NAME: Venkata Sai Siva Pruthvi. Ustepalle S-ID: 800957126

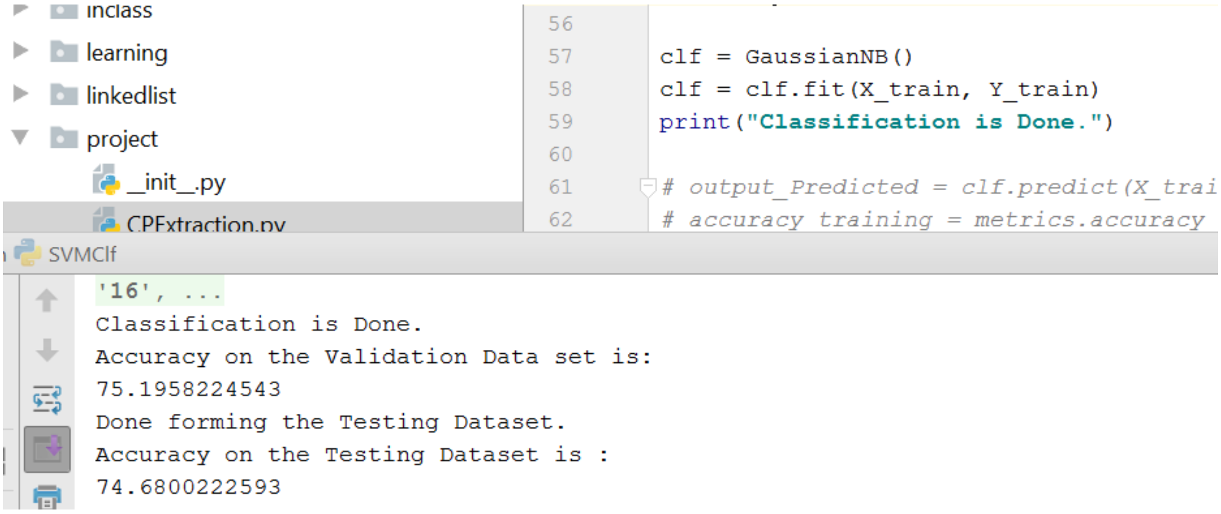
(Worked together with Chaitanya Sri Krishna Lolla)

Digit Recognition dataset

- RESULTS:

Splitting the given data into 2 parts to perform cross validation (as a step for pruning the data).

Training on anything below 60% for cross validation which is resulting in a 100% accuracy on training set and 74.6% accuracy on the validation data.



```
clf = GaussianNB()
clf = clf.fit(X_train, Y_train)
print("Classification is Done.")

# output_Predicted = clf.predict(X_trai
# accuracy training = metrics.accuracy
```

```
'16', ...
Classification is Done.
Accuracy on the Validation Data set is:
75.1958224543
Done forming the Testing Dataset.
Accuracy on the Testing Dataset is :
74.6800222593
```

Any parameter tuning did not perform better than default values for this dataset.

Expected low accuracy with Navies Bayes classifier as it depends on the prior classified results and each instance classification is interlinked.

Amazon review dataset

- Result:

Using nltk library services to lemmatize, tokenize the reviews along with removing the punctuation from the review texts as a part of preprocessing steps as in done in the previous assignment.

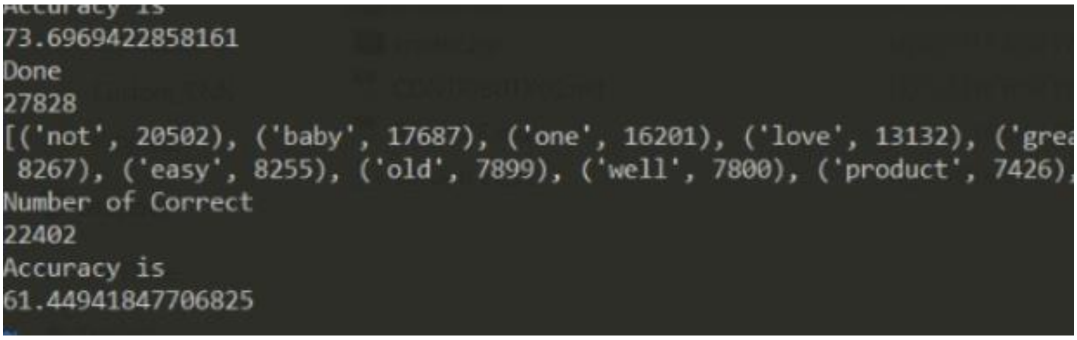
Pre-processing the text and making it into two sets of positive and negative reviews.

Picking top 5000 of the most commonly occurring words and assigning them numerical values to pass as parameters for the decision tree classifier method.

Used TF-IDF to build a matrix of numerical representation of the words in each sentence.

And the same procedure is repeated on the test dataset and the accuracy is measured.

- Used a normal GaussianNB() method and checked for the accuracy of model (which has default values as mentioned above) .



```
Accuracy is
73.6969422858161
Done
27828
[('not', 20502), ('baby', 17687), ('one', 16201), ('love', 13132), ('grea
 8267), ('easy', 8255), ('old', 7899), ('well', 7800), ('product', 7426),
Number of Correct
22402
Accuracy is
61.44941847706825
```

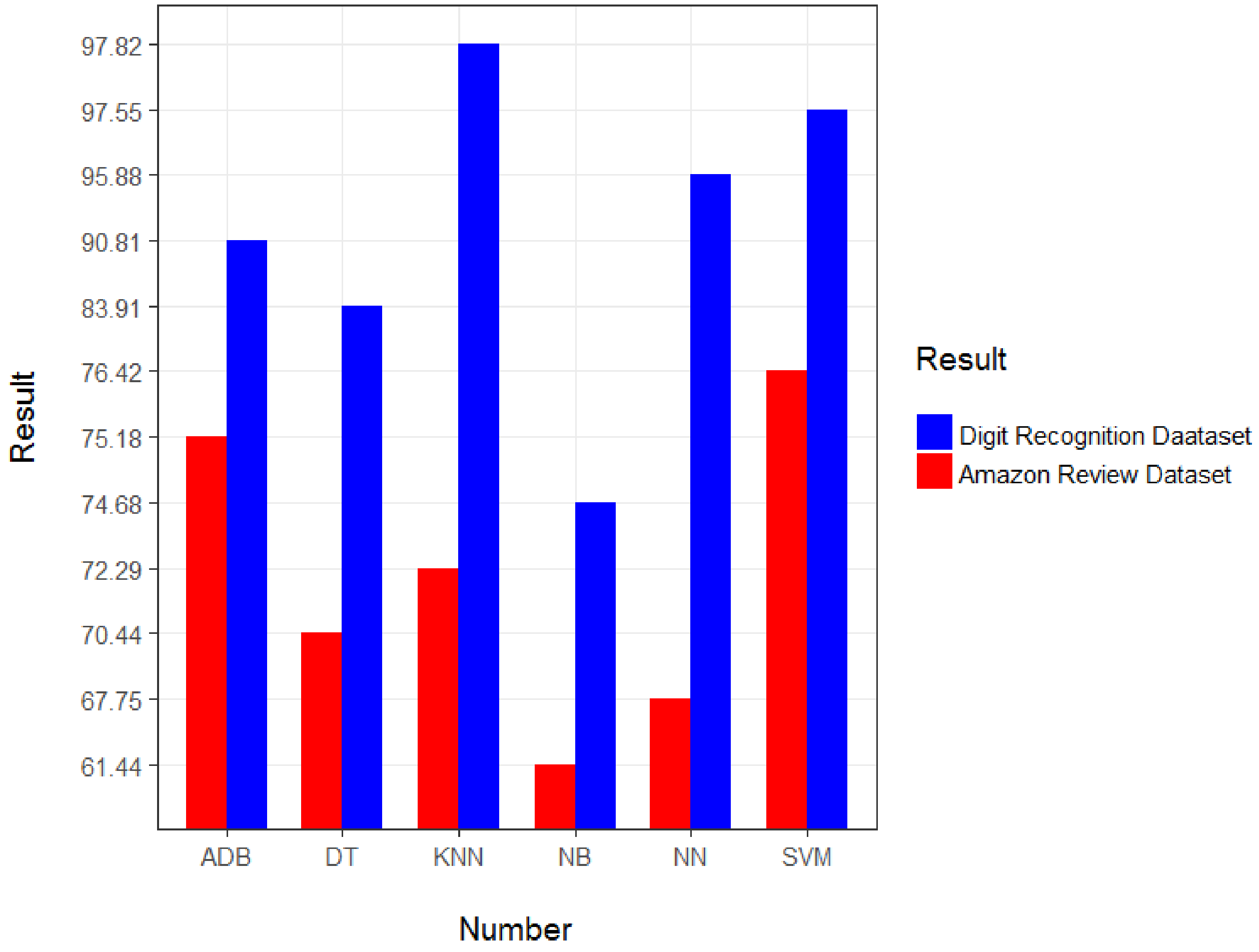
Even this dataset did not give any extraordinary results than default values of classifier.

It is quiet expecting from NB classifier to give a low accuracy as it need a well delineated data set to work up on.

Preprocessing of dataset is a relying point for best accuracy scores in NB algorithm.

Comparison of supervised machine learning algorithms implemented (Digit recognition, Amazon Reviews)   :

Decision Trees	83.91,70.44
Neural Networks	95.88,67.75
KNN	97.82,72.29
Ada boost	90.81,75.18
SVM	97.55,76.42
Naïve Bayes	74.68,61.44



Interpretation on each of the algorithms implemented

Decision trees:

- As a beginner algorithm in machine learning it helped me to learn basics of how learning happens in a machine.
- It performs best for a binary classifier as the probability between either of the classification is obvious when it’s a tree.
- It works for either classification or regression problems.
- It did not give a great accuracy when compared to other algorithms.
- Very fast performance time and need no tuning for most of the cases unless it is really a huge dataset with imbalanced number of classes.

Neural Networks:

- Neural Networks same works for both classification and regression tasks.
- It worked better than decision trees for digit recognition dataset with an accuracy of 95.88% which is a good accuracy and an accuracy of 67.75 on amazon review dataset.
- Neural networks should perform better on any kind of processes data given significant amount of time and fine tuning the parameters.
- With increase in number of hidden layers and change of activation functions it may be a slow running algorithm as it learns with every iteration.
- It took ages for running neural networks on Amazon dataset. 8 hours of wall clock running time on a High performance cluster for training the best possible classifier.

Ada Boost:

- Boosting algorithms are one popular kind used in many data science hackathons or for any data scientists.
- Concept of boosting is pragmatic when it comes to a real-world scenario as it is learning based on a voting system.
- Adaptive Boosting fits a sequence of weak learners on different weighted training data.
- It ca be used for both classification and regression problems.
- Results cannot be easily interpreted.
- It gives a good accuracy score. A real higher score with a slow training speed and a fast predicting speed.
- For Amazon dataset, it really depicted a higher accuracy value when compared to all other algorithms, though it did not perform extraordinary when compared to svm. An uncomplainably low difference in their accuracy scores made it stand second in performance.

K Nearest Neighbors:

- It works for both classification and regression.
- KNN is one algorithm which is relatively easily to understand and interpret it to others.
- It's like a commonsense algorithm which fetches instances closer to a classified instance and classifies these into a group based on more number of similar features.
- This is the reason why it stands third in classification of Amazon dataset classification.
- It performed well on Digit recognition dataset as it ignores the outlier data in each instance and classify them based on most commonly occurring features.
- Lesser calculation time.

Scalable Vector Machines:

- It the best classification algorithm I've come across amongst the implemented six.
- The concept of interpreting instances in a high dimension variable is one best of a classification we can expect from an algorithm.
- Basic depiction of SVM on internet is visualizing a three-feature instance on a X-Y axes plane is an easiest way to understand this algorithm.
- Concept of placing a hyper plane to classify instances makes predicting testing values really a cake walk.
- It performed extremely well on Amazon dataset and it gave higher accuracy as it has classified all the instances considering all possible features on a multidimensional plane and classified it into a good learner.

Naïve Bayes:

- It works on Bayes theorem with an assumption of independence among predictors.
- Naïve Bayes algorithm works good on large datasets. It outperforms even highly sophisticated classification methods.
- It provides a way of calculating posterior probability. (given probability of occurrence and predicting a new instance).
- This classification method is purely mathematical as it uses the concept of conditional probability.
- It took 3-4 hours to run this algorithm for Amazon dataset.

Overall understanding of performance of classification on given datasets:

- For Digit Recognition dataset, I observed that the different ways of writing a digit might become a problem for an algorithm to correctly classify input. Discarding outliers might really give best results.
- For Amazon dataset, more accuracy can be expected by correctly interpreting the textual reviews into numerical values. Count Vectorizer alone cannot perform the job well.