Artificial Intelligence Mini Project

Team Members

1. P.S.A.BHASKAR REDDY - [072]
2. V.R.S.MURTHY REDDY - [075]

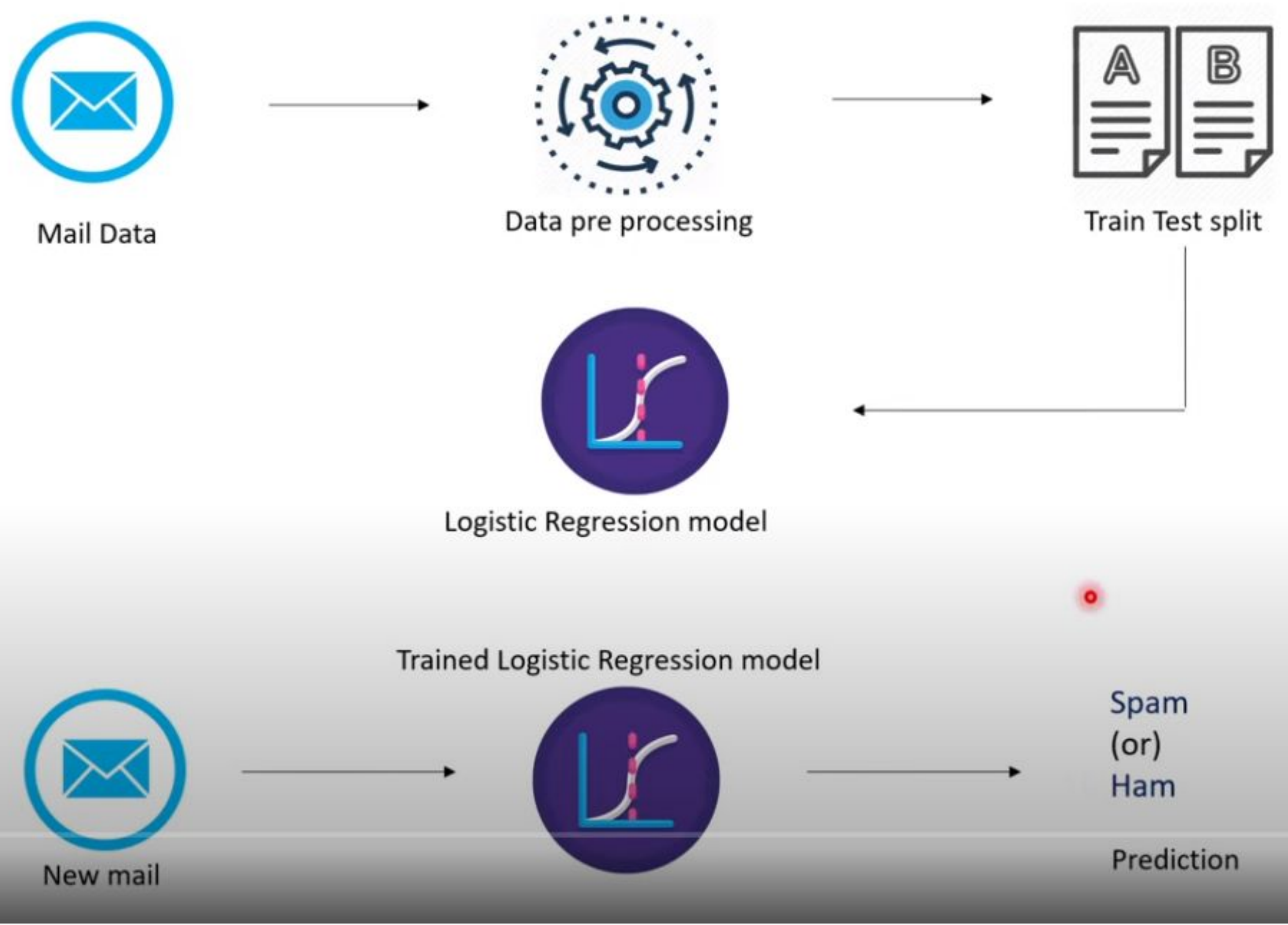
Topic : Gmail Spam Detection

Gmail is the worldwide use of communication application. It is because of the ease of use and faster than other communication application. However, its inability to detect whether the mail content is either spam or ham degrade its performance. Nowadays, lot of cases have been reported regarding stealing of personal information or phishing activities via gmail from the user. This project will discuss how machine learning help in spam detection. Machine learning is an artificial intelligence application that provides the ability to automatically learn and improve data without being explicitly programmed. The algorithm will predict the score more accurately. The objective of developing this model is to detect and score word faster and accurately.

