ÇANKAYA UNIVERSITY

Faculty of Engineering

Computer Engineering

CENG 464 Term Project

Instructor: Gül TOKDEMİR

prepared by:

Ayşe Simge ALMAŞ 201432005
Yaren İPEK 201711038

1.INTRODUCTION

In this project, we are analyzing the dataset1. At the end of the project, we are expected to be analyzed the dataset and its attributes, and make predictions in light of these analyses.

2.METHODS

2.1 Libraries

Readxl: They are used to get data out of Excel into R. Since our dataset is in xlsx format, we imported this packages.

Caret: The caret package can be used for data splitting, pre-processing, feature selection, etc. These functionalities are necessary for our project.

Factoextra: This package is used to extract and visualize the results of data analyses.

Ggplot: The ggplot package is useful for creating graphics. Data visualization is an important step to see our results clearly.

Dplyr: The dplyr package helps solving the most common data manipulation challenges.

Cluster: This package includes methods for cluster analysis.

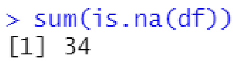
Fpc: This package includes various methods for clustering and cluster validation.

NbClust: The NbClust package helps proposing the best clustering scheme from the different results.

Class: The class package includes functions for classification.

2.2 Preprocessing the Data

In our data set, there were 82 attributes with 999 observations and there were some missing values. So we started with removing these by the mean value. After this process, some of the features have eliminated.

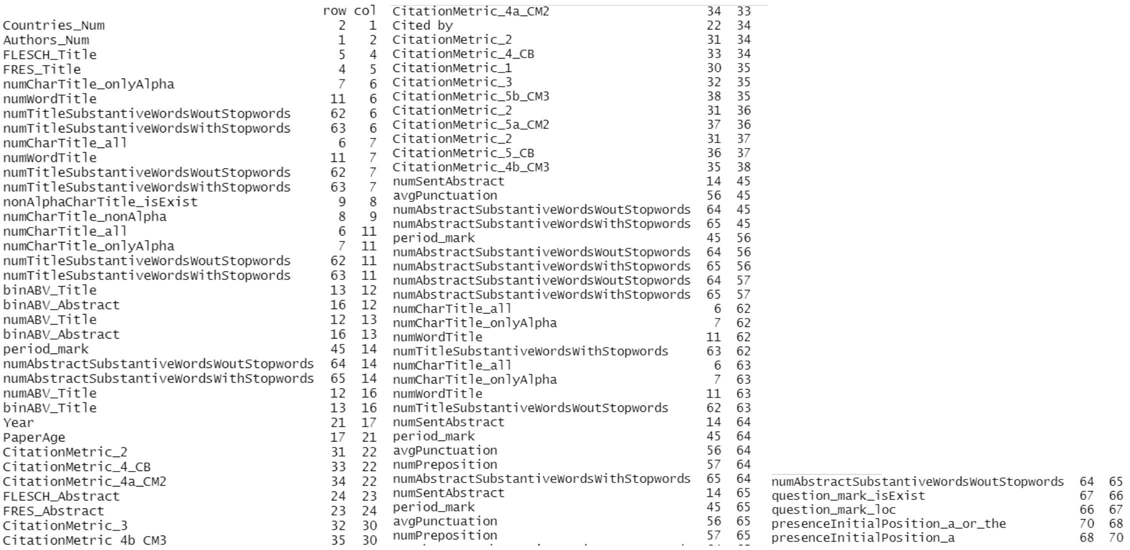


```
> sum(is.na(df))
[1] 34
```

Figure 1-*Missing Values*

When the data set was examined, it was observed that 34 data were missing. Missing values were filled in with mean values.We prepared a correlation matrix and selected features according to this matrix. Categorical variables have been converted to numeric variables.

