Predicting the Type of Forest Cover using Classification Techniques.

Tanay Kasar
dept-School of Computing
National College of Ireland
Dublin, Ireland
x17156700@student.ncirl.ie

***Abstract*— In the recent world due to deforestation the percentage of land occupied by forests has gone down. There are some forests in the world which haven't been intervened by humans. Analysis can be done on such a piece of land to predict the type of forest cover, type of trees. Also the features which have an influence on the land can be studied. The aim is to find the type of trees covers which dominate a specific portion of land in a forest. The research is implemented on the data obtained from the 'Roosevelt National Forest of Northern Colorado'. Each sample of land is 30m x 30m in area. There are seven types of different forest covers mentioned in the data. The objective is to predict a certain type of forest cover for each of the 15120 sample observations of the land in the forest. Machine learning models like Random Forest, Support Vector Machine, Decision trees have been used for classification. The results obtained were evaluated briefly to select the best model.**

Keywords—Random Forest, SVM, Decision tree, Feature importance

I. INTRODUCTION

The tropical covering over the world varies from region to region. Every region has some unique features of its own. The features and the factors that influence the forest cover of a region can be studied. It is very vital to study these features. By achieving this the trends of the forests can be studied. This can prove very beneficial for nature preservation. This research focuses on the use of Advance Data Mining and Machine Learning concepts to study and predict the type of forest cover. Forest Resource Management is a challenging task. The difficulties arise while making decisions which involve different species and cover types. In the past years the forest authorities find it difficult to make right decisions. The data available is limited. Therefore the data should be studied in dept to gain knowledge from it. This where Machine learning and Data Analytics concepts can be applied. Machine learning models like Random Forest, SVM (Support Vector Machine) and Decision Tree can be used. The above mentioned models are specifically used for classification tasks. All the implementations and model training was done on R-Studio using R-programming language. The implementation is divided into stages. Starting from data loading and pre-processing followed by Data mining processes. In the final stage the classification models are trained and implemented.

The land in the national forest is divided into four territories, depending on their aspects. These areas are named as '**Rawah'**, **'Comanche Peak'**, **'Neota'** and **'Cache la poudre'.** There are 7 types of forests covers and 40 types of soils in the region of our study. These types are discussed briefly in the later sections of the report. The research objectives for this research were formulated on the data that was acquired from the national forest in Colorado.

Research Objectives:

1. To predict the type of Forest cover in the National Forest.
2. To find, which features have more influence on the forest.
3. To find the correlation between the features.

Feature importance technique is used to show the influence of the features on the forest cover. As for finding the co-relation between features corrplot will be used. For predicting the type of forest cover classification models will be implemented. Their results will be evaluated and the better model will be selected on the basis of its accuracy. The data cleaning and pre-processing steps, Exploratory Data Analysis and the model implementation processes are described in the Implementation section. For this research **KDD** (Knowledge Discovery in Databases) methodology was adopted because our goal is to harness knowledge from the forest data. This research will benefit the forest management department in decision making and resource managing. The results obtained after the execution are discussed in the evaluation section. In this section the model results are compared and the output from EDA are explained briefly. In the next section the related work done in the domain will be discussed. The approaches used in the past will be critically reviewed. The techniques using the classification models will also be reviewed. The techniques that are applied and the undergoing work on the topic are discussed and reviewed in the coming section. The data for accomplishing this particular research was obtained from the UCI machine learning repository. Due to the reducing percentage of forest cover on the planet, the remaining forest lands have become endangered. Thus getting insights on such type of data is very crucial. The outcomes of this research will prove useful to develop measures for forest conservations. The task of easy and right decision making for the well-being of the forests lands is the goal. The next section discusses the related work that has been already done on the domain.

II. RELATED WORK

In this section the related applications of advance data mining techniques on the forests has been discussed. The applications of the classification models are also discussed. To show why the specific models were selected.