Objective :

The objective of this project is to build a prediction model to predict whether a mail is spam or not.

Work Flow


Mail Data
Data pre processing
Train Test split
Logistic Regression model
Trained Logistic Regression model
New mail
Spam
(or)
Ham
Prediction

Label Encoding

STEPS INVOLVED:

1.Importing Dataset

2.Preprocessing Dataset

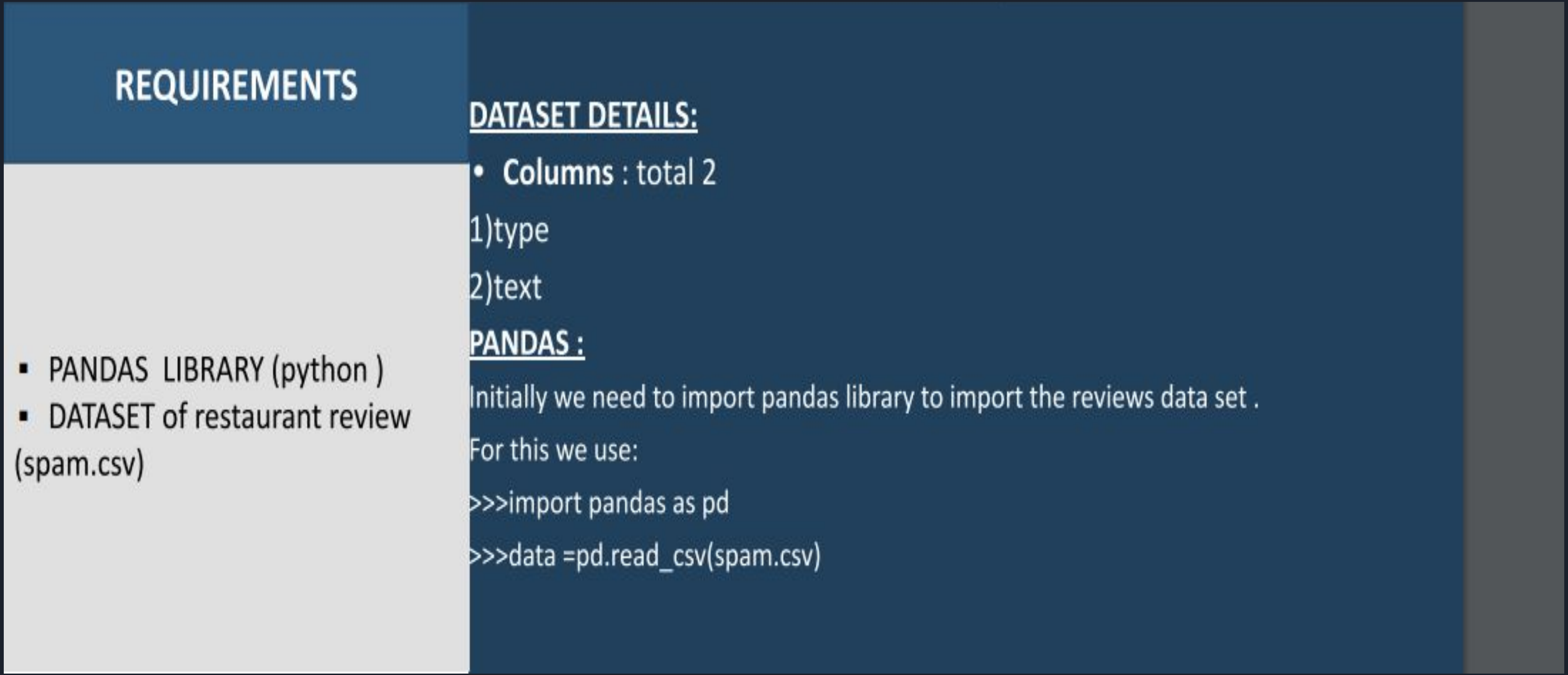
3.Label Encoding

4.Splitting the data into training data & test data

5.Logistic regression

6.Evaluating the trained model

1.Importing Dataset



REQUIREMENTS

- PANDAS LIBRARY (python)
- DATASET of restaurant review (spam.csv)

DATASET DETAILS:

- **Columns** : total 2

1)type

2)text

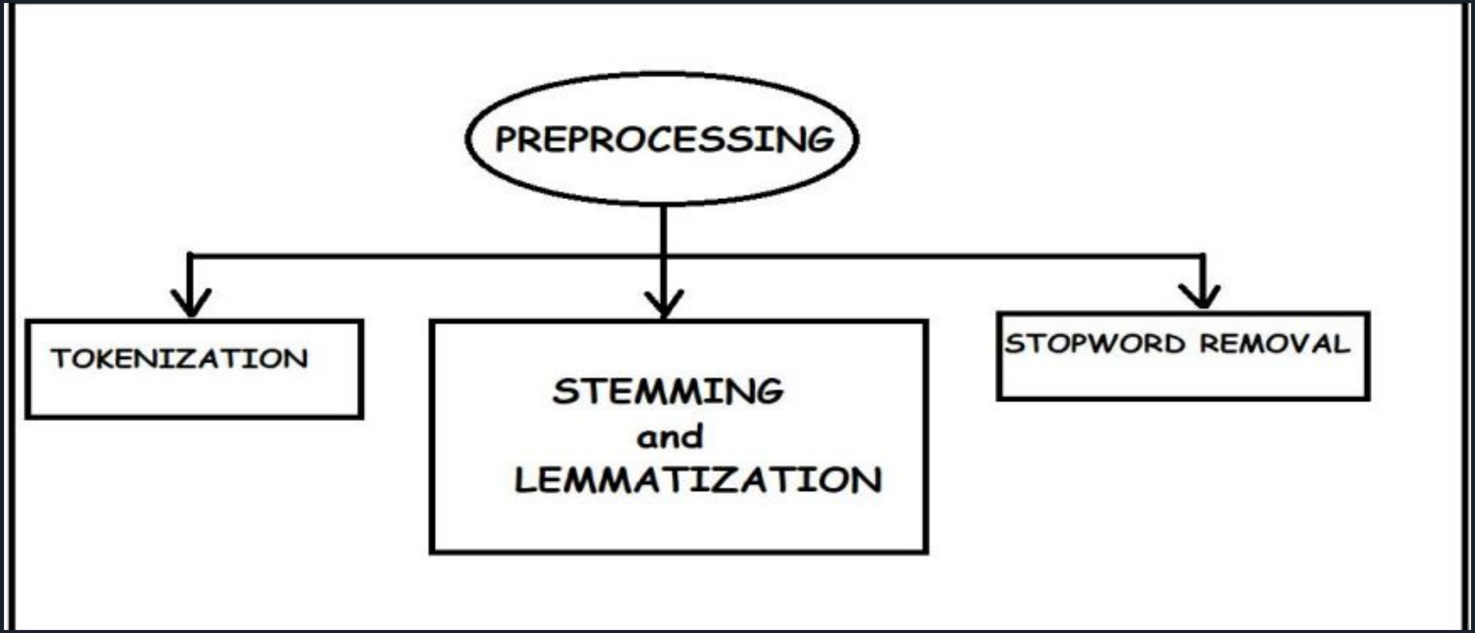
PANDAS :

Initially we need to import pandas library to import the reviews data set .

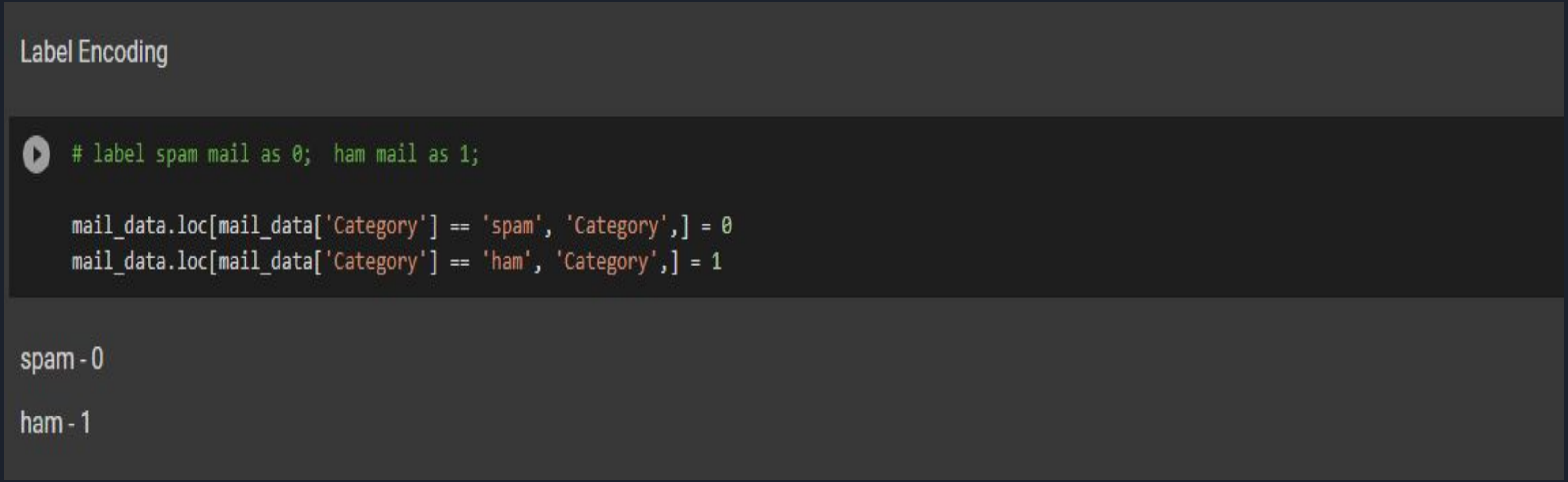
For this we use:

```
>>>import pandas as pd
>>>data =pd.read_csv(spam.csv)
```

