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

(Worked together with Chaitanya Sri Krishna Lolla)

# Digit Recognition dataset

Exploring the data:

- Checking the first 5 tuples of the data for an overview,

```
Console D:/pruthvi/Spring-17/ML/ITCS6156_SLProject/DigitRecognition/
> head(train_data,5)
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29
1  0  1  6 15 12  1  0  0  0   7  16   6   6  10   0   0   0   8  16   2   0  11   2   0   0   5  16   3   0
2  0  0 10 16  6  0  0  0  0   7  16   8  16   5   0   0   0  11  16   0   6  14   3   0   0  12  12   0   0
3  0  0  8 15 16 13  0  0  0   1  11   9  11  16   1   0   0   0   0   0   7  14   0   0   0   0   3   4  14
4  0  0  0  3 11 16  0  0  0   0   5  16  11  13   7   0   0   3  15   8   1  15   6   0   0  11  16  16  16
5  0  0  5 14  4  0  0  0  0   0  13   8   0   0   0   0   0   3  14   4   0   0   0   0   0   6  16  14   9
  V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 V41 V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56
1   5   7   0   0   7  13   3   0   8   7   0   0   4  12   0   1  13   5   0   0   0  14   9  15   9   0   0
2  11  11   0   0  12  12   0   0   8  12   0   0   7  15   1   0  13  11   0   0   0  16   8  10  15   3   0
3  12   2   0   0   1  16  16  16  16  10   0   0   2  12  16  10   0   0   0   0   0   2  16   4   0   0   0
4  16  10   0   0   1   4   4  13  10   2   0   0   0   0   0  15   4   0   0   0   0   0   3  16   0   0   0
5   2   0   0   0   4  16   3   4  11   2   0   0   0  14   3   0   4  11   0   0   0  10   8   4  11  12   0
  V57 V58 V59 V60 V61 V62 V63 V64
1   0   0   6  14   7   1   0   0
2   0   0  10  16  15   3   0   0
3   0   0   9  14   0   0   0   0
4   0   0   0   1  15   2   0   0
5   0   0   4  12  14   7   0   0
```

- Number of attributes in the dataset,

```
Console D:/pruthvi/Spring-17/ML/ITCS6156_SLProject/DigitRecognition/
> ncol(train_data)
[1] 64
>
```

- Checking for the number of total tuples in the data set,

```
Console D:/pruthvi/Spring-17/ML/ITCS6156_SLProject/DigitRecognition/
> nrow(train_data)
[1] 3823
>
```

- Checking the structure of data,

```
Console D:/pruthvi/Spring-17/ML/ITCS6156_SLProject/DigitRecognition/
> str(train_data)
'data.frame':   3823 obs. of  64 variables:
 $ V1 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ V2 : int  1 0 0 0 0 0 0 0 0 0 ...
 $ V3 : int  6 10 8 0 5 11 1 8 15 3 ...
 $ V4 : int  15 16 15 3 14 16 11 10 2 13 ...
 $ V5 : int  12 6 16 11 4 10 13 8 14 13 ...
 $ V6 : int  1 0 13 16 0 1 11 7 13 2 ...
 $ V7 : int  0 0 0 0 0 0 7 2 2 0 ...
 $ V8 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ V9 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ V10: int  7 7 1 0 0 4 0 1 0 6 ...
 $ V11: int  16 16 11 5 13 16 9 15 16 16 ...
 $ V12: int  6 8 9 16 8 10 14 14 15 12 ...
 $ V13: int  6 16 11 11 0 15 6 12 12 10 ...
 $ V14: int  10 5 16 13 0 8 4 12 13 8 ...
 $ V15: int  0 0 1 7 0 0 3 4 8 0 ...
 $ V16: int  0 0 0 0 0 0 0 0 0 0 ...
 $ V17: int  0 0 0 0 0 0 0 0 0 0 ...
 $ V18: int  8 11 0 3 3 4 0 7 2 9 ...
 $ V19: int  16 16 0 15 14 16 16 15 16 15 ...
 $ V20: int  2 0 0 8 4 3 12 12 12 12 ...
 $ V21: int  0 6 7 1 0 11 16 5 1 16 ...
```

- Did not consider calculating mean and standard deviation of each attribute because each column is a pixel values and calculating a statistical measure of them is not relevant in establishing an observation.

- RESULTS:

Calculating the accuracy of the predicted values to the existing values trough the output of cost matrix. Taking in the diagonal values form the cost matrix and summing up the entire cost matrix to calculate the difference between them to give the accurate values predicted and converting the result into percentage.

```
Console D:/pruthvi/Spring-17/ML/ITCS6156_SLProject/DigitRecognition/
> diag<-diag(cm)
> n<-sum(cm)
> acc<-sum(diag)/n
> print(acc*100)
[1] 83.91764
>
```

The complexity parameter (cp) is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue. Least the CP more the tree can accommodate values and increases the accuracy of the model. For instance, I've used cp=0.00001 and the accuracy of the predictions is at its maximum or else it was approximately 75% which is behind the present accuracy of around 83%.

