BIG DATA

Group

-Carlos Hernandez ID: 986636
-Sebastian Valencia ID: 986775

Naive Bayes Algorithm Tutorial

What you Will see(?)

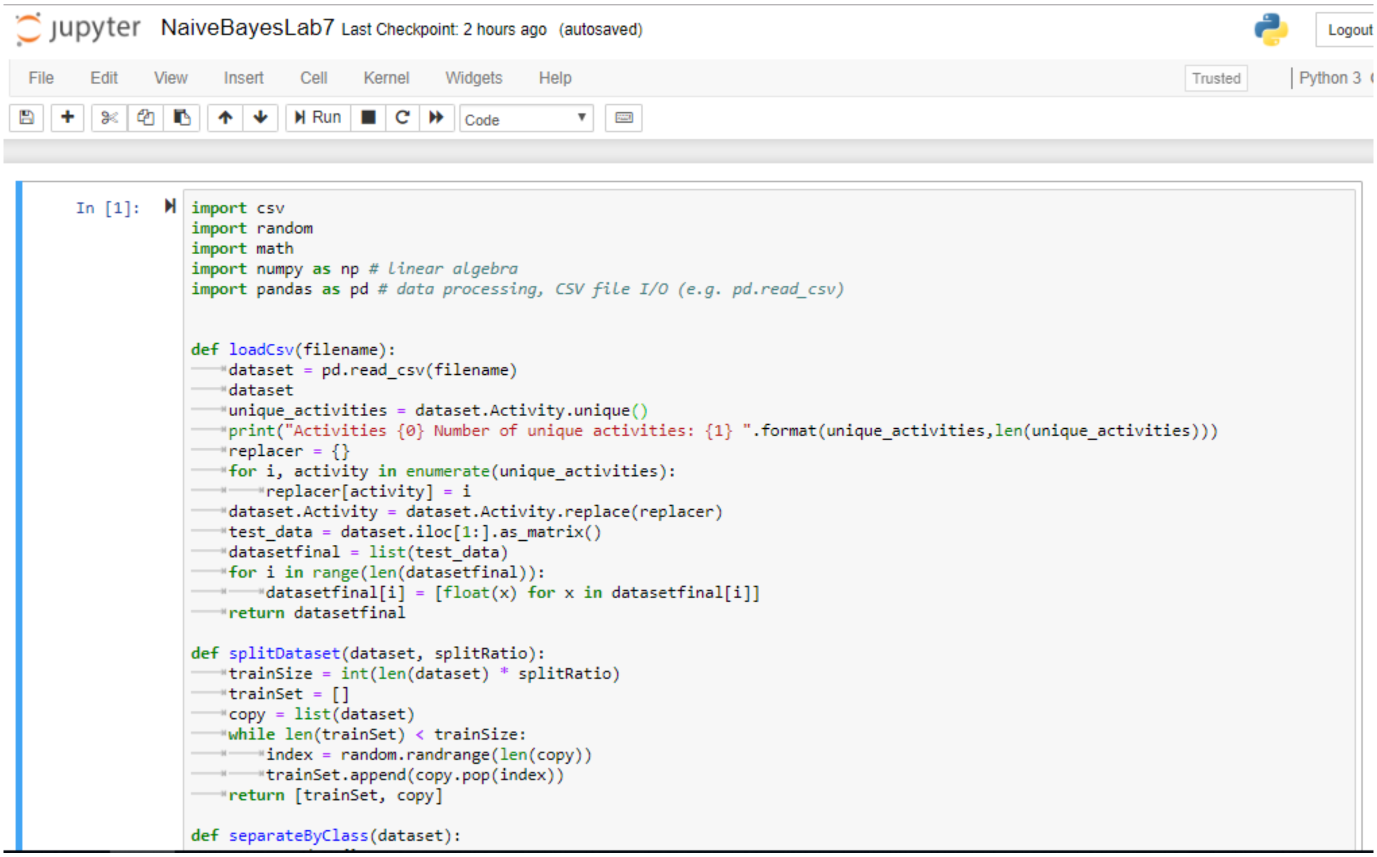
- **Handle Data**: Load the data from CSV file and split it into training and test datasets (67%,33% respectively).
- **Summarize Data**: summarize the properties in the training dataset so that we can calculate probabilities and make predictions.
- **Make a Prediction**: Use the summaries of the dataset to generate a single prediction.
- **Make Predictions**: Generate predictions given a test dataset and a summarized training dataset.
- **Evaluate Accuracy**: Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

