Chronic Kidney Disease Prediction Using Machine Learning: A New Approach

[1]Sujata Drall, [2]Gurdeep Singh Drall, [3]Sugandha Singh,

[4]Bharat Bhushan Naib

[1]Research Scholar, C.S.E, PDMCE, India
[2]Research Scholar
[3]Prof. & Head. CSE, PDMCE, India
[4]Assistant Prof., CSE, PDMCE
[1]drallsujata@gmail.com, [2]gurdeepdrall@gmail.com, [3]sugandha_engg@pdm.ac.in,
[4]bharat_engg@pdm.ac.in

Abstract

Chronic kidney disease (CDK) is defined by the presence of kidney damage which lasts longer than three months with decreased glomerular filtration rate (GFR). Chronic kidney disease involves condition like heart disease, high blood pressure or anemia. CDK can be caused by various reasons which include diabetes, high blood pressure, and polycystic kidney disease. People with glomerular filtration rate <60 ml\min/1.73 m^2 for 3 months are classified as having CDK. In this current work focus is on predicting that the patient is CKD or non CKD. To predict this various machine learning algorithms have been use. Different classifications models have been built using different classification algorithms to predict weather the patient is suffering from Chronic disease or not. This Prediction is performed using Naive Bayes Classifier and K-Nearest Neigbour algorithm. The data used is collected from the UCI Repository with 400 data sets with 25 attributes. This data has been fed into Classification algorithms.

The experimental results show that Naïve Bayes Algorithm gives an accuracy of 96.25%, whereas K-Nearest Neigbour came up with an accuracy of 100%.

Keywords – Chronic kidney disease, Data Mining, Machine Learning, Classification algorithm's, Naïve Bayes, K-Nearest Neigbour, feature selection.

1. Introduction

Chronic Kidney disease can also be termed as kidney failure. One in Ten people worldwide are suffering from kidney disease. 10% of the population worldwide suffers from chronic kidney disease; one in five men and one in four women from age group from 60 to 75 have CKD as per National Kidney Foundation [15]. The present work focuses on predicting weather a person is suffering from CKD or not using Data mining using Machine Learning. Data Mining [10] is the process of searching large data sets and discovering patterns and trends and transforming it into understandable data using Data pre-processing, Visualization. Machine Learning is the field of computer which uses statistical techniques to give the ability to learn to computer. Machine learning can be both Supervised and Un-supervised learning. Supervised learning can be defined as when we map an input to a desired output [12], [13]. Machine learning algorithms are provided to support future predictions. There are various

supervised machine learning algorithms like Logistic regression, multi-class classification, support vector machine, K-nearest neighbour, Naïve Bayes, Random forest and many more. In un-Supervised learning algorithm we train the data using info which is not labelled. In this algorithm we divide the data into 2 groups based on similarity and reducing the dimensionality. Most common unsupervised learning approaches are clustering algorithms. There are various clustering algorithms like Hierarchical clustering, K-means clustering and many more [12], [14]. Feature Selection also called as variable selection, attribute selection or feature extraction. It uses relevant data sets and avoids redundant and irrelevant data. It reduces the dimensionality by using small subsets from the original dataset it helps in easy calculation of results and attaining Shorter training times [11]

In this research we have used machine learning to detect CKD and Non-CKD by using 10 attributed which Contribute to kidney disease. The data used consist of record of 400 people. The data set has various missing data. We have used this data sets and classification algorithms to build a classification model for prediction of CKD. The model with the best accuracy prediction is taken. This will help to achieve fast and accurate results for CKD prediction, which will reduce the time for disease prediction and provide benefits to both doctors and patients in providing early treatment and speedy recovery.

Stage of CKD	Clinical Characteristics	GFR (mL/min/1.73 m²)
1	Persistent kidney damage; normal GFR or increase in GFR	≥90
2	Persistent kidney damage; mild decrease in GFR	60–89
3	Moderate decrease in GFR (moderate CKD)	30–59
4	Severe decrease in GFR (severe CKD)	15–29
5	Kidney failure	<15

Figure 1: Stages of chronic kidney disease

Figure 1 explains various stages of chronic kidney disease based on glomerular filtration rate. There are five stages in chronic kidney disease. In Stage 1 the GFR value is equal or less than 90ml/min and the patient do not feel any kind of problem and this is the reason why this stage goes undetected. In stage 2 the GFR value decreases from 60-89ml/min. In this stage kidney suffers from mild damage. In stage 3 the kidney suffers from moderate damage and the GFR value now ranges from 30-59 ml/min. When the GFR value reduces to 15-29 ml/min the patient is suffering from severe chronic kidney disease and this stage is called as stage 4. The patient may now face health issues like high blood pressure, heart problems. In stage 5 the kidney loses its ability to filter out waste material from the blood. The GFR value in stage 5 is less than 15 ml/min, at this stage the patient is advised to undergo medical treatments as the kidneys have lost their ability to filter out waste.