2.Preprocessing Dataset



3.Label Encoding



Label Encoding

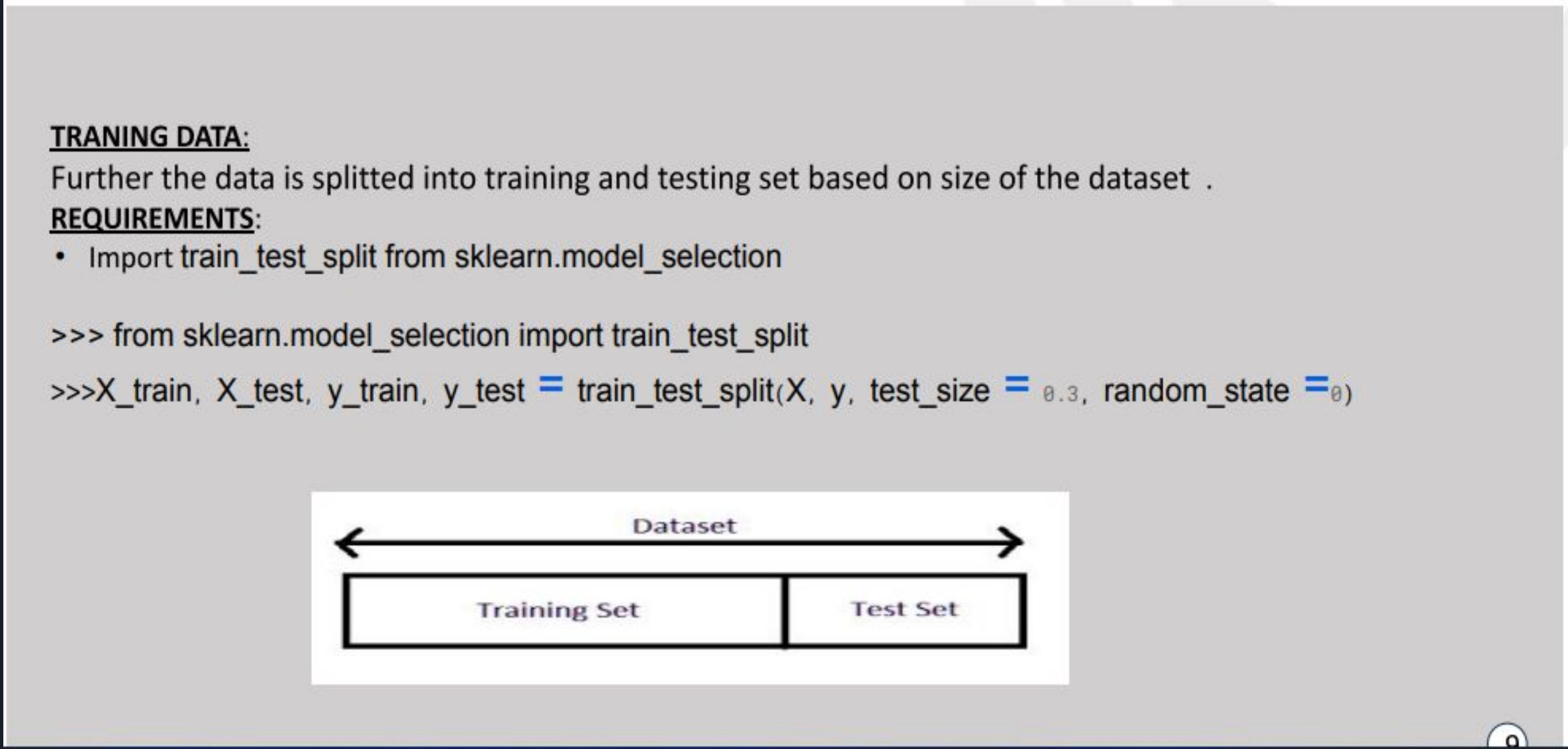
```
# label spam mail as 0;  ham mail as 1;

mail_data.loc[mail_data['Category'] == 'spam', 'Category',] = 0
mail_data.loc[mail_data['Category'] == 'ham', 'Category',] = 1
```