While the correlation process was performed in the data set, it was determined that some features had the same values.These features were deleted from the data set.The correlation process has been done. Below are the similarity features of over 75 percent.



```
                                            row col
Countries_Num                                 2   1
Authors_Num                                   1   2
FLESCH_Title                                  5   4
FRES_Title                                    4   5
numCharTitle_onlyAlpha                        7   6
numWordTitle                                 11   6
numTitleSubstantiveWordsWoutStopwords        62   6
numTitleSubstantiveWordsWithStopwords        63   6
numCharTitle_all                              6   7
numWordTitle                                 11   7
numTitleSubstantiveWordsWoutStopwords        62   7
numTitleSubstantiveWordsWithStopwords        63   7
nonAlphaCharTitle_isExist                     9   8
numCharTitle_nonAlpha                         8   9
numCharTitle_all                              6  11
numCharTitle_onlyAlpha                        7  11
numTitleSubstantiveWordsWoutStopwords        62  11
numTitleSubstantiveWordsWithStopwords        63  11
binABV_Title                                 13  12
binABV_Abstract                              16  12
numABV_Title                                 12  13
binABV_Abstract                              16  13
period_mark                                  45  14
numAbstractSubstantiveWordsWoutStopwords     64  14
numAbstractSubstantiveWordsWithStopwords     65  14
numABV_Title                                 12  16
binABV_Title                                 13  16
Year                                         21  17
PaperAge                                     17  21
CitationMetric_2                             31  22
CitationMetric_4_CB                          33  22
CitationMetric_4a_CM2                        34  22
FLESCH_Abstract                              24  23
FRES_Abstract                                23  24
CitationMetric_3                             32  30
CitationMetric 4b CM3                        35  30
```

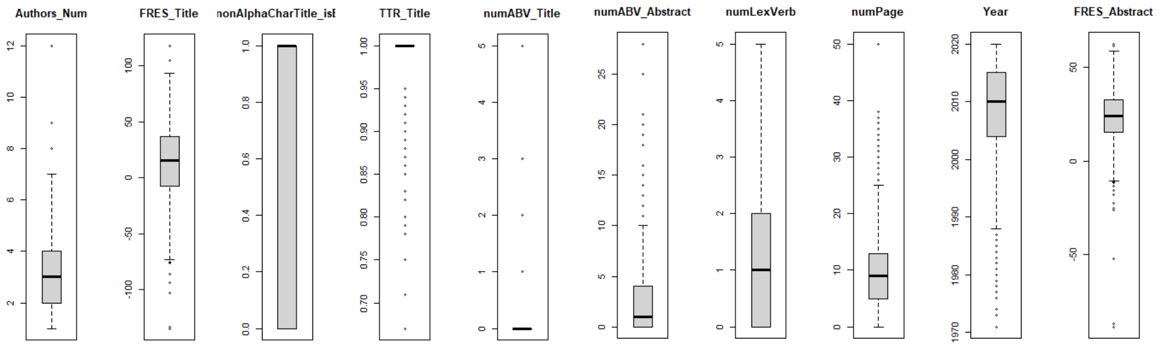
```
CitationMetric_4a_CM2                       34  33
Cited by                                    22  34
CitationMetric_2                            31  34
CitationMetric_4_CB                         33  34
CitationMetric_1                            30  35
CitationMetric_3                            32  35
CitationMetric_5b_CM3                       38  35
CitationMetric_2                            31  36
CitationMetric_5a_CM2                       37  36
CitationMetric_2                            31  37
CitationMetric_5_CB                         36  37
CitationMetric_4b_CM3                       35  38
numSentAbstract                             14  45
avgPunctuation                              56  45
numAbstractSubstantiveWordsWoutStopwords    64  45
numAbstractSubstantiveWordsWithStopwords    65  45
period_mark                                 45  56
numAbstractSubstantiveWordsWoutStopwords    64  56
numAbstractSubstantiveWordsWithStopwords    65  56
numAbstractSubstantiveWordsWoutStopwords    64  57
numAbstractSubstantiveWordsWithStopwords    65  57
numCharTitle_all                             6  62
numCharTitle_onlyAlpha                       7  62
numWordTitle                                11  62
numTitleSubstantiveWordsWithStopwords       63  62
numCharTitle_all                             6  63
numCharTitle_onlyAlpha                       7  63
numWordTitle                                11  63
numTitleSubstantiveWordsWoutStopwords       62  63
numSentAbstract                             14  64
period_mark                                 45  64
avgPunctuation                              56  64
numPreposition                              57  64
numAbstractSubstantiveWordsWithStopwords    65  64
numSentAbstract                             14  65
period_mark                                 45  65
avgPunctuation                              56  65
numPreposition                              57  65
```

```
numAbstractSubstantiveWordsWoutStopwords    64  65
question_mark_isExist                       67  66
question_mark_loc                           66  67
presenceInitialPosition_a_or_the            70  68
presenceInitialPosition_a                   68  70
```

Figure 2- *>0.75% Correlation*

We cleared the attributes above 75 percent.The features that were thought to affect the prediction results negatively were cleared.There were 18 attributes left.

