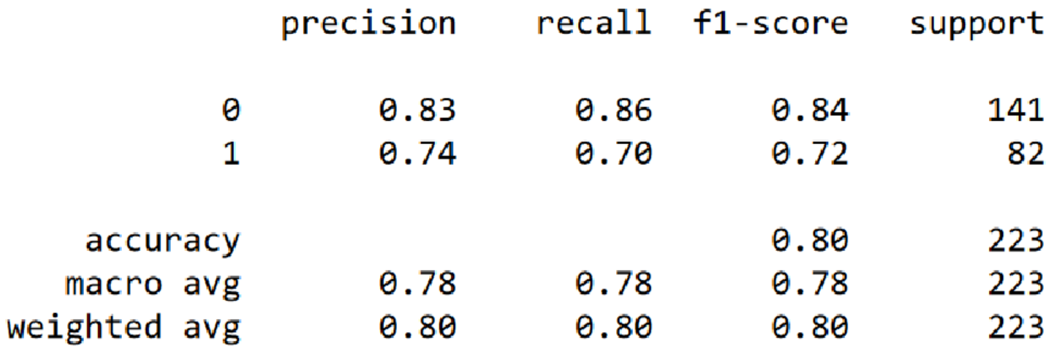
Question

Create a model that predicts whether emails are spam/not spam based on the words used in the email. To do this, you use information from a data set in which the output variable is Spam (0 = no spam, and 1 = spam) and a set of k input variables X (word categories) is included. You program in the programming language Python.

i. The data is available as a spam.csv -file. What are the programming commands for the following: 1. Load the data into Python. 2. Display the first 10 lines on the screen. 3. Show descriptive statistics of the data set.

ii. You estimate the model using logistic regression. The output variable is available as Y, and the matrix of predictors as X. What are the programming commands for splitting the data set into training and test data set, for using logistic regression and for fitting the model to the training data?

iii. You receive the following output for the performance of the model:



```
              precision    recall  f1-score   support

           0       0.83      0.86      0.84       141
           1       0.74      0.70      0.72        82

    accuracy                           0.80       223
   macro avg       0.78      0.78      0.78       223
weighted avg       0.80      0.80      0.80       223
```

Which programming command do you use to generate this output in Python? Explain the concepts of precision and recall and the connection to true/ false positives or negatives.

Solution

i.
```
dataset = pd.read_csv(“C:/spam.csv”)

dataset.head(10)

dataset.describe()
```

ii.
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = .25, random_state=25)

logit = LogisticRegression()

logit.fit(X_train, Y_train)
```

iii.
```
print(metrics.classification_report(Y_test, Y_pred))
```

Precision is the share of predicted positives which were correctly classified (TP / P*). Recall is the share of true positives which were correctly classified (TP / P).