spam - 0

ham - 1

4.Splitting the data into training data & test data



TRANING DATA:

Further the data is splitted into training and testing set based on size of the dataset .

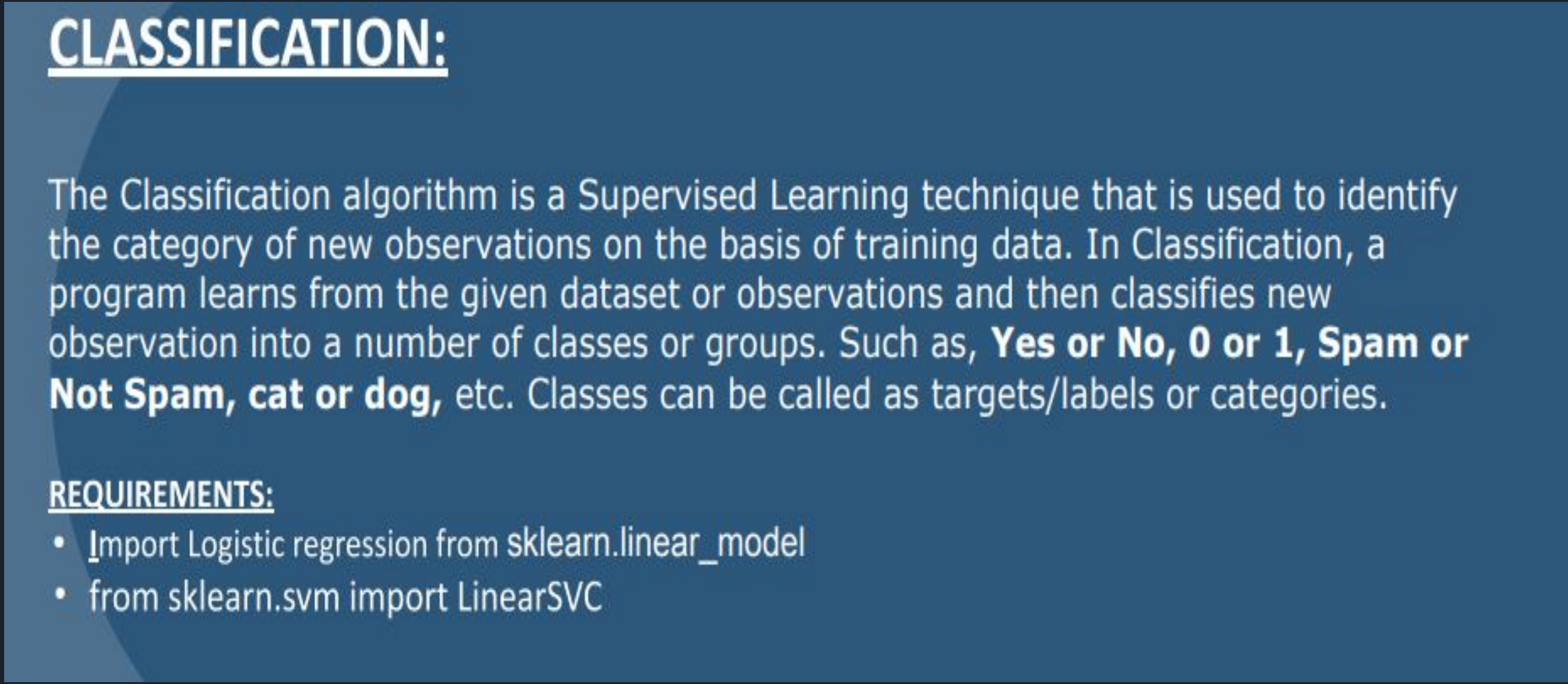
REQUIREMENTS:

- Import train_test_split from sklearn.model_selection

```
>>> from sklearn.model_selection import train_test_split
>>>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state =0)
```

5.**Logistic Regression** :

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.



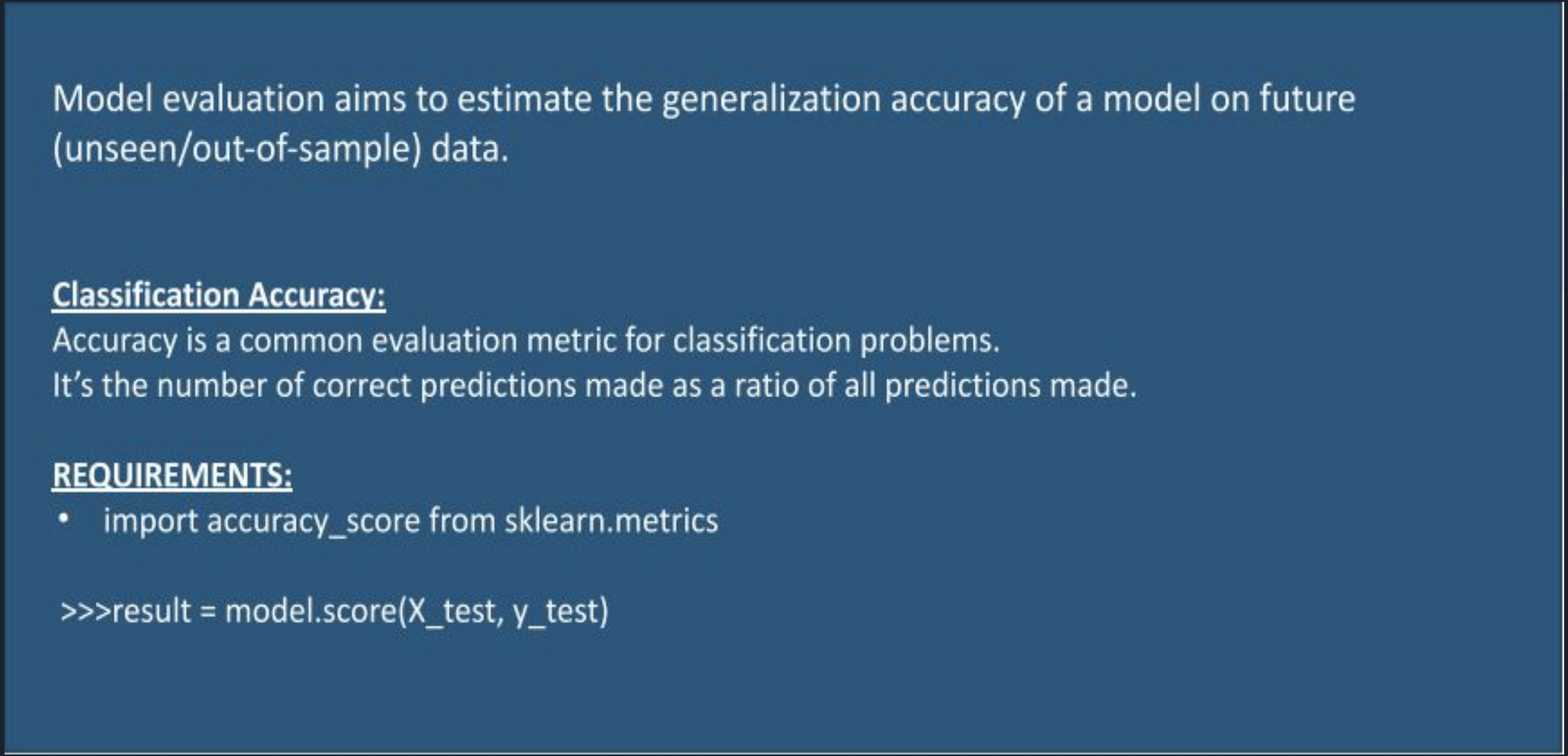
CLASSIFICATION:

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

REQUIREMENTS:

- Import Logistic regression from sklearn.linear_model
- from sklearn.svm import LinearSVC

6.Evaluating the trained model



Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.

Classification Accuracy:

Accuracy is a common evaluation metric for classification problems.
It's the number of correct predictions made as a ratio of all predictions made.

REQUIREMENTS:

- import accuracy_score from sklearn.metrics

```
>>>result = model.score(X_test, y_test)
```

Thank
you