In this subsection all the research carried on forest areas are reviewed. Techniques like 'Support Vector Machine' and 'Artificial Neural Network' are used to predict the vegetation. The combination of these concepts leads to the formation of a new model. The accuracy obtained using this model was 87.3%. It has better accuracy as compared to the results of SVM model [1]. Random Forest model was used for classification to study the cover type of an island in the Mediterranean sea. The task was to forecast the changes that can occur in the next year. The output of random forest was further executed using Neural Networks. The accuracy obtained with this particular approach was 93%. The accuracy was high due the use of limited data [2]. Models like nearest neighbor, SVM, random forest were used to predict the land cover using satellite images. All of the above mentioned models were trained on the same data. Nearest neighbor and SVM performed the best giving good accuracy. The use of technique known as SMOTE which is used for dealing with class imbalance [3].

The characteristics of the land are studied using images. 8 different types of classifiers were used to find which features have the most influence on the land. Classification models like random forest. The number of images used as a data to train the model was low. The overall accuracy of the model was 82%. Values kappa co-efficient, recall were used to determine the performance of the model [4]. Techniques like SVM are used to find the effect of human population growth on the forest areas. SVM was used to classify the land samples depending the percent of forest on the particular piece of land. Before the developments of machine learning models, statistical approach were used for solving these kinds of tasks. It was concluded that the results of SVM model was better than the statistical approach. Statistical approach was found to be less flexible than the machine learning models [5]. Random forest, k-NN and Naïve Bayes algorithms are often used for classifying different cover types in the forest. The above specified models were trained along with feature selection. In the end the results of the model were compared, in order to select the suitable model. It was observed that Random Forest was the model which gave the highest accuracy. It was concluded that random forest is a reliable model when it comes to prediction of cover type in the forest [6].

Decision trees were used to reduce the error caused in billing of electric city. The data was obtained from an electric company in Beijing. The focus was to add an extra level of training set for the decision tree. The random forest with such decision tree proved to be promising. The results were compared with SVM. The model was beneficial for the electric company as it reduced the rate of error [7]. SVM is one of the most widely used supervised model. Binary SVM was used for multiclass classification. The multiclass data was divided into a number of binary classifications. This approach was compared with the performance of Binary SVM results on the same data. The novel approach gave an accuracy of 92%, while the accuracy for binary SVM was 89% [8]. In Ensemble approach multiple models and concepts are used to solve one task. An ensemble model was implemented, the model was compared with other techniques. The evaluation was carried out using feature importance. It was observed that the use of feature importance was beneficial [9].

In the following paper the author has implemented an SVM model to classify the threats in a network. Decision tree was implemented to train the SVM model. This leads to faster implementation with good accuracy. This implementation proved to be faster in response and precise in threat detection [10]. The data from hyperspectral images was used for study. In order to classify the data from the images n new algorithm was proposed which was based on SVM concepts. Classifying the images by categorizing the pixels, the images contained the data of piece of land. The results of this algorithm were evaluated with the results of SVM and ANN. The accuracy of the new approach was slightly more than the conventional methods. It was concluded that SVM was still a reliable technique when it come to classification of images [11]. An approach to improve the performance of Decision trees was proposed. It includes addition of a balance factor for the trees and the setting the node impurity. The random forest model was implemented on these accurately obtained trees. The performance was evaluated using maximum likelihood method. The accuracy of the model was higher than the previously used technique, the accuracy observed was 87.53% [12].

An attempt was made to improve the performance of LS-SVM. The task was to predict the ATE parameters. After the implementations experiment were carried on the model. It shows that LS-SVM has better accuracy in terms prediction. This approach was accepted by an organization, and now it is used for stability prediction of the parameters [13]. SVM finds it applications in micro-tech domain. SVM is used to classify the images of samples obtained from electron microscopes. SVM was selected due to its precision in classifying the slightest of values [14]. In a particular research the modern classification techniques are discussed. It mentions that there are two types of techniques active learning technique and semi-supervised technique. Semi-supervised technique is better at handling unlabelled data. Active learning is better at giving high precision. After carrying out experiments using both techniques it was found that active learning technique reached convergence with few iterations, and achieved better accuracy [15]. A model was developed for the purpose of text classification. This model was built using concepts of k-NN and SVM. The results of the implemented model was compared with other techniques like naïve bayes and SVM. The new model out performed to the conventional techniques [16].

After reviewing all the referred researches Random Forest, Decision Trees, SVM models will be used as classification models. On the points mentioned in this section. In the next section the methodology which was adopted is briefly described.

III. METHODOLOGY

Depending on the aspects of this research the methodology adopted was KDD (Knowledge Discovery in Databases. KDD is selected when the implementation is carried out for research prospects. It is adopted to observe pattern recognition and to acquire statistical from the system or database. The objective of KDD is to harness the vital knowledge from huge databases. The kdd methodology includes the execution various steps, these steps must be followed in proper order as they are mentioned.

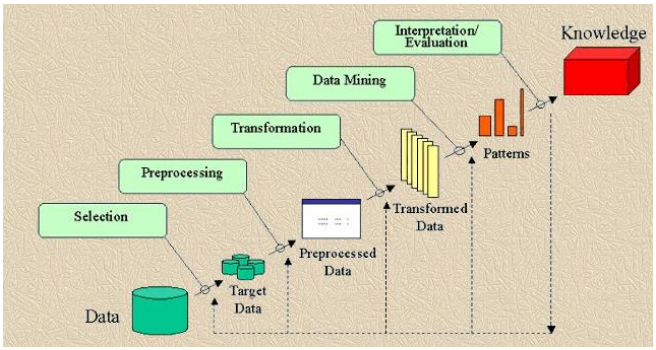

Figure 1: KDD methodology process flow[1]

From the above displayed image it is visible that there are a number steps in the methodology. In the next step the application of these steps on this particular research.

Our main aim is to harness as much as knowledge from the forest data. There are 5 steps mentioned in the methodology that has to be followed to achieve the goal. These steps and the execution involved in these steps have been discussed in this section.

1. Understanding the Data

This includes understanding the data structure. All the important factors and variable in the data should be studied to acquire a better understanding of the data. The domain of the application must also be studied. The benefits of the proposal must be analysed. The objective of the research must be clear. After the clarity is achieved the next step is to select a target variable. In our case we have to predict the type of forest cover. This indicates that our target data is cover type.

2. Studying the target variable

It is clear that cover type is our target variable. In the data of the national forest there are 7 different types of covers they are named as follows Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-Fir, Krummholz [14]. Our task is to predict the cover types for all samples of land. In the next point how the target variable was targeted and the transformation of the dataset is described.

3. Data Pre-processing and Transformation

In this sub section all the steps that were executed from acquiring to its final form is described. Before the splitting the data into training and test sets the data needs to be transformed. The forest is divided into 4 areas of wilderness, mentioned in the earlier sections. Therefore a number of columns were renamed according to the preference. There are 40 types of soils in the data, thus new columns were drafted to bring clarity in understanding the data. There were a number of columns which didn't had any relevance, were not considered. For training the random forest model 2 sample sets were created. For training the SVM model feature engineering was carried out on the data to achieve good accuracy. The cleaned data was split into train and test sets and SVM model was implemented. The classification tree model was implemented using the random forest model to improve its accuracy further.

4. Feature Importance.

Our objective is to find the influence of the features on the cover type of the national forest. For finding the importance of each feature the use of feature importance plot. The feature importance plot was plotted on the random forest model. For plotting the command varImpPlot() was used. The output of the importace plot is shown below. It can be concluded that Elevation has most influence on the cover type of forest, followed by horizontal distance between roads and fire-points.

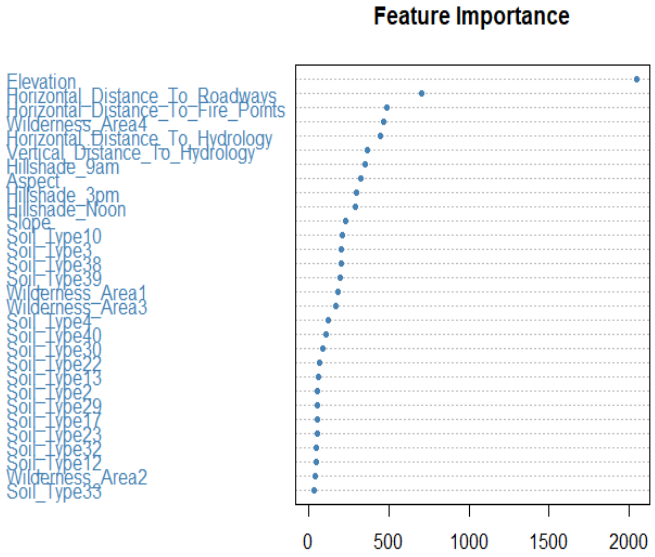

Figure 2: Feature Importance Plot

5. Finding Correlation between features

The next objective is find the correlation between the features. This is achieved using corrplot(). This function is available in the library corrplot. The plot obtainedis displayed below. In the plot the blue dots represent the positive correlation between features. While the red dots represent negative correlation. The next objective is find the correlation between the features. This is achieved using corrplot(). This function is available in the library corrplot. The plot obtainedis displayed below. In the plot the blue dots represent the positive correlation between features. While the red dots represent negative correlation. From the plot displayed below we can conclude Hill shade at noon and hill shade at 3pm are positively correlated. Aspect and hill shade at 3pm are negatively correlated. Elevation and horizontal distance to road are positively correlated. Thus the we have obtained the correlation between the feature

[1]http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

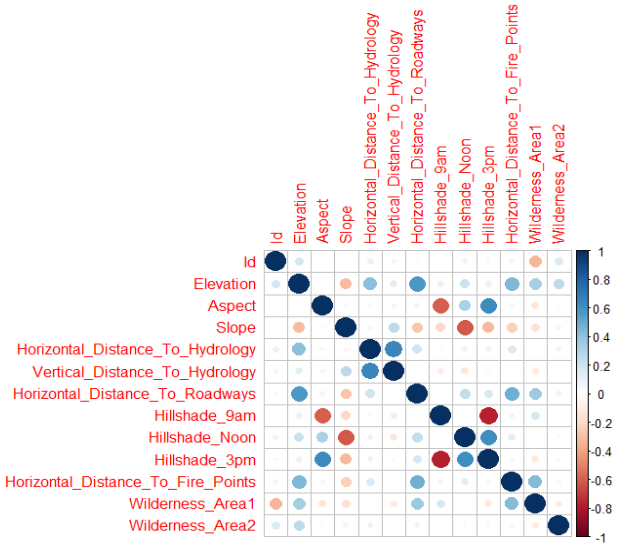

Figure 3: Correlation plot

IV. IMPLEMENTATION AND EVALUATION

After the data has been processed it is ready for training the classification models. These models are implemented to predict the cover type, which is the target variable. The implementation of models like Random Forest, Classification trees and SVM have been discussed in this particular section.

Random Forest:

It is considered as one the most widely used supervised algorithm. It is used for classification problems. Forests are created in which a number of decision trees are implemented. Depending on the value of the predictor, the output is provided randomly by the model. In our case to get good accuracy 300 trees were trained. For the implementation of the model the randomForest() function was used, which is available in the randomForest library of R-studio. After the successful implantation of the model. The confusion matrix was obtained to see the performance of the model. The accuracy achieved was **82.71%.**

Classification Tree:

This model is the extension of the random forest model. It is implemented only to see whether the performance of the random forest can be improved or not. For achieving this rpart() function was used which in rpart package in R-studio. For pruning the tree the prune function was prune function was used. The tree was plotted and the following result was obtained. The confusion matrix for the model obtained and the accuracy obtained was **91.22%.**

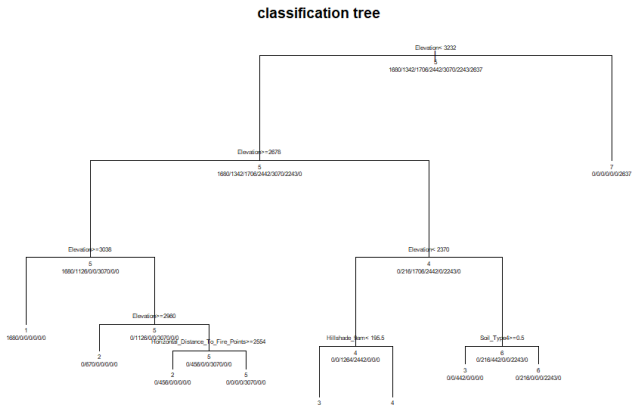

Figure 4: Classification Tree

Support Vector Machine:

SVM is a model used for classification tasks. It is classified as supervised model. It makes the use of statistical nature of support vectors to classify the data which is unlabeled. For implementing the model svm function was used from the e10171 package in R-studio. The target variable used was the cover type. After the implementation of the model, values were predicted. The confusion matrix was obtained to see the results. The accuracy observed was **83.38%.** In the later part of this section all the results of the model implemented will be compared to select the best model.

