MACHINE LEARNING PROJECT PHASE 2

SUPERVISED/Unsupervised LEARNING ALGORITHMS

1. Problem Definition

- The problem statement which we are dealing with is ***Sentiment analysis*** .We want to analyse huge volumes of data and want to detect the sentiment of it whether it is positive or negative

  TASK(T): To classify if a given text/sentence is positive or negative

  EXPERIENCE(E): Corpus files having both positives and negatives

  PERFORMANCE(P): Accuracy score.Accuracy is used as a score of performance

2. Datasets

- ***Restaurant reviews:*** Dataset having reviews from restaurants. The training set contains reviews as well as their labels, whereas the testing set only reviews.

3. Prepare Data

- As our input data is text,we used text related preprocessing
- In preprocessing step we have done
  - Removal of stopwords,wild characters,converting uppercase to lowercase letters.
  - Stemming,Tf-IDF/ bag-of-words

- “Stop words” are commonly used words that are unlikely to have any benefit in natural language processing.So remove them and wild characters
- TF-IDF is a statistical measure that evaluates how relevant a word is to a document in collection of documents
- Removed null values from the dataset
- Removed duplicates
- Applied standardization

4. Python packages :

- Numpy and Pandas
- Regular expressions
- Scikit-learn package for including various classification algorithms like Naive bayes , SVM, logistic regression.
- NLTK package for text preprocessing

  example removing stopwords, stemming ..

5. Supervised Learning Algorithms :

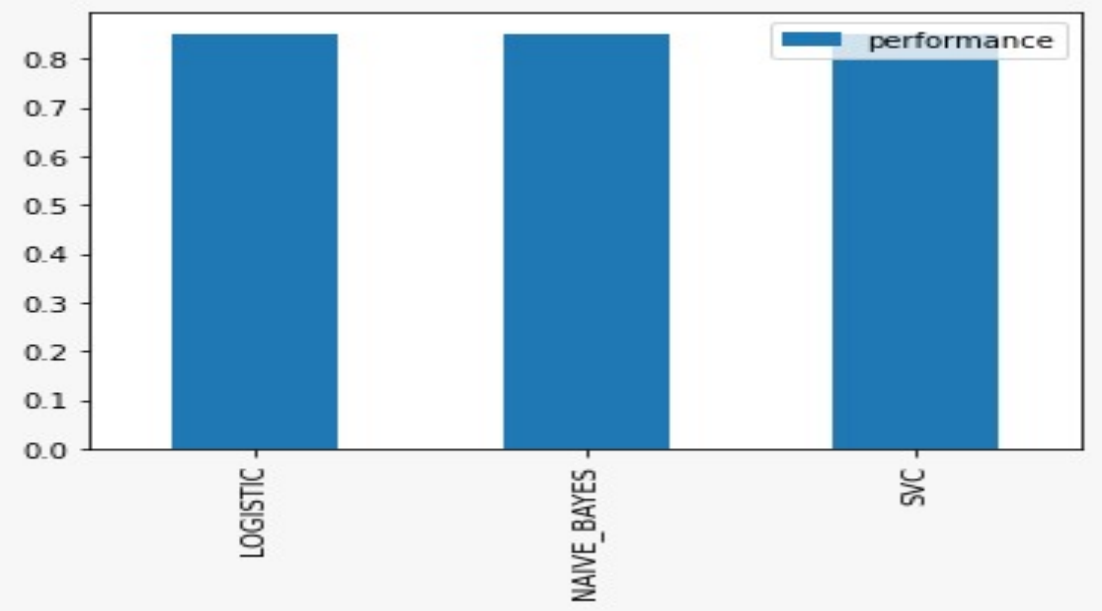
- **Naive Bayes:***Naive Bayes is used in text classification problems.It predicts probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class and is ruled as true and less likely class is ruled as fake*
- **Logistic Regression:***Logistic regression is used to predict the probability of a target variable.It predicts based on probability.It estimates the probability of an event occurring having been given some previous data.It is used to predict the*

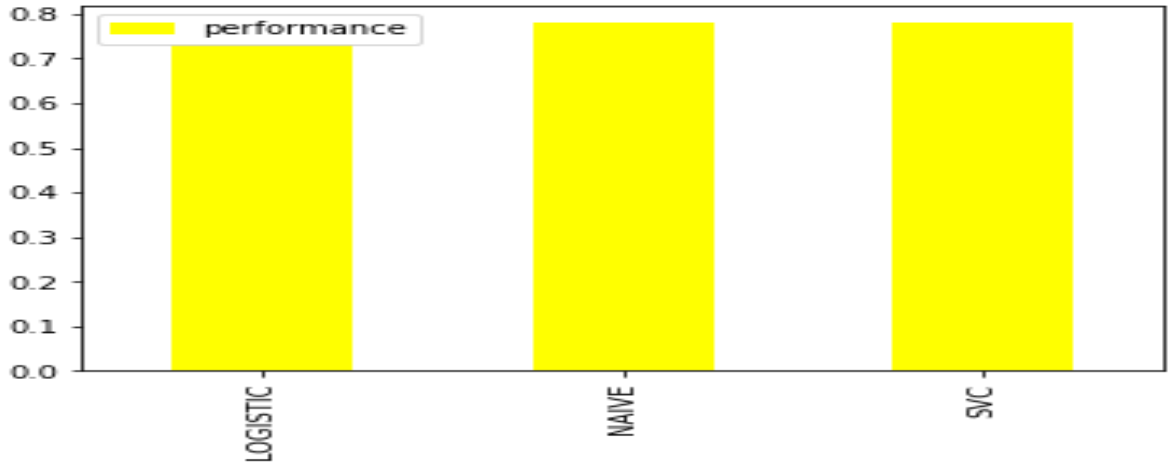
odds of the text data positive or negative

- **SVM :** *The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine*

The results of each of these classifiers on datasets are:

C.Sumanth Reddy:




performance
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
LOGISTIC
NAIVE
SVC