DATA MINING

CS6405

Anish Viswanathan **119220053**

"I have read and understand the UCC academic policy on plagiarism and I agree to the requirements set out thereby in relation to plagiarism and referencing. I confirm that I have referenced and acknowledged properly all sources used in preparation of this assignment. I declare that this assignment is entirely my own work based on my personal study. I further declare that I have not engaged the services of another to either assist me in, or complete this assignment"

E-Signature

Anish Viswanathan

Exploration of the dataset

In this part of the project I created a corpus and constructed bag of words. If you are facing memory issues, there is an additional impute parameter "word_sample" which takes an integer value and reduces the corpus. (I have taken word_sample=4)

Note: If you are not facing memory issues then input word_sample=0

Basic Evaluation

Split the dataset for training and testing, where training is 70% of data and testing is 30% of the data.

Naïve Bayes:

Self Implementation

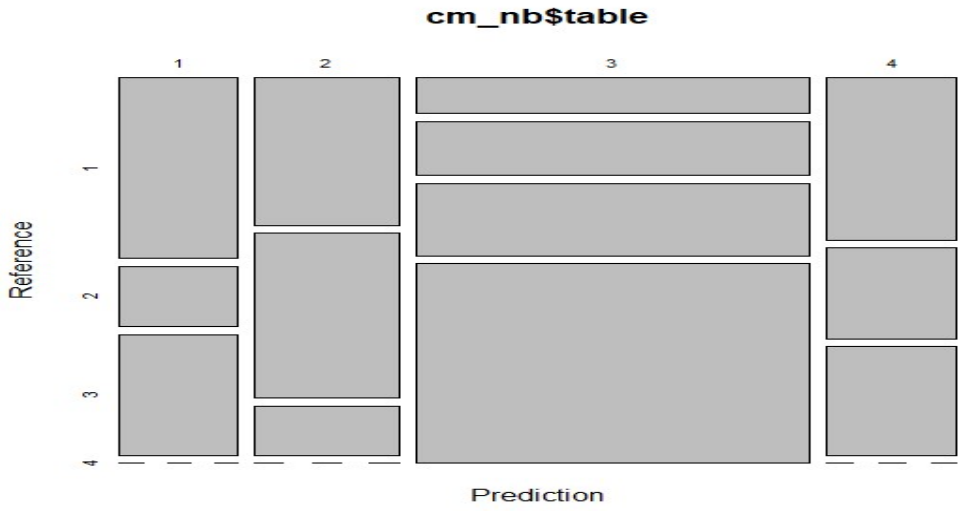
```
Overall Statistics

          Reference
Prediction  1  2  3  4
         1  9  3  6  0
         2  9 10  3  0
         3  6  9 12 33
         4  9  5  6  0

               Accuracy : 0.2583
                 95% CI : (0.1828, 0.3462)
    No Information Rate : 0.275
    P-Value [Acc > NIR] : 0.6912

                  Kappa : 0.0231

 Mcnemar's Test P-Value : 8.223e-07
```




```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.2727  0.37037   0.4444   0.0000
Specificity            0.8966  0.87097   0.4839   0.7701
Pos Pred Value         0.5000  0.45455   0.2000   0.0000
Neg Pred Value         0.7647  0.82653   0.7500   0.6700
Prevalence             0.2750  0.22500   0.2250   0.2750
Detection Rate         0.0750  0.08333   0.1000   0.0000
Detection Prevalence   0.1500  0.18333   0.5000   0.1667
Balanced Accuracy      0.5846  0.62067   0.4642   0.3851
```