Detected outliers in the boxplots:



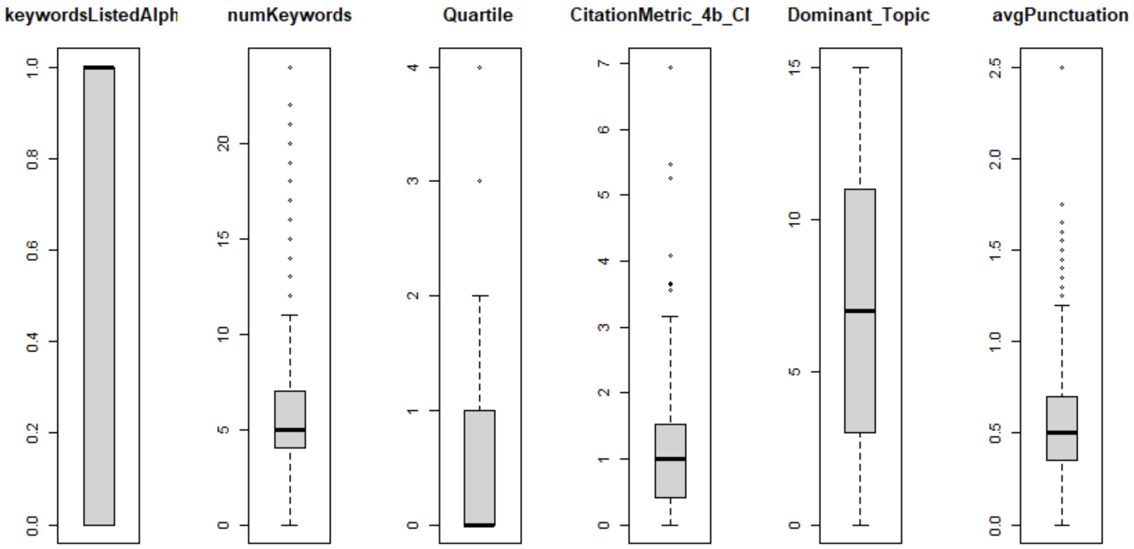


Figure 3*- Box Plot before Preprocessing*

An average value was assigned to some outliers entity. Some entities were removed from the dataset.(4 values) . We also normalized attributes. There were 964 observations and 15 attributes left. The boxplots of the attributes after these operations are below:

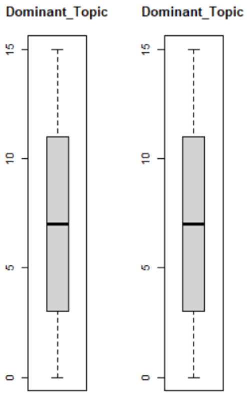


Figure 4*- Pre-processed sample properties*

When the data set is examined, the distribution range of the values in the data set is very wide.Therefore, the normalization process has been done.

2.3 Clustering Methods

2.3.1 K-Means (for Quartile)

Elbow method was used to get high accuracy in the K-means algorithm. It was decided that the value of 5 was optimal.

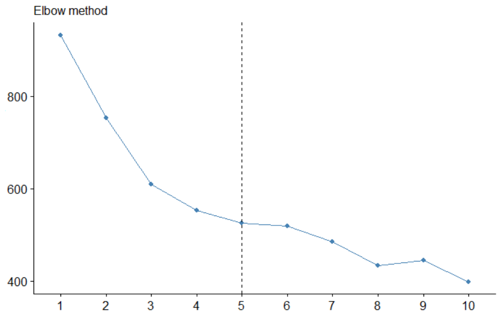


Figure 5*-Elbow Method*

We used k-means method to apply clustering to our data set. We used kmeans function, chose the k value as 5 and the nstart as 20. Accuracy was 48.1%. Here is the cluster plot of the k-means.

