Project Report

Personalized Medicine: Redefining Cancer Treatment

1. Problem Statement:

Cancer Tumor can have thousands of genetic mutation but the challenge is distinguishing the mutation that contribute to tumor growth from natural mutators by developing a model to classify genetic mutations based on clinical evidence (text). There are nine different classes a genetic mutation can be classified on.

2. Data Files provided:-

Training_Variants.csv : A comma separated file containing the description of the genetic mutations used for training.

Fields :-

- ID -The id of the row used to link the clinical evidence to the genetic mutation
- Gene -the gene where this genetic mutation is located
- Variation -the aminoacid change for this mutations
- Class - 1-9 the class this genetic mutation has been classified on

Training_text: A double pipe (||) delimited file that contains the clinical evidence (text) used to classify genetic mutations

Fields :-

- ID -The id of the row used to link the clinical evidence to the genetic mutation
- Text -the clinical evidence used to classify the genetic mutation

Test_variants - A comma separated file containing the description of the genetic mutations used for training

Fields :-

- ID -The id of the row used to link the clinical evidence to the genetic mutation
- Gene - The gene where this genetic mutation is located
- Variation - The aminoacid change for this mutations
- Class - 1-9 the class this genetic mutation has been classified on

Test_text: A double pipe (||) delimited file that contains the clinical evidence (text) used to classify genetic mutations

Fields :-

- ID -The id of the row used to link the clinical evidence to the genetic mutation
- Text -the clinical evidence used to classify the genetic mutation

3.Tools and Library's used

1. R 3.4.1

2. Rstudio-1.0.153

3. Library's used:

- utils
- magrittr
- tm
- wordcloud
- ggplot2
- scales
- caTools
- e1071
- caret

4.Exploratory Data analysis

After successfully loading the file into the environment

1. structure of all file were analyzed
2. top rows of data were analyzed to get idea about variables
3. Bar Plot of class variable was made for checking the distribution of class over the data set

5.Features Selection / Engineering

1. Gene , Variation variables from the Variants file were converted into Numeric data structure
2. ID and Class variables were converted into Factor data structure
3. New variable class (Target variable) was added into the test data with NA values

6.Text Mining

1. The training_text and test_text files were combined for building the text corpus

2. Pre-processing steps performed:

    1. Removing Punctuation Marks
    2. Removing Numbers
    3. Case folding
    4. Removing Stop words
    5. Removing White spaces
    6. Stemming

3. Wordcloud was made over corpus to observe the importance of words

4. Document term matrix of the text corpus was made on TF-IDF as weighting criteria

5. The generated Document term matrix was spliited into train and test set

6. The train Documentterm matrix and test Document term matrix was cbinded with their Variants file respectfully.

7.Model buliding and validation

1. Sample was taken from the train data set for validation.

2.Predictive model was build using Naive bayes technique upon the sample train data(70% train and 30%test) and prediction was done on sample test data taking Class as target variable and all other variable as predictors.

3. Model took 37.476 seconds to train.

4 Confusion matrix was Calculated for validation of predicted values against original values of class

Overall Statistics Of Confusion matrix

Accuracy : 0.5521
95% CI : (0.5206, 0.5833)
No Information Rate : 0.3287

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4417
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6
Sensitivity	0.48990	0.38650	0.000000	0.6393	0.45455	0.63158
Specificity	0.90875	0.91257	0.972919	0.8908	0.94444	0.96204
Pos Pred Value	0.57059	0.46324	0.000000	0.5680	0.27397	0.57831
Neg Pred Value	0.87802	0.88399	0.998970	0.9167	0.97405	0.96940
Prevalence	0.19840	0.16333	0.001002	0.1834	0.04409	0.07615
Detection Rate	0.09719	0.06313	0.000000	0.1172	0.02004	0.04810
Detection Prevalence	0.17034	0.13627	0.027054	0.2064	0.07315	0.08317
Balanced Accuracy	0.69932	0.64954	0.486459	0.7651	0.69949	0.79681

	Class: 7	Class: 8	Class: 9
Sensitivity	0.6159	1.000000	0.750000
Specificity	0.8746	0.994985	0.991952
Pos Pred Value	0.7063	0.166667	0.272727
Neg Pred Value	0.8230	1.000000	0.998987
Prevalence	0.3287	0.001002	0.004008
Detection Rate	0.2024	0.001002	0.003006
Detection Prevalence	0.2866	0.006012	0.011022
Balanced Accuracy	0.7452	0.997492	0.870976

8.Final prediction on test set and output file generation

1. Class values for the orignal test data were predicted using the Predictive model.

2. Output of predicted values of all of Nine clases was Column binded with ID variable from the test dataset.

3. Final Output File was saved in workspace in csv format.