Testing accuracy :-27.5%

Library Implementation

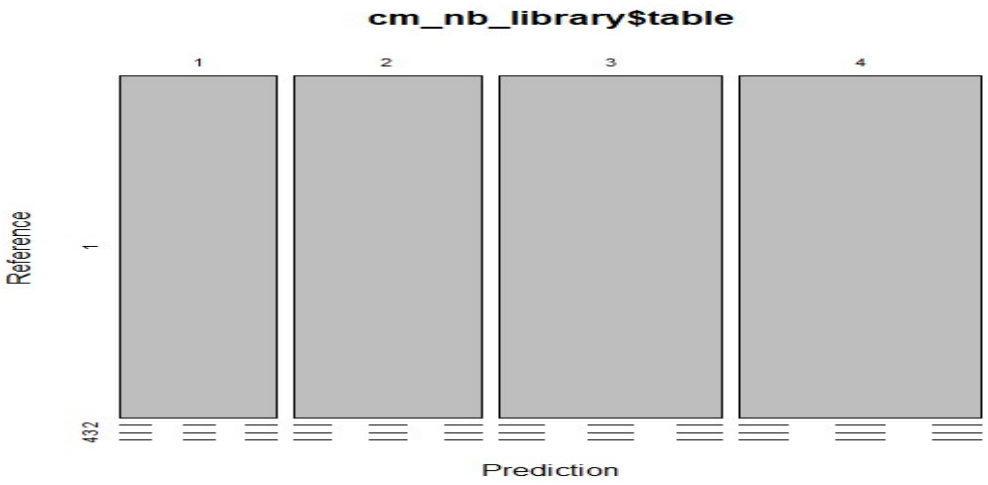
```
Overall Statistics

          Reference
Prediction  1  2  3  4
         1 23  0  0  0
         2 28  0  0  0
         3 33  0  0  0
         4 36  0  0  0

               Accuracy : 0.1917
                 95% CI : (0.1256, 0.2736)
    No Information Rate : 1
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : NA
```




```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.1917       NA       NA       NA
Specificity                NA   0.7667    0.725      0.7
Pos Pred Value             NA       NA       NA       NA
Neg Pred Value             NA       NA       NA       NA
Prevalence             1.0000   0.0000    0.000      0.0
Detection Rate         0.1917   0.0000    0.000      0.0
Detection Prevalence   0.1917   0.2333    0.275      0.3
Balanced Accuracy          NA       NA       NA       NA
```

K-Nearest Neighbours:

Confusion matrix, precision and recall:

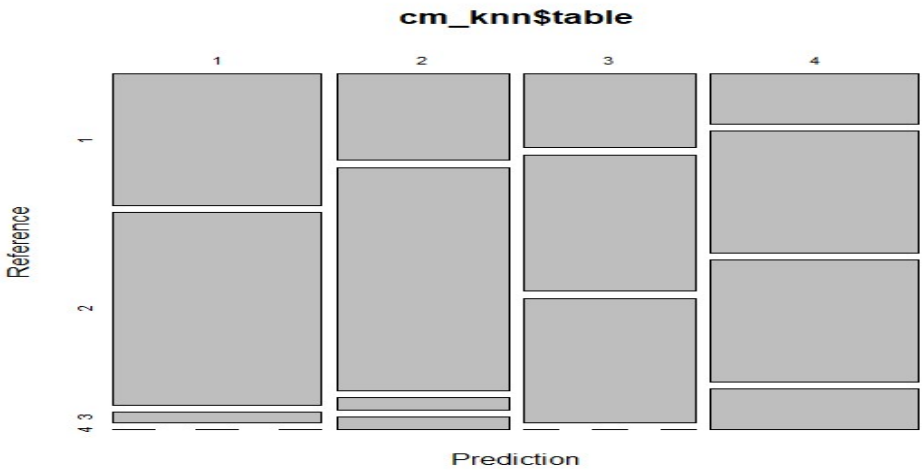
```
          Reference
Prediction  1  2  3  4
         1 13 19  1  0
         2  7 18  1  1
         3  6 11 10  0
         4  5 12 12  4
```

```
Overall Statistics

               Accuracy : 0.375
                 95% CI : (0.2883, 0.468)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.9978

                  Kappa : 0.1776

 Mcnemar's Test P-Value : 8.282e-08
```



Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.4194	0.3000	0.41667	0.80000
Specificity	0.7753	0.8500	0.82292	0.74783
Pos Pred Value	0.3939	0.6667	0.37037	0.12121
Neg Pred Value	0.7931	0.5484	0.84946	0.98851
Prevalence	0.2583	0.5000	0.20000	0.04167
Detection Rate	0.1083	0.1500	0.08333	0.03333
Detection Prevalence	0.2750	0.2250	0.22500	0.27500
Balanced Accuracy	0.5973	0.5750	0.61979	0.77391

Testing accuracy :-37.5%

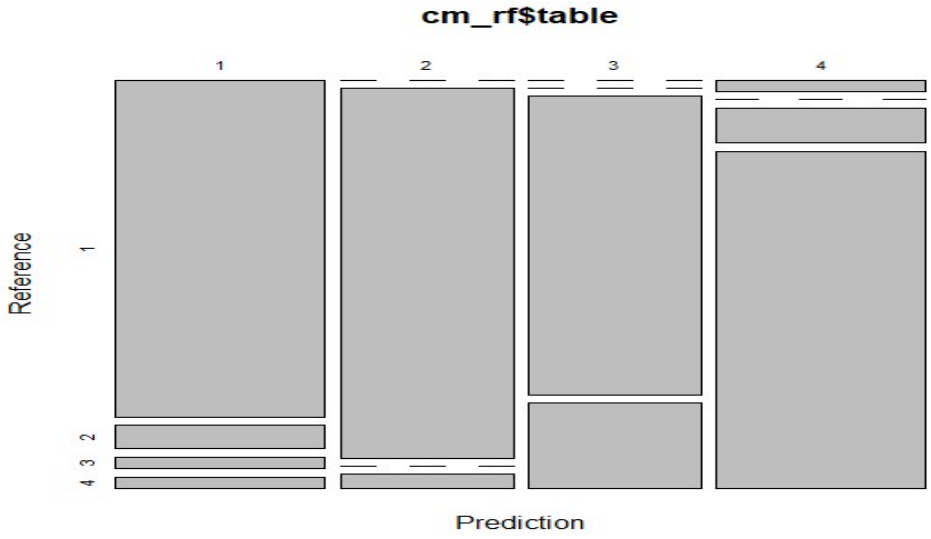
Random Forest:

Confusion matrix, precision and recall:

Prediction \ Reference	1	2	3	4
1	29	2	1	1
2	0	26	0	1
3	0	0	21	6
4	1	0	3	29

Overall Statistics

Accuracy	: 0.875
95% CI	: (0.8022, 0.9283)
No Information Rate	: 0.3083
P-Value [Acc > NIR]	: < 2.2e-16
Kappa	: 0.8327
Mcnemar's Test P-Value	: NA



Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.9667	0.9286	0.8400	0.7838
Specificity	0.9556	0.9891	0.9368	0.9518
Pos Pred Value	0.8788	0.9630	0.7778	0.8788
Neg Pred Value	0.9885	0.9785	0.9570	0.9080
Prevalence	0.2500	0.2333	0.2083	0.3083
Detection Rate	0.2417	0.2167	0.1750	0.2417
Detection Prevalence	0.2750	0.2250	0.2250	0.2750
Balanced Accuracy	0.9611	0.9589	0.8884	0.8678

Testing accuracy :- 87.5

Robust Evaluation

- Pre-processing techniques:
  To clean the data set the following methods were applied:
  1. Stop words removal
  2. Punctuation removal
  3. Converted all text to lower case
  4. White space removal

- Feature Selection:
  Applied feature selection and extracted 10% of the most important features.

- Cross validation:
  It is applied on all models separately.

- Performance Matrix:
  It is applied on all models separately.

K-Nearest Neighbours:

Cross Validation and Hyper parameter tuning:

```
> res_knnr$x
$k
[1] 7
```

```
> res_knnr$y
acc.test.mean
     0.457487
```

Confusion matrix, precision and recall:

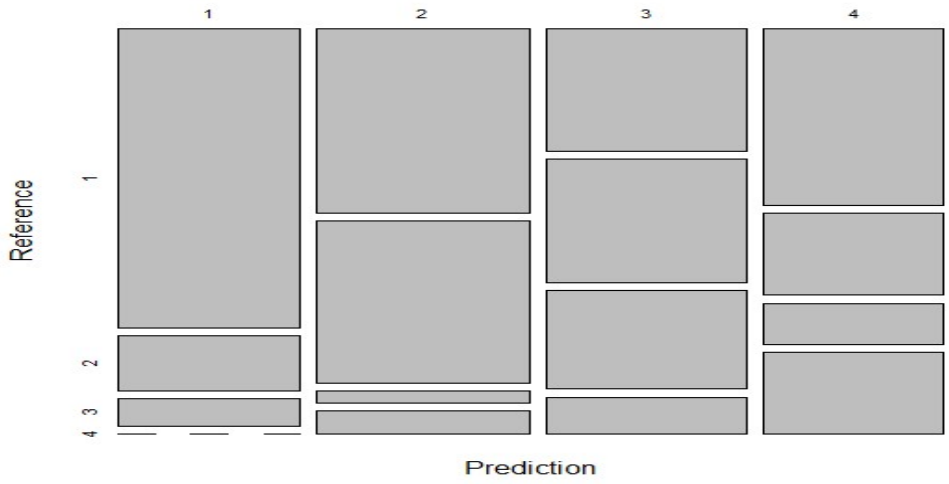
```
          Reference
Prediction  1  2  3  4
         1 22  4  2  0
         2 16 14  1  2
         3 10 10  8  3
         4 13  6  3  6
```

```
Overall Statistics

               Accuracy : 0.4167
                 95% CI : (0.3274, 0.5102)
    No Information Rate : 0.5083
    P-Value [Acc > NIR] : 0.9823

                  Kappa : 0.2242

 Mcnemar's Test P-Value : 4.512e-06
```




```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.3607   0.4118  0.57143  0.54545
Specificity            0.8983   0.7791  0.78302  0.79817
Pos Pred Value         0.7857   0.4242  0.25806  0.21429
Neg Pred Value         0.5761   0.7701  0.93258  0.94565
Prevalence             0.5083   0.2833  0.11667  0.09167
Detection Rate         0.1833   0.1167  0.06667  0.05000
Detection Prevalence   0.2333   0.2750  0.25833  0.23333
Balanced Accuracy      0.6295   0.5954  0.67722  0.67181
```