2. Literature Survey

In 2015 Parul et al. conducted a Comparative Study for predicting CKD by using classification algorithms K-Nearest Neighbour and SVM [1]. Performance of classification algorithm has been compared on the basis of accuracy, precision and total execution time for prediction of Chronic Kidney disease. MATLAB was used for this classification model. Performance of K-Nearest Neighbour classifier was 78.75% which was better than Support Vector machine with an accuracy of 73.75%

In 2017 Pinar Yildirim made a classification model for chronic kidney Disease prediction on Imbalanced Data by using Multilayer Perceptron [2]. This work focuses on after effect of class imbalance in training data for predicting CKD or Non CKD. Multilayer perceptron algorithm is used to calculate accuracy. Resample, SMOTE algorithms have been used. The work was performed using 0.8 WEKA 3.7.3 Software, and data for research was taken from UCI Machine Learning Repository of 400 patients with 25 attributes. As per the result Resample Method with Multilayer Perceptron was more accurate, but for Execution time Spread Sub Sample Algorithm was fast with time of 0.0509

In 2017 Gunarathne et al. has made a Performance Evaluation on Machine Learning Classification Techniques for Disease Forecasting and classification through Data Analytics for Chronic Kidney Disease (CKD) [3]. Their area of research is to predict CKD and Non CKD of a patient. Number of Classification algorithms has been used– Multiclass - Decision Forests, Jungle, Logistic Regression and NN. Results were obtained using Microsoft Azure Machine Learning Studio. Result of Multiclass Decision Forest was with highest accuracy of 99.1%

Dr.Uma N Dulhare et al. in 2016 performed Extraction of action Rules for Chronic Kidney Disease Prediction using Naïve Bayes. Naïve Bayes with OneR attribute Selector was used for prediction CKD status of a patient [4]. The idea was to select a subset from input data by elimination idle data which carried little or no predictive knowledge. Data sets were taken from UCI ML Repository. The result and analysis proposed Naïve Bayes with OneR with highest improved accuracy and also reduced number of attributed to 80% which is 05 from total of 25 attributes compared to other attribute evaluators

Dr.N.Radha, S.Ramya in 2016 performed a diagnosis of chronic kidney disease using Machine learning [5] using R tool and algorithms like Back Propagation neural network, Random forest, Radial Basis function, ANN. The data for this research was medical reports of patients taken from different labs in South India. They have used 1000 instances with 15 CKD related attributes. Their model is evaluated on different measures like Sensitivity, Accuracy, and Specificity. The experimental results proved that Radial Basis Function performed better than other algorithms and obtained an accuracy of 85.3%.

Dr.S.Vijayarani, S.Dhayanand in 2015 used Data Mining Classification algorithms for Prediction Kidney Disease [6] using MATLAB tool. Their work focuses on finding best classification algorithm on basis of accuracy and execution time for prediction of Kidney Disease. They have used Naïve Bayes and Support Vector Machine algorithms. In their Prediction model SVM classification algorithm performed better than Naïve Bayes with an accuracy of 76.32

M.P.N.M. Wickramasinghe et al. in 2017 proposed Dietary prediction of patients with CKD by considering Blood Potassium Level [7]. Their work suggests diet plans by taking patients potassium

level in consideration. The experiment is performed using Multiclass Jungle, Forests, and Neural networks in Microsoft Azure Machine Learning Studio. In their results Multiclass Decision Forest performed with an accuracy of 99.17%

Torgyn Shaikhina et al. in 2017 developed classification model for outcome prediction in antibody incompatible kidney transplantation [8]. The base objective is to independently identify risk linked with kidney transplant within first 30 days of transplant that how much the kidney is accepted by the patient's body. This work would help doctors to predict outcomes of kidney transplant at early stage. Decision Tree, Random Forest classification algorithms were used for this prediction. Their work for predicting kidney transplant failure performed with an accuracy of 85%

Radha, N, Ramya,S. in 2015 predicted occurrence of chronic kidney disease using machine learning classification algorithms [9]. In their work the data collected is real data and belongs to the laboratories of south India and consist of record of 1000 people with their respective 14 attributes. In their work they have taken Naïve Bayes, KNN, SVM and decision tree as their classification algorithm for CKD prediction. They have used the same data for all the algorithms. The results were obtained Naïve Bayes classifier performed with an accuracy of 61.85, KNN performed with an accuracy of 98%, SVM performed with an accuracy of 83.9% and decision tree with an accuracy of 78.6%. All the predicted values were compared and KNN algorithms was chosen with best accuracy

3. Dataset And Attributes

We have downloaded Chronic Kidney Disease datasets from publically available data from UCI Machine Learning Repository [16]. Table 1 gives a list of all the attributes taken