This section focuses on the comparing the results discussed in the previous section. The model performances will be evaluated by studying the confusion matrix of each model. The confusion Matrices of the random forest model, Classification Tree and SVM are shown below.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6   7
         1 304  89   0   0   1   0  22
         2  49 262   0   0  21   3   0
         3   1  20 318   5   8  40   0
         4   0   0  25 432   0  17   0
         5  14  49   4   0 399   5   0
         6   3  14  79   8   5 363   0
         7  34   3   0   0   0   0 405

Overall Statistics

               Accuracy : 0.8271
                 95% CI : (0.8131, 0.8405)
    No Information Rate : 0.1482
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7983
 Mcnemar's Test P-Value : NA
```

Figure 5.a: Random forest confusion matrix

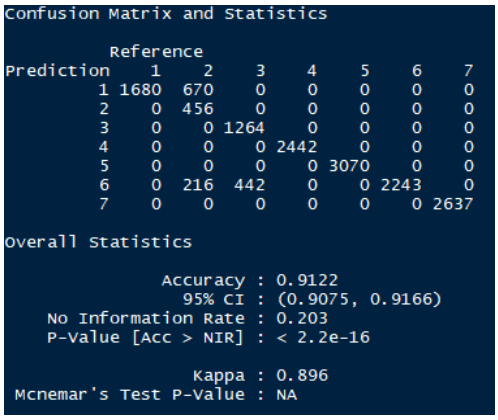

Figure 5.b: Confusion Matrix of decision tree model

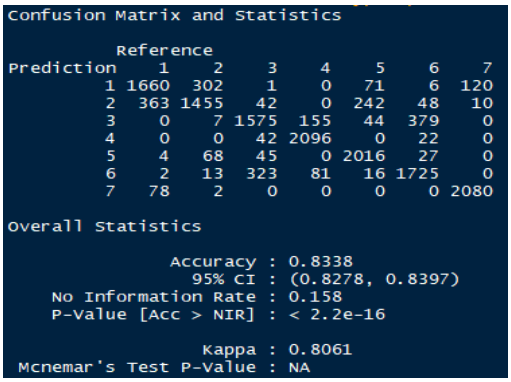

Figure 5.c: Confusion Matrix for SVM model

Significance of kappa value:

It is value which shows the extent of inter-agreement between classifiable terms. The value for must be high as possible.

After studying the confusion matrices we can compare the models for their performances. The kappa value and accuracy obtained will be used to compare the models.

Model Name	Accuracy	Kappa
Random Forest	82.71	0.7983
Decision Tree	91.22	0.896
SVM	83.38	0.8061

Table 1: Comparison of Model Performance

Thus after the comparing the values mentioned in the table it can be concluded that Decision trees model has the highest accuracy and k value. The predictions made by this model will be more precise. The output of this model was converted into data frame and stored in the system. The output file contains the IDs of the sample land areas and the predicted cover variables of the cover type of forest.

V. CONCLUSION AND FUTURE WORK

The objectives described in the introduction were to predict the type of forest cover, to find the influence of the features on the cover type, to find the correlation between the features. From the implementations described earlier in the report we conclude that Elevation has the most influence on the cover type. Followed by horizontal distances to road and fire points. It was found that the values of hill shade at noon and hill shade at 3pm are positively correlated. Elevation and horizontal distance to roads are positively correlated. While Aspect and hill shade are negatively correlated. After comparing the model performances in the evaluation part it can be concluded that Decision tree model gave the best accuracy.

In future, complex clustering techniques can be used to classify the data. The implementation of clustering on this data will prove very beneficial. The land samples can be clustered on the basis of similarity or correlation. Thus better insights can be obtained on the data.

REFERENCES

1. S. Feng, "Predicting forest cover types with immune and genetic", *2012 IEEE 14th International Conference on Communication Technology*, 2012. https://ieeexplore.ieee.org/document/6511338

2. E. Symeonakis, P. Caccetta, J. Wallace, E. Arnau-Rosalen, A. Calvo-Cases and S. Koukoulas, "Multi-temporal Forest Cover Change and Forest Density Trend Detection in a Mediterranean Environment", *Land Degradation & Development*, vol. 28, no. 4, pp. 1188-1198, 2015 https://ieeexplore.ieee.org/document/7730423

3. A. Panda, A. Singh, K. Kumar, A. Kumar, Uddeshya and A. Swetapadma, "Land Cover Prediction from Satellite Imagery Using Machine Learning Techniques", *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018 https://ieeexplore.ieee.org/document/8473241

4. A. Sitthi, M. Nagai, M. Dailey and S. Ninsawat, "Exploring Land Use and Land Cover of Geotagged Social-Sensing Images Using Naive Bayes Classifier", *Sustainability*, vol. 8, no. 9, p. 921, 2016.

5. M.Pratama and A.Arymurthy, "Automatic land cover classification of geotagged images using ID3, Naïve Bayes and Random Forest", *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017.

6. R. Kishore, S. Narayan, S. Lal and M. Rashid, "Comparative Accuracy of Different Classification Algorithms for Forest Cover Type Prediction", *2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, 2016.

7. Y. Zhao and X. Ma, "Study on credit evaluation of electricity users based on random forest", *2017 Chinese Automation Congress (CAC)*, 2017.

8. A. Mathur and G. Foody, "Multiclass and Binary SVM Classification: Implications for Training and Classification Users", *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 2, pp. 241-245, 2008.

9. F. Yang, C. Huang, M. Habibullah, X. Yang, Y. Shen and R. Neo, "Feature importance-guided multi-regression ensemble with application to remaining useful life prediction", *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2017.

10. P. Wang, K. Chao, H. Lin, W. Lin and C. Lo, "An Efficient Flow Control Approach for SDN-Based Network Threat Detection and Migration Using Support Vector Machine", *2016 IEEE 13th International Conference on e-Business Engineering (ICEBE)*, 2016.

11. J. Li, H. Zhang, Y. Huang and L. Zhang, "Classifiction for hyperspectral imagery based on nonlocal weighted joint sparsity model", *2012 4th Workshop on Hyperspectral Image and Signal Processing (WHISPERS)*, 2012.

12. H. Xu, M. Yang and L. Liang, "An improved random decision trees algorithm with application to land cover classification", *2010 18th International Conference on Geoinformatics*, 2010.

13. M. Hongyu, A. Shaolong, Z. Yuchuan and H. Zhuolin, "Prediction for ATE state parameters based on improved LS-SVM", *2013 IEEE 11th International Conference on Electronic Measurement & Instruments*, 2013.

14. Y. Yang, J. Wang and Y. Yang, "Improving SVM classifier with prior knowledge in microcalcification detection1", *2012 19th IEEE International Conference on Image Processing*, 2012.

15. L. Bruzzo and C. Persello, "Recent trends in classification of remote sensing data: active and semisupervised machine learning paradigms", *2010 IEEE International Geoscience and Remote Sensing Symposium*, 2010.

16. F. Miao, P. Zhang, L. Jin and H. Wu, "Chinese News Text Classification Based on Machine Learning Algorithm", *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 2018.

17. Yi Tan and Guo-Ji Zhang, "The application of machine learning algorithm in underwriting process", *2005 International Conference on Machine Learning and Cybernetics*, 2005.

18. X.Wang and S.Lu, "Improved Fuzzy Multicategory Support Vector Machines Classifier", *2006 International Conference on Machine Learning and Cybernetics*, 2006.

19. C. Li, P. Hsieh and B. Kuo, "Multiple SVMS based on random subspaces from kernel feature importance for hyperspectral image classification", *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

20. B.Li, "Importance weighted feature selection strategy for text classification", *2016 International Conference on Asian Language Processing (IALP)*, 2016.

21. http://archive.ics.uci.edu/ml/index.php