

```
((18, 83), (18,))
```

TASK 4

Create a logistic regression object then create a GridSearchCV object `logreg_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters` .

We output the `GridSearchCV` object for logistic regression. We display the best parameters using the data attribute `best_params_` and the accuracy on the validation data using the data attribute `best_score_` .

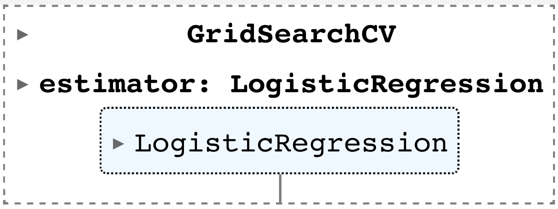
In [415…

```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

In [416…

```
lr=LogisticRegression()
logreg_cv = GridSearchCV(lr, parameters, cv=10)
logreg_cv.fit(X_train, Y_train)
```

Out[416]:



```
▸          GridSearchCV
▸ estimator: LogisticRegression
      ▸ LogisticRegression
```

In [417…

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

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

TASK 5

Calculate the accuracy on the test data using the method `score` :

In [418…

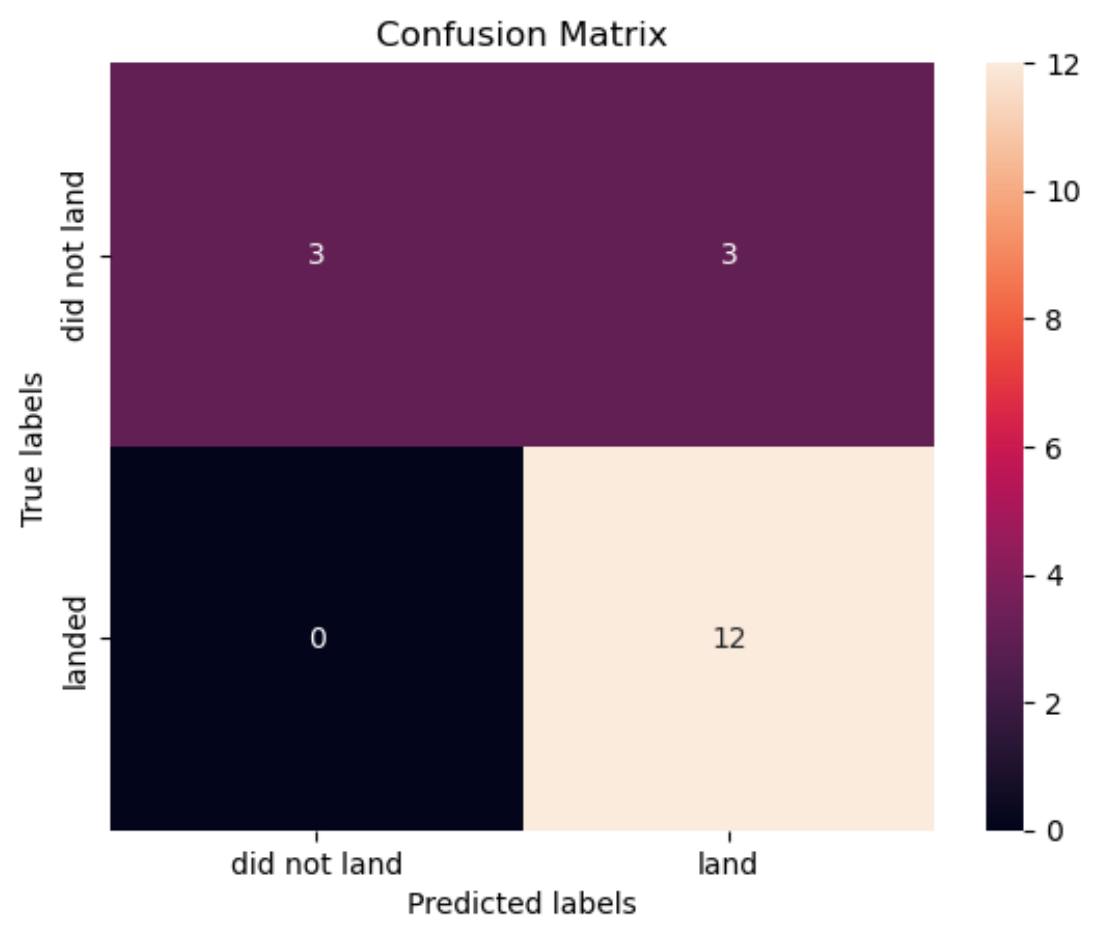
```
accuracy_score_logreg = logreg_cv.score(X_test, Y_test)
print("Accuracy on test data using the method score is: ", accuracy_score_logreg
```

```
Accuracy on test data using the method score is:  0.8333333333333334
```

Lets look at the confusion matrix:

In [419…

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

TASK 6

Create a support vector machine object then create a `GridSearchCV` object `svm_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters`.

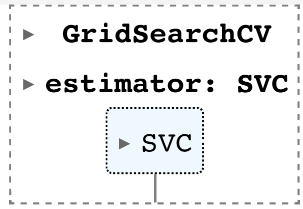
In [420…

```
parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'),
              'C': np.logspace(-3, 3, 5),
              'gamma':np.logspace(-3, 3, 5)}
svm = SVC()
```

In [421…

```
svm_cv = GridSearchCV(svm, parameters ,cv=10)
svm_cv.fit(X_train,Y_train)
```

Out[421]:



In [422…

```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.0316227766016837
9, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

TASK 7

Calculate the accuracy on the test data using the method `score` :

In [423…

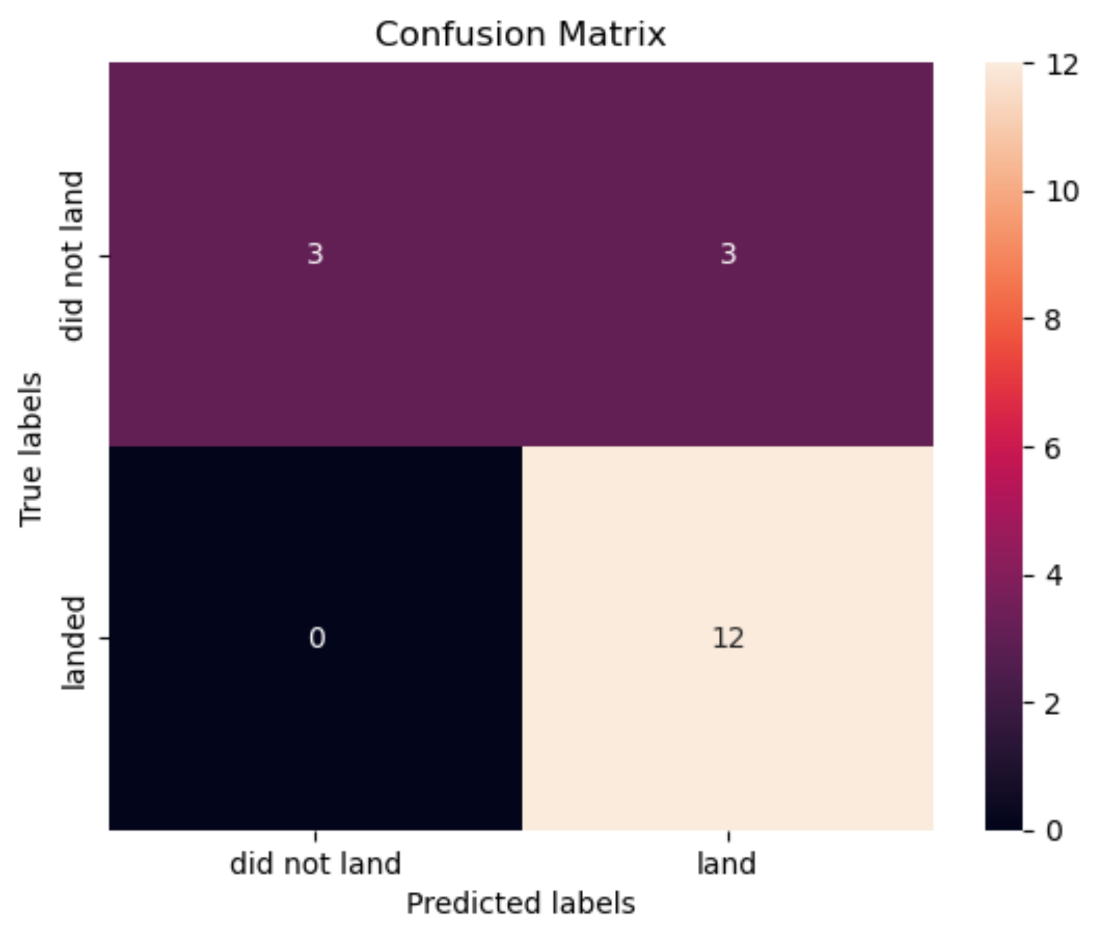
```
accuarcy_svm_test = svm_cv.score(X_test, Y_test)
print("Accuracy on the test data is:", accuarcy_svm_test)
```

```
Accuracy on the test data is: 0.8333333333333334
```

We can plot the confusion matrix

In [424…

```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



TASK 8

Create a decision tree classifier object then create a `GridSearchCV` object `tree_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters` .

In [425…

```
parameters = {'criterion': ['gini', 'entropy'],
     'splitter': ['best', 'random'],
     'max_depth': [2*n for n in range(1,10)],
     'max_features': ['sqrt', 'log2'],
     'min_samples_leaf': [1, 2, 4],
     'min_samples_split': [2, 5, 10]}

tree = DecisionTreeClassifier()
```

In [426…

```
tree_cv = GridSearchCV(tree, parameters, cv=10)
tree_cv.fit(X_train, Y_train)
```

Out[426]:

```
▸              GridSearchCV
▸ estimator: DecisionTreeClassifier
       ▸ DecisionTreeClassifier
```

In [427…

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth':
8, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'split
ter': 'best'}
accuracy : 0.8892857142857142
```

TASK 9

Calculate the accuracy of tree_cv on the test data using the method `score` :

In [428…

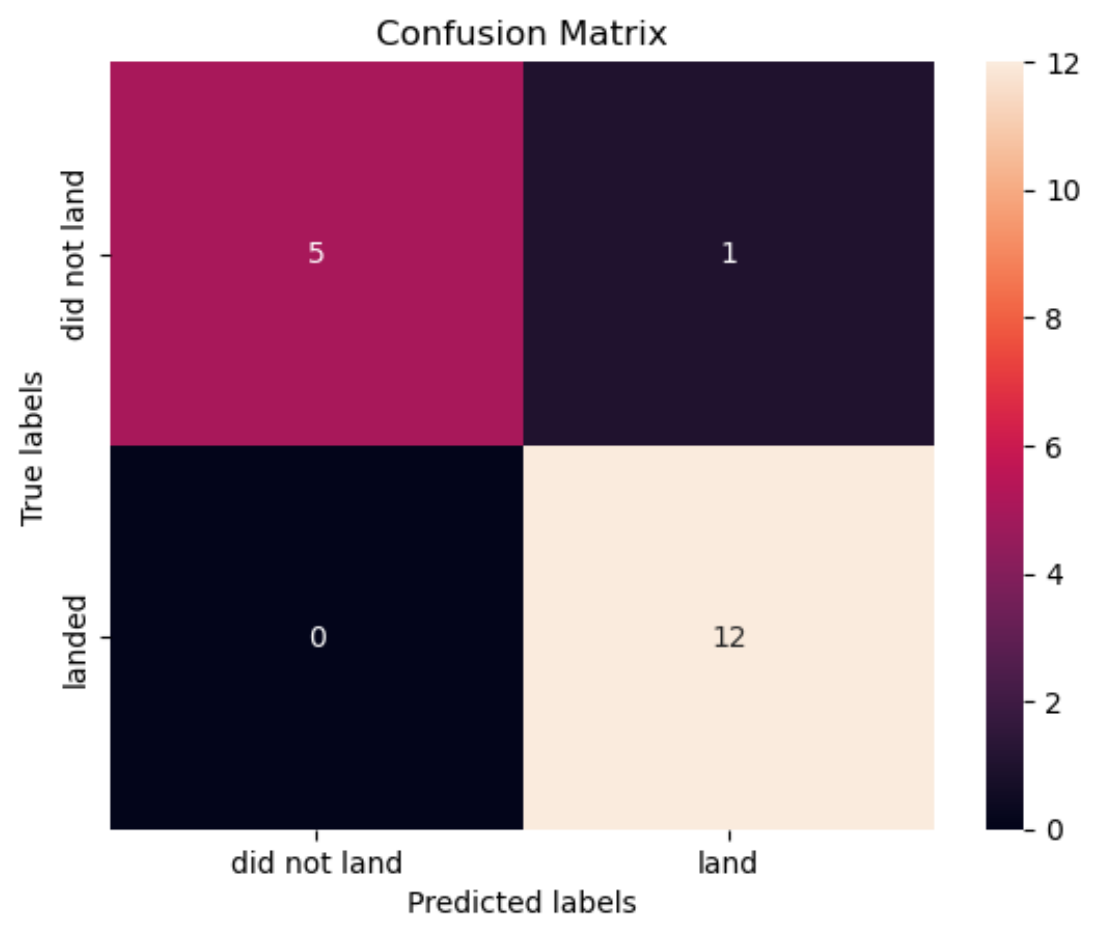
```
accuracy_tree_test = tree_cv.score(X_test, Y_test)
print("Accuracy on test data :", accuracy_tree_test)
```

```
Accuracy on test data : 0.9444444444444444
```

We can plot the confusion matrix

In [429…

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



TASK 10

Create a k nearest neighbors object then create a `GridSearchCV` object `knn_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters`.

In [430…

```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()
```

In [431…

```
knn_cv = GridSearchCV(KNN, parameters, scoring='accuracy', cv=10)
knn_cv = knn_cv.fit(X_train, Y_train)
```

In [432…

```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 1
0, 'p': 1}
accuracy : 0.8482142857142858
```

TASK 11

Calculate the accuracy of tree_cv on the test data using the method `score`:

In [433…

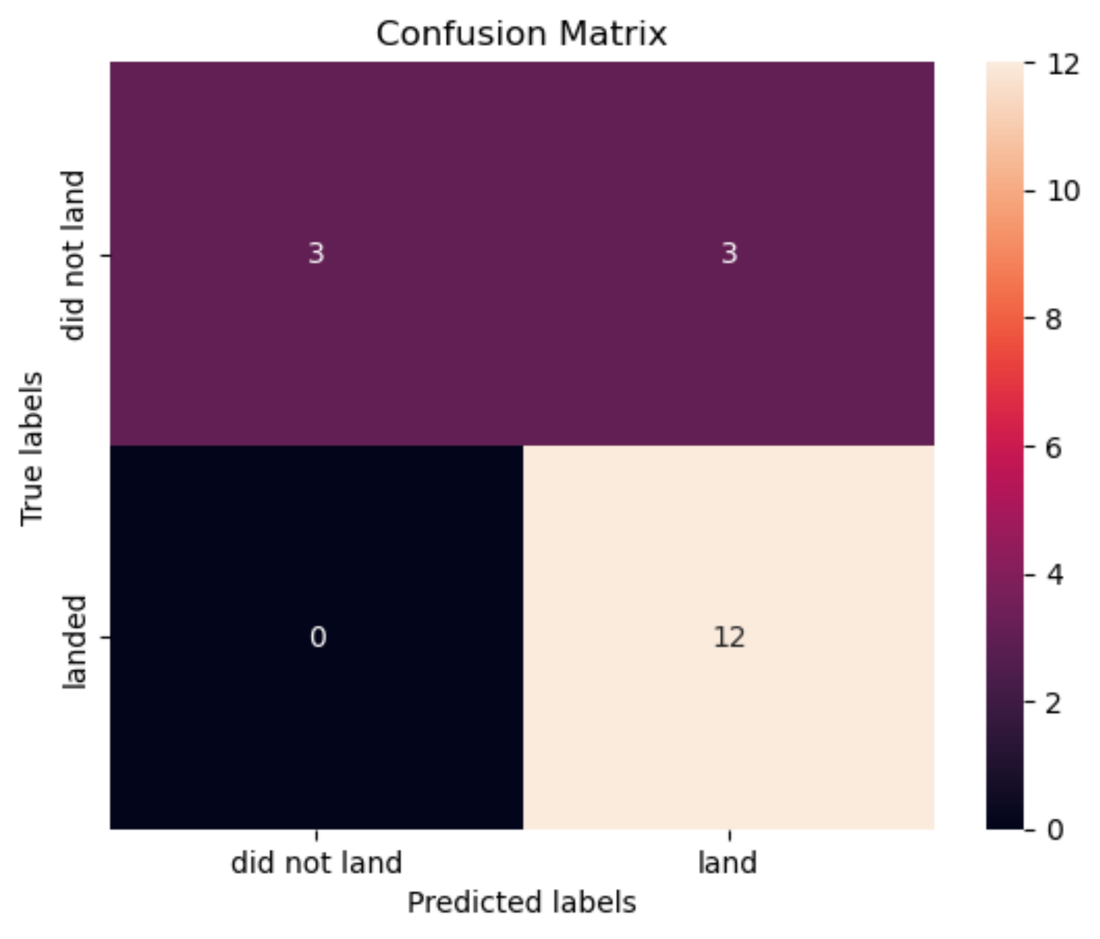
```
accuracy_score_test_knn = knn_cv.score(X_test, Y_test)
print("Accuracy on test data :", accuracy_score_test_knn)
```

```
Accuracy on test data : 0.8333333333333334
```

We can plot the confusion matrix

```
In [434…
```

```python
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



TASK 12

Find the method that performs best:

```
In [435…
```

```python
methods = ['Logreg','Svm','Tree','Knn']
accs_train = [logreg_cv.best_score_, svm_cv.best_score_, tree_cv.best_score_, kn
accs_test = [accuracy_score_logreg, accuarcy_svm_test, accuracy_tree_test, accur

dict_meth_accs = {}

for i in range(len(methods)):
    dict_meth_accs[methods[i]] = [accs_train[i], accs_test[i]]

df = pd.DataFrame.from_dict(dict_meth_accs, orient='index')
df.rename(columns={0: 'Accuracy Train', 1: 'Accuracy Test'}, inplace = True)

df.head()
```

Out[435]:

	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.889286	0.944444
Knn	0.848214	0.833333

Authors

Joseph Santarcangelo has a PhD in Electrical Engineering, his research focused on using machine learning, signal processing, and computer vision to determine how videos impact human cognition. Joseph has been working for IBM since he completed his PhD.

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-08-31	1.1	Lakshmi Holla	Modified markdown
2020-09-20	1.0	Joseph	Modified Multiple Areas