Table 1. Attributes for chronic kidney disease prediction

Attribute	Values
Age	Numerical
blood pressure	Numerical
specific gravity	Nominal sg(.005,1.010,1.015,1.020,1.025)
Albumin	Nominal al - (0,1,2,3,4,5)
Sugar	Nominal su - (0,1,2,3,4,5)
red blood cells	Nominal rbc - (normal,abnormal)
pus cell	Nominal pc - (normal,abnormal)
pus cell clumps	Nominal pcc- (present,notpresent)
Bacteria	Nominal ba - (present,notpresent)
blood glucose random	Numerical
blood urea	Numerical
serum creatinine	Numerical
Sodium	Numerical
Potassium	Numerical
Haemoglobin	Numerical
Packed cell volume	Numerical
white blood cell count	Numerical
red blood cell count	Numerical

Hypertension	Nominal	htn - (yes,no)
diabetes mellitus	Nominal	dm - (yes,no)
coronary artery disease	Nominal	cad - (yes,no)
Appetite	Nominal	appet - (good,poor)
pedal edema	Nominal	pe - (yes,no)
Anemia	Nominal	ane- (yes,no)
Class	Nominal	class - (ckd,notckd)

Table 1 describes 25 chronic kidney disease related attributes which are taken form UCI repository, it consist of Record of 400 Patients with 25 attributes. The Data Set is real and consists of Nominal, Numerical and Class attributes.

4. Methodology

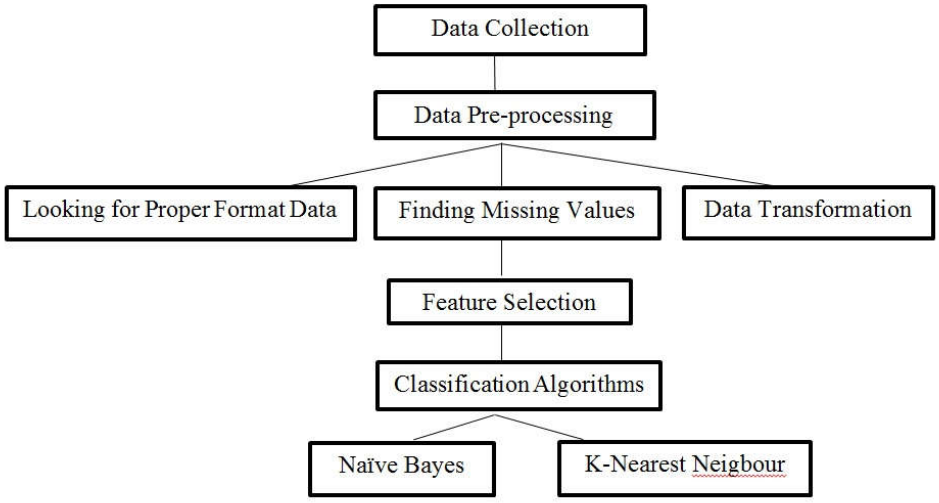

Figure 2. Components of methodology for chronic kidney disease prediction

Figure 2 describes the components of methodology used for CKD prediction. Steps involved in this process of chronic kidney disease prediction are:

4.1. Data Collection

In this research paper we have used Real world data set for predicting CKD status of a patient. The data collected is widely used data and is available at UCI Machine Learning Repository. This Real data belongs to Apollo Hospital in Tamilnadu, India over a period of 2 months. The data set available is specifically used for Chronic Kidney Disease research. It consists of record of 400 people with their respective 25 CKD related attributes. The data consisted of real numbers, Decimal values and Nominal values.

4.2. Data Pre-Processing

Data pre-processing is a way to convert the noisy and huge data into relevant and clean data, as the data available is Real world data, so it contains inaccurate data, missing values and other Noisy data, for removing this inconsistent data from the Dataset, the proposed system have to clean the raw data.

This is an important part to complete the prediction model. It reduces the dimensionality and helps the machine to achieve better results. This is one of the most time consuming stage in building a classification model.

Following data pre-processing steps are followed:

4.2.1. Looking Up For Proper Format: As we have made our model using python, so we need a csv file (comma separated value) for our code. The data downloaded is in the form of RAR file, so we extract the data from the text file available and save it into a csv file so that our python code can read it. This is the first most important step, if the data is not available in requires format then we cannot design the classification model.

4.2.2. Finding Missing Values: When the data collected is real world data, and then it will contain missing values. This brings more change in the prediction accuracy. Sometimes these missing values can be simply deleted or ignored if they are not large in number. It is the simplest way to handle the missing data but it is not considered healthy for the model as the missing value can be an important attribute contributing to the disease. The missing values can also be replaced by zero this will not bring any change as whole, but this method cannot be much yielding. So an efficient way to handle missing values is to use mean, average of the observed attribute or value. This way we lead to more genuine data and better prediction results.

4.2.3. Data Transformation: In this step we transform the given real data into required format. The data downloaded consist of Nominal, Real and Decimal values. In this step we convert the Nominal data into numerical data of the form 0 and 1. The positive value is assigned the value of 1 and the negative value is assigned the value of 0. Now the resultant csv file comprises of all the integer and decimