NAME: SUNKARI KRISHNAVENI

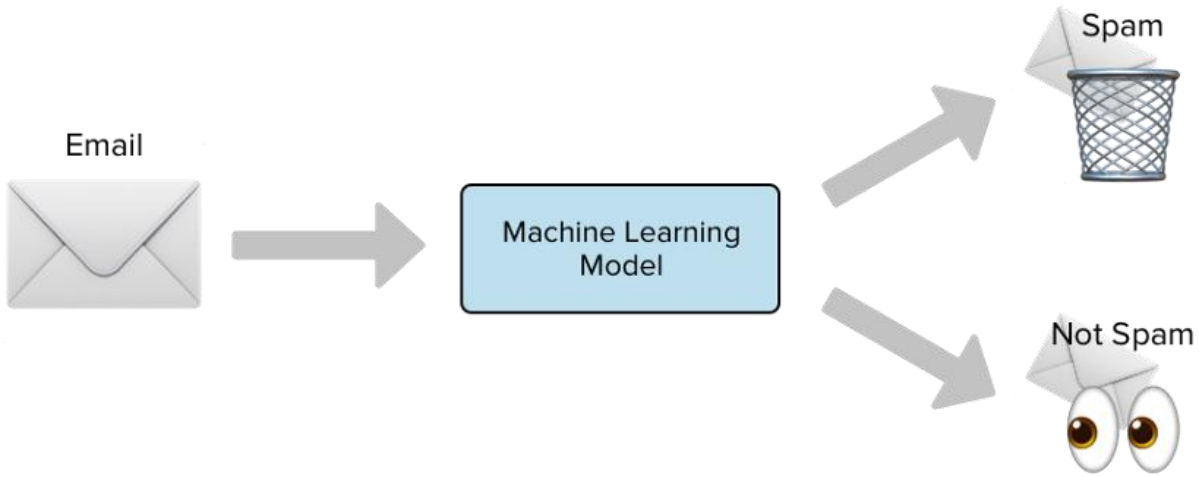
EMAIL: 190101120070@cutm.ac.in

Task 6: Data Science Example

Implement a sample machine learning program for a problem statement of your choice.

Solution 6:

Problem Statement: Spam Mail Detection using Support Vector Machine



SOFTWARE USED: Anaconda, Jupyter Notebook

LANGUAGE USED: Python

STEPS:

1) Download the dataset containing emails of **Spam** and **Non-Spam.**
2) In the dataset, there are two features:
   a. Label – Ham or Spam
   b. Email text – Actual Email

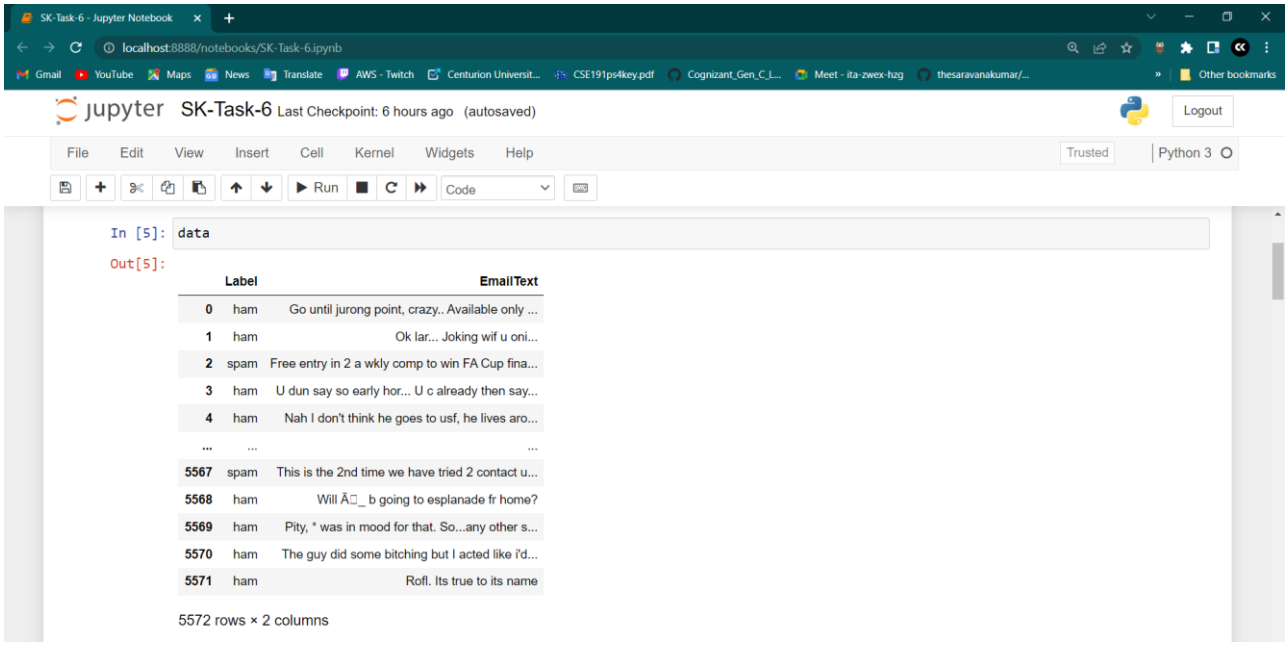
3) I will use **SVM algorithm** to build the model and that will recognize the pattern and will predict whether the mail is spam or genuine**.** "**Support Vector Machine**" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used for in classification problems. In the SVM algorithm, we will plot each data item as a point in n-dimensional space (where n is a number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.
4) At first, open the Jupyter Notebook app, then it will open the editor in your default browser.
5) Import **Important Libraries**

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
from sklearn import svm
```

6) Load the **dataset**

```
data = pd.read_csv('spam.csv')
```

READ THE DATA:



	Label	EmailText
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will Ã□_ b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

5572 rows × 2 columns

7) Retrieve the information of the dataset.

data.info()

```
In [8]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Label      5572 non-null   object
 1   EmailText  5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

8) Split the data into **X** and **y.**

X = data['EmailText'].values

y = data['Label'].values

9) Split the data into **training** and **testing.**

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=0)
```

10) Now, convert the **text** into **integer** using **CountVectorizer()**

```
cv = CountVectorizer()
X_train = cv.fit_transform(X_train)
X_test = cv.transform(X_test)
```

11) Apply **SVM algorithm**

```
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 10)
classifier.fit(X_train, y_train)
```

12) Check the **accuracy**.

```
print(classifier.score(X_test,y_test))
```

OUTPUT:- 0.9766816143497757

This produces the best accuracy for my model and it will predict best results.