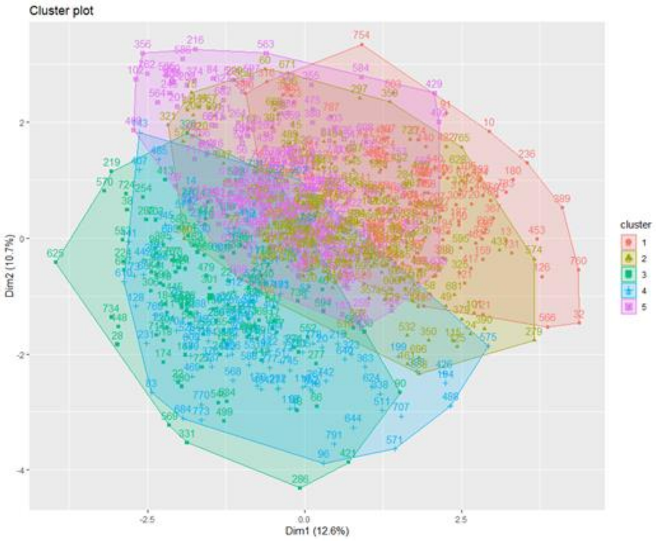


Figure 6*-K-Means Clustering*

2.3.2 Hierarchical Clustering (for Quartile)

We applied hierarchical clustering to our data set by using the ward's method. Hclust and dist functions have used. The dendrogram is below:

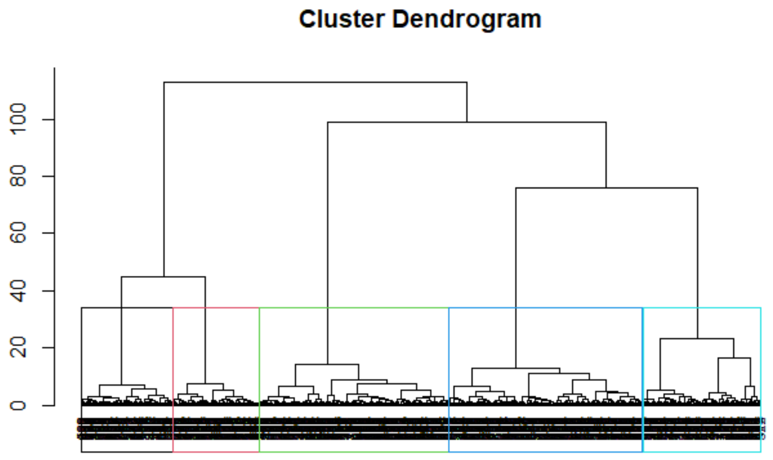


Figure 7 *-Hierarchical Clustering*

2.3.3 K-Means (for J_or_C)

We used k-means method to apply clustering to our data set. We used kmeans function, chose the k value as 5 and the nstart as 20. Accuracy was 20.4%. Here is the cluster plot of the k-means.

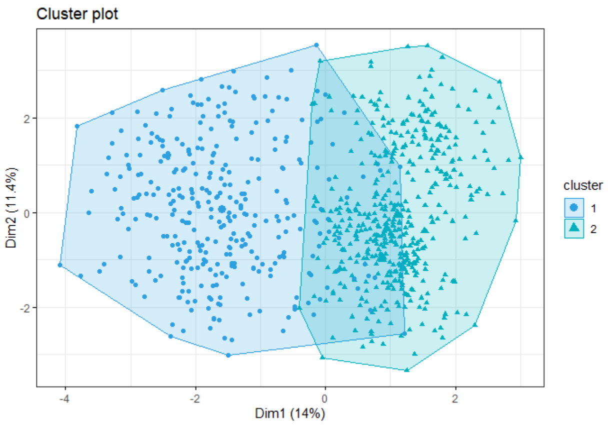


Figure 8*-K-Means Clustering Plot*

2.3.4 Hierarchical Clustering (for J_or_C)

We applied hierarchical clustering to our data set by using the ward's method. Hclust and dist functions have used. The dendrogram is below:

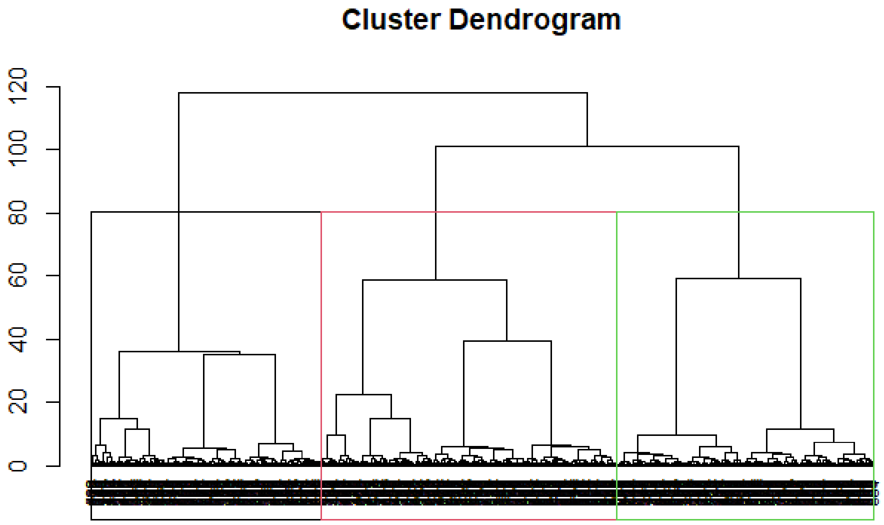


Figure 9-Hierarchical Clustering

2.4 Classification Methods

2.4.1 K-Nearest Neighbour (for Quartile)

We applied k-nearest neighbour algorithm to classify our data. We used knn function for this method. We randomly splitted our data set into train and test sets. The train set contains 75% and the test set contains 25% of the data. We calculated the accuracy and got 83.33% accuracy.

2.4.2 Decision Tree (for Quartile)

We applied decision tree algorithm to classify our data. We used decision tree function for this method. We randomly splitted our data set into train and test sets. The train set contains 75% and the test set contains 25% of the data. We calculated the accuracy and got 76.76% accuracy.

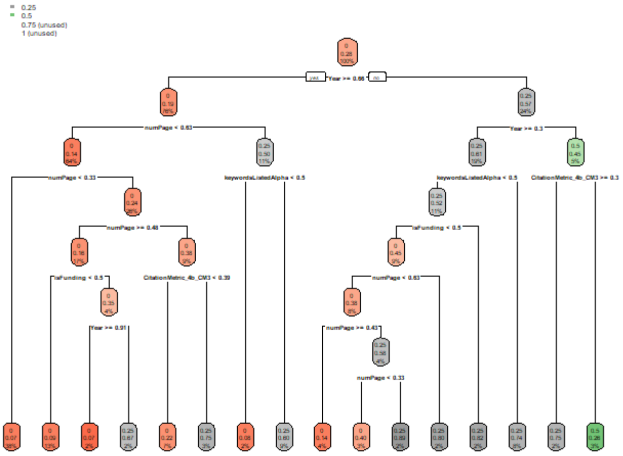


Figure 10 *-Decision Tree*

2.4.3 Decision Tree (for J_or_C)

We applied decision tree algorithm to classify our data. We used knn function for this method. We randomly splitted our data set into train and test sets. The train set contains 75% and the test set contains 25% of the data. We calculated the accuracy and got 95.95% accuracy.

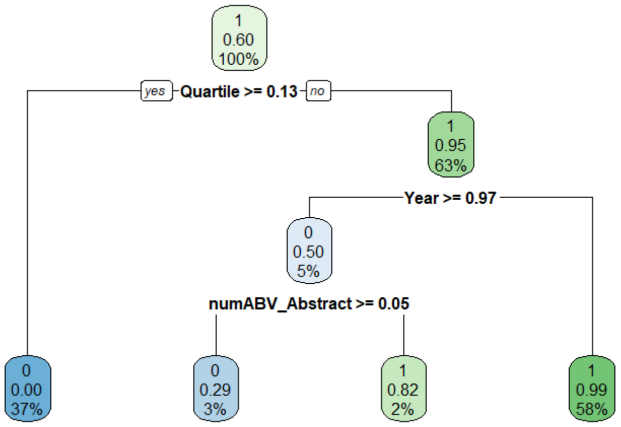
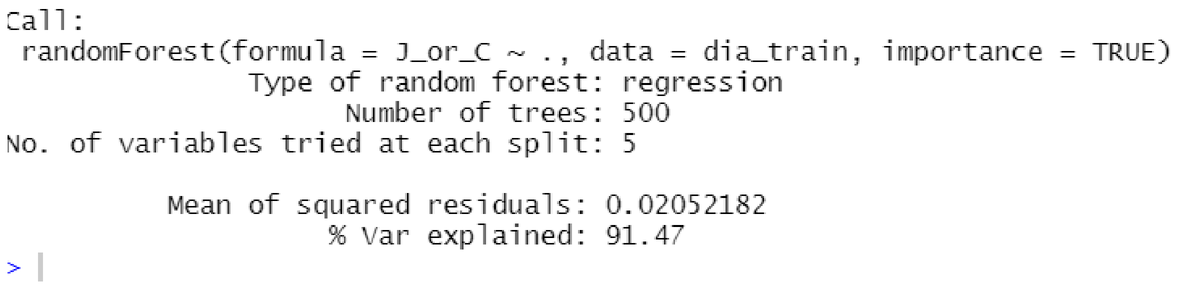


Figure 11*- Decision Tree*

2.4.4 Random Forest(for J_or_C)

We applied random forest algorithm to classify our data. We used random forest function for this method. We randomly splitted our data set into train and test sets. The train set contains 75% and the test set contains 25% of the data. We calculated the accuracy and got 91.47% accuracy.

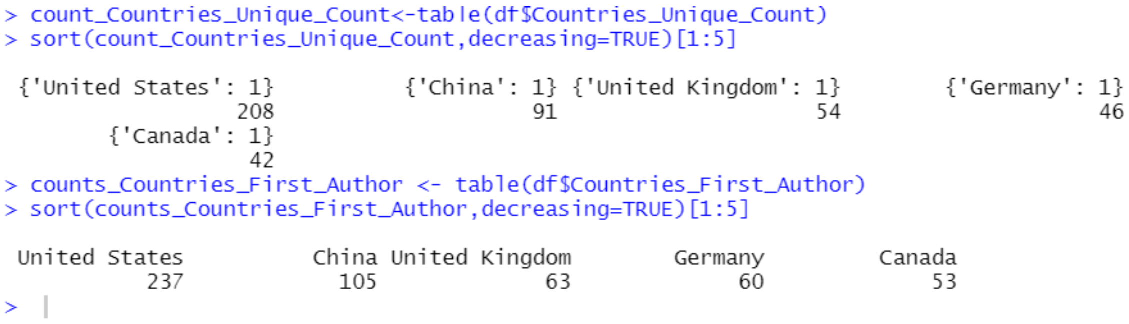


```
Call:
 randomForest(formula = J_or_C ~ ., data = dia_train, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 0.02052182
                    % Var explained: 91.47
> 
```

Figure 10*-Random Forest Result*

3.Exploring Data



```
> count_Countries_Unique_Count<-table(df$Countries_Unique_Count)
> sort(count_Countries_Unique_Count,decreasing=TRUE)[1:5]

 {'United States': 1}          {'China': 1} {'United Kingdom': 1}        {'Germany': 1} 
                  208                    91                    54                    46 
        {'Canada': 1} 
                   42 
> counts_Countries_First_Author <- table(df$Countries_First_Author)
> sort(counts_Countries_First_Author,decreasing=TRUE)[1:5]

 United States          China United Kingdom        Germany         Canada 
           237            105             63             60             53 
> 
```

Figure 11- *Top 5 countries and top 5 first author's countries*

4.RESULTS

In this project, we tried to predict two attributes: the Quartile and the J or C. There are 3 methods we used. For the Quartile attribute, KNN is the best method since we got 83.33% accuracy. For the J or C attribute, the Decision Tree is the best method since we got 95.95% accuracy.