Testing accuracy :- 34.2%

Random Forest:

Cross Validation and Hyper parameter tuning:

```
> res_rfr$x
$ntree
[1] 180
```

```
> res_rfr$y
acc.test.mean
    0.8526353
```

Confusion matrix, precision and recall:

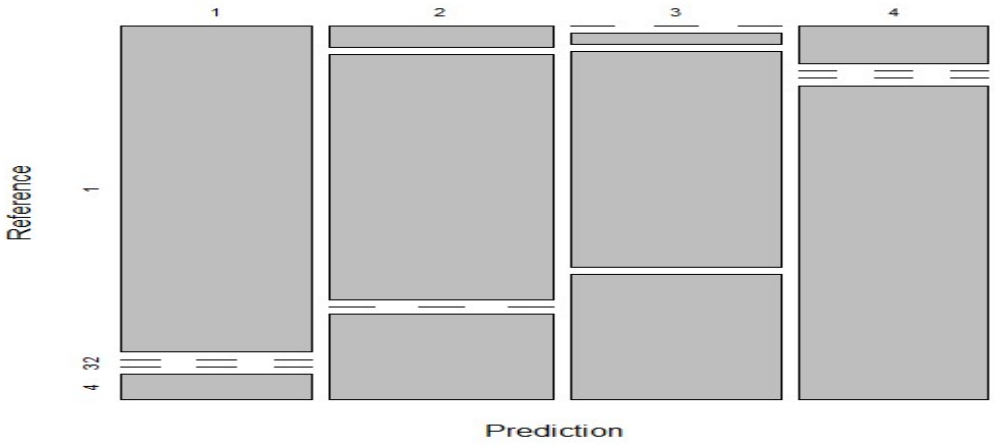
```
          Reference
Prediction  1  2  3  4
         1 26  0  0  2
         2  2 23  0  8
         3  0  1 19 11
         4  3  0  0 25
```

```
Overall Statistics

               Accuracy : 0.775
                 95% CI : (0.6898, 0.8462)
    No Information Rate : 0.3833
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7017

 Mcnemar's Test P-Value : NA
```




```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8387   0.9583   1.0000   0.5435
Specificity            0.9775   0.8958   0.8812   0.9595
Pos Pred Value         0.9286   0.6970   0.6129   0.8929
Neg Pred Value         0.9457   0.9885   1.0000   0.7717
Prevalence             0.2583   0.2000   0.1583   0.3833
Detection Rate         0.2167   0.1917   0.1583   0.2083
Detection Prevalence   0.2333   0.2750   0.2583   0.2333
Balanced Accuracy      0.9081   0.9271   0.9406   0.7515
```

Testing accuracy :- 88.3%

Decision Tree:

Cross Validation and Hyper parameter tuning:

```
> print(res_dt$x)
$minsplit
[1] 12

$maxdepth
[1] 7

$cp
[1] 0.001
```

```
> print(res_dt$y)
acc.test.mean
    0.8451913
```

Confusion matrix, precision and recall:

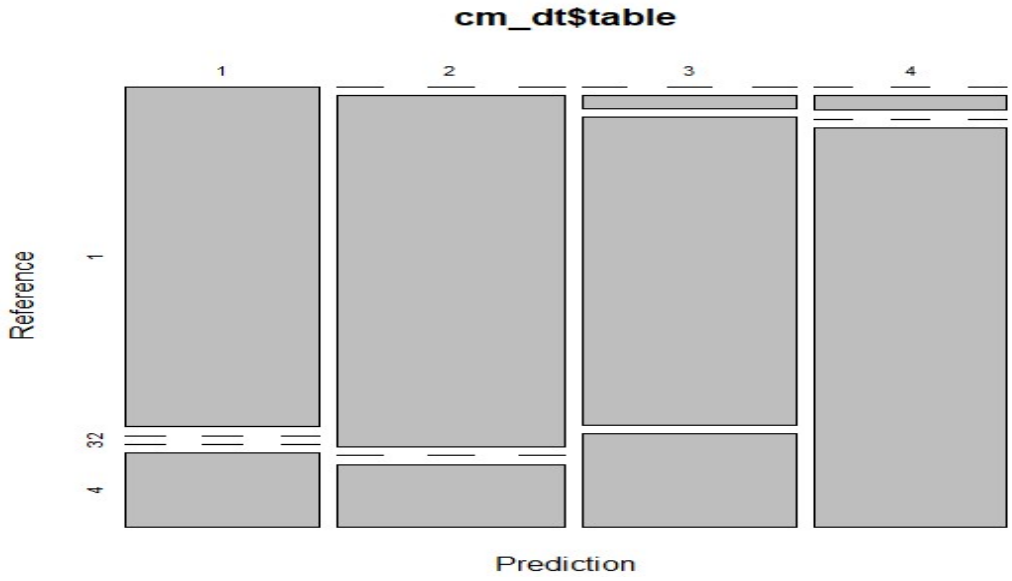
```
          Reference
Prediction  1  2  3  4
         1 23  0  0  5
         2  0 28  0  5
         3  0  1 23  7
         4  0  1  0 27
```

```
Overall Statistics

               Accuracy : 0.8417
                 95% CI : (0.7638, 0.9019)
    No Information Rate : 0.3667
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7893

 Mcnemar's Test P-Value : NA
```




```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            1.0000   0.9333   1.0000   0.6136
Specificity            0.9485   0.9444   0.9175   0.9868
Pos Pred Value         0.8214   0.8485   0.7419   0.9643
Neg Pred Value         1.0000   0.9770   1.0000   0.8152
Prevalence             0.1917   0.2500   0.1917   0.3667
Detection Rate         0.1917   0.2333   0.1917   0.2250
Detection Prevalence   0.2333   0.2750   0.2583   0.2333
Balanced Accuracy      0.9742   0.9389   0.9588   0.8002
```

Testing accuracy :- 88.3%

Support Vector Machines:

Cross Validation and Hyper parameter tuning:

```
> res_dt$x
$minsplit
[1] 12

$maxdepth
[1] 7

$cp
[1] 0.001
```

```
> res_dt$y
acc.test.mean
    0.8451913
```

Confusion matrix, precision and recall

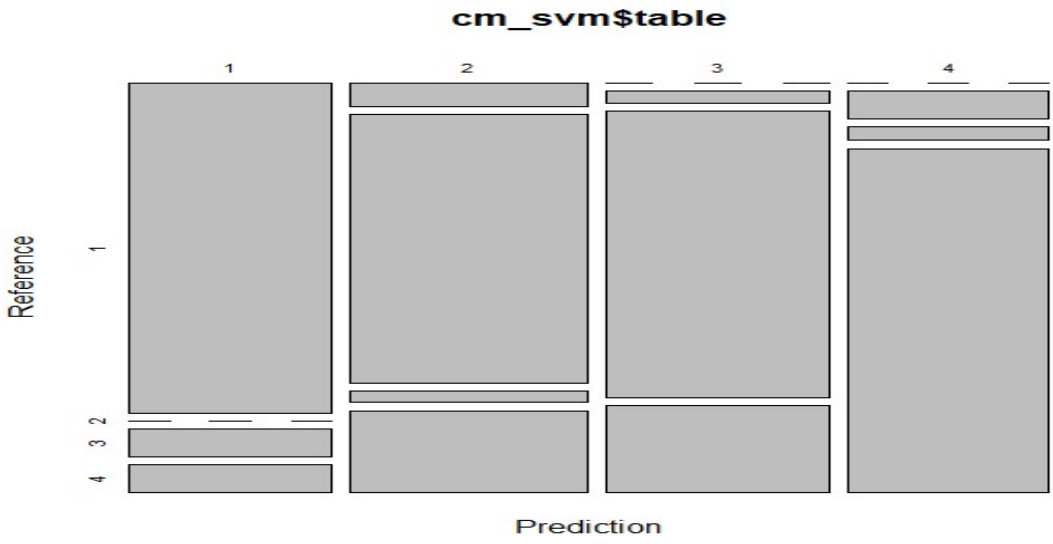
```
          Reference
Prediction  1  2  3  4
         1 24  0  2  2
         2  2 23  1  7
         3  0  1 23  7
         4  0  2  1 25
```

```
Overall Statistics

               Accuracy : 0.7917
                 95% CI : (0.708, 0.8604)
    No Information Rate : 0.3417
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.723

 Mcnemar's Test P-Value : 0.03883
```




```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.9231   0.8846   0.8519   0.6098
Specificity            0.9574   0.8936   0.9140   0.9620
Pos Pred Value         0.8571   0.6970   0.7419   0.8929
Neg Pred Value         0.9783   0.9655   0.9551   0.8261
Prevalence             0.2167   0.2167   0.2250   0.3417
Detection Rate         0.2000   0.1917   0.1917   0.2083
Detection Prevalence   0.2333   0.2750   0.2583   0.2333
Balanced Accuracy      0.9403   0.8891   0.8829   0.7859
```

Testing accuracy :- 81.7%

Naïve Bayes:

Self Implementation

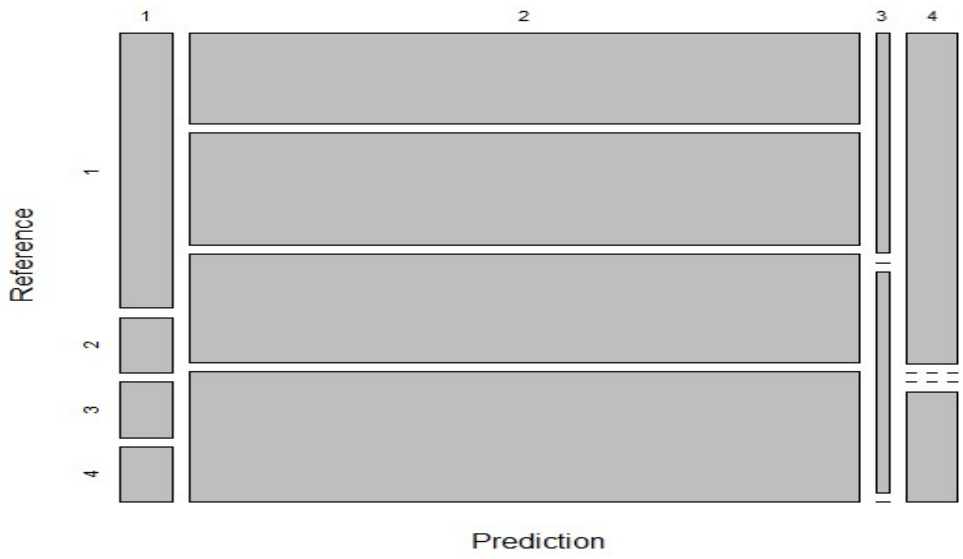
```
          Reference
Prediction  1  2  3  4
         1  5  1  1  1
         2 21 26 25 30
         3  1  0  1  0
         4  6  0  0  2
```

```
Overall Statistics

               Accuracy : 0.2833
                 95% CI : (0.2049, 0.3728)
    No Information Rate : 0.275
    P-Value [Acc > NIR] : 0.4533

                  Kappa : 0.0672

 Mcnemar's Test P-Value : NA
```




```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.15152   0.9630 0.037037  0.06061
Specificity           0.96552   0.1828 0.989247  0.93103
Pos Pred Value        0.62500   0.2549 0.500000  0.25000
Neg Pred Value        0.75000   0.9444 0.779661  0.72321
Prevalence            0.27500   0.2250 0.225000  0.27500
Detection Rate        0.04167   0.2167 0.008333  0.01667
Detection Prevalence  0.06667   0.8500 0.016667  0.06667
Balanced Accuracy     0.55852   0.5729 0.513142  0.49582
```

Testing accuracy :-28.5%

Considering all the testing accuracy it can be said that Random Forest outperforms others and is the best model.

Naïve Bayes performs well for text classification.

There is sparsity in our data which is leading to low performance of Naïve Bayes with large number of features

Reducing the features can give us better result.