Prerequisites

Be sure that you already has installed the next programs in your machine

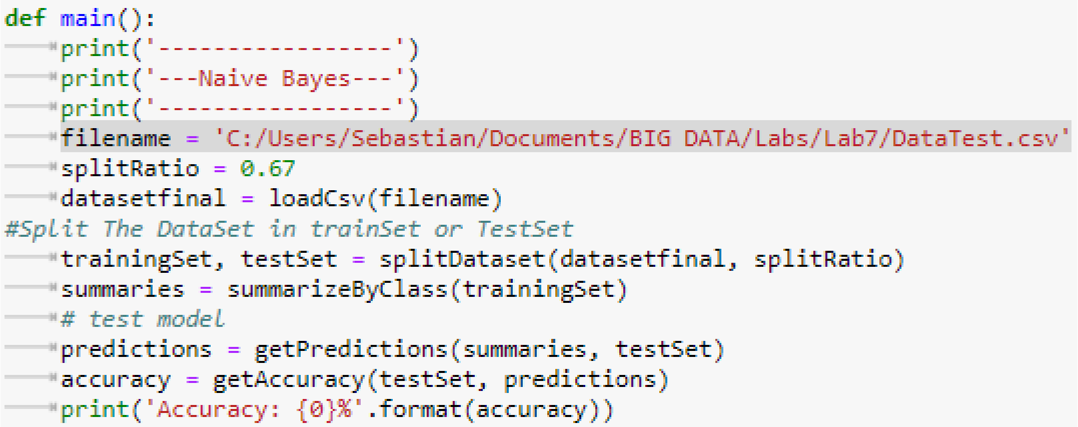
-Anaconda3 (python 3.5.1)
-Jupyter Notebook
-These librarys (csv, random, math, numpy, pandas)

Open the file 'NaiveBayesLab7' in Jupyter Notebook. (You should see this)



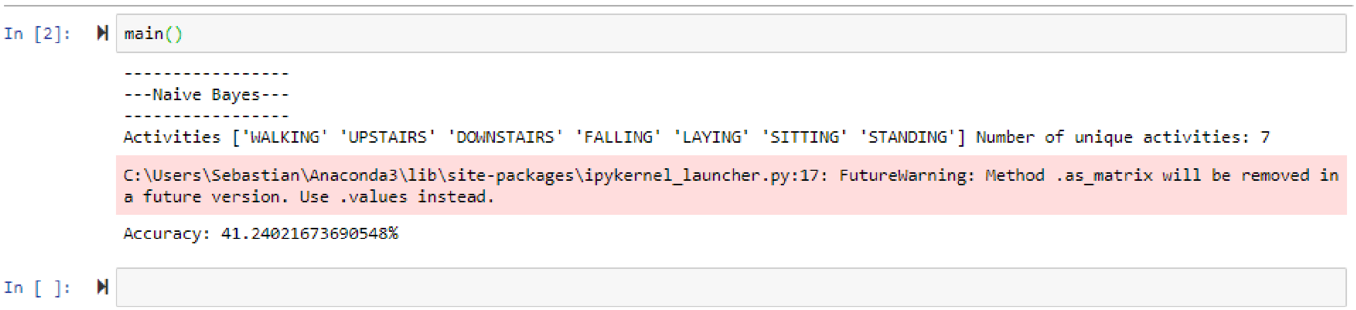
```
import csv
import random
import math
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)


def loadCsv(filename):
    dataset = pd.read_csv(filename)
    dataset
    unique_activities = dataset.Activity.unique()
    print("Activities {0} Number of unique activities: {1} ".format(unique_activities,len(unique_activities)))
    replacer = {}
    for i, activity in enumerate(unique_activities):
        replacer[activity] = i
    dataset.Activity = dataset.Activity.replace(replacer)
    test_data = dataset.iloc[1:].as_matrix()
    datasetfinal = list(test_data)
    for i in range(len(datasetfinal)):
        datasetfinal[i] = [float(x) for x in datasetfinal[i]]
    return datasetfinal

def splitDataset(dataset, splitRatio):
    trainSize = int(len(dataset) * splitRatio)
    trainSet = []
    copy = list(dataset)
    while len(trainSet) < trainSize:
        index = random.randrange(len(copy))
        trainSet.append(copy.pop(index))
    return [trainSet, copy]

def separateByClass(dataset):
```

You should change the path of the file



```
def main():
    print('-----------------')
    print('---Naive Bayes---')
    print('-----------------')
    filename = 'C:/Users/Sebastian/Documents/BIG DATA/Labs/Lab7/DataTest.csv'
    splitRatio = 0.67
    datasetfinal = loadCsv(filename)
#Split The DataSet in trainSet or TestSet
    trainingSet, testSet = splitDataset(datasetfinal, splitRatio)
    summaries = summarizeByClass(trainingSet)
    # test model
    predictions = getPredictions(summaries, testSet)
    accuracy = getAccuracy(testSet, predictions)
    print('Accuracy: {0}%'.format(accuracy))
```

# Run code (should look something like this)



```
In [2]: main()

-----------------
---Naive Bayes---
-----------------
Activities ['WALKING' 'UPSTAIRS' 'DOWNSTAIRS' 'FALLING' 'LAYING' 'SITTING' 'STANDING'] Number of unique activities: 7

C:\Users\Sebastian\Anaconda3\lib\site-packages\ipykernel_launcher.py:17: FutureWarning: Method .as_matrix will be removed in
a future version. Use .values instead.

Accuracy: 41.24021673690548%

In [ ]:
```

Enjoy !!! :D