R programming language provides two libraries rpart and ctree(partykit) to implement decision tree. Both provide a pruned result. However, I've used the prune method on the model created to just check if it betters the fit by any means but it resulted in the same predictions.

```
# pruning the above model to check the accuracy.
pfit<- prune(fit, cp= fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])
prune_predict<-predict(pfit,test_data,type="class")
cm_prune<-table(prune_predict,class_label_test)
```

## Amazon review dataset

Data Exploration:

- Structure of the dataset in rows and columns:

```
C:\Users\pruth\Anaconda3\lib\site-pack
  "This module will be removed in 0.20
(145927, 3)

Process finished with exit code 0
```

- Removing null spaces and checking the rows and columns,

```
C:\Users\pruth\Anaconda3\lib\site-packages\sklearn\cross_validation.py
  "This module will be removed in 0.20.", DeprecationWarning)
(146824, 3)

Process finished with exit code 0
```

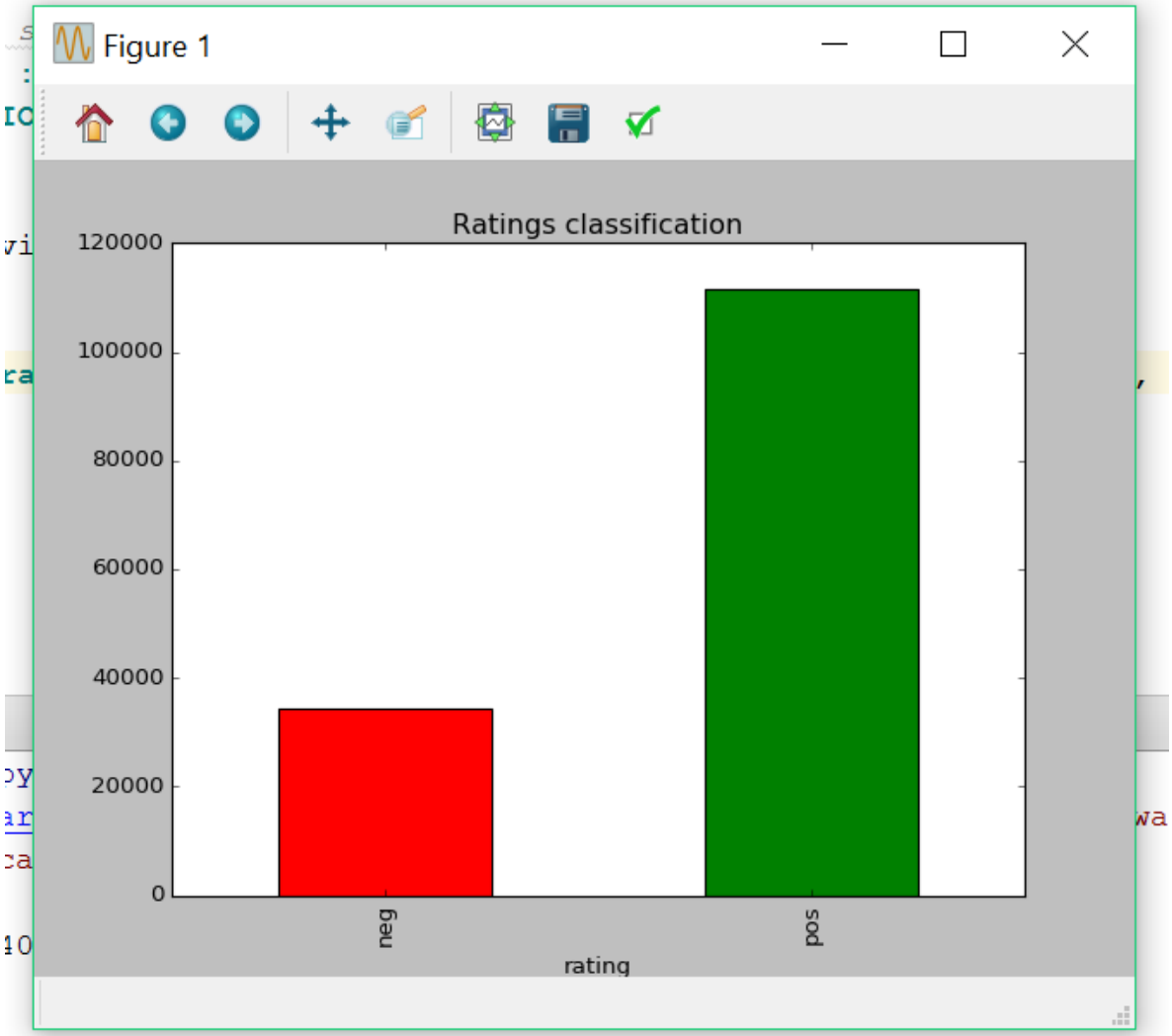
- Calculating mean and standard deviation on the ratings column,

```
C:\Users\pruth\Anaconda3\lib\site-packages\sklearn\cross_valid
  "This module will be removed in 0.20.", DeprecationWarning)
MEAN OF RATINGS :  4.120430078052725
STANDARD DEVIATION OF RATINGS :  1.2853703237434095

Process finished with exit code 0
```

- Dividing the dataset into negative and positive reviews based on rating values. I considered 1,2,3 ratings as negative and 4,5 as positive.

```
C:\Users\pruth\Anaconda3\lib\site-packages\sklearn\cross_valic
  "This module will be removed in 0.20.", DeprecationWarning)
MEAN OF RATINGS :  4.120430078052725
STANDARD DEVIATION OF RATINGS :  1.2853703237434095
rating
neg     34398
pos    111529
Name: review, dtype: int64

Process finished with exit code 0
```

- Plotting positive and negative reviews.




- 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 and training the model on a 25% split data from the given data set as a cross validation.

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

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

```
[('the', 152974), ('it', 99372), ('i',
Accuracy of test :  70.4418904463 %

Process finished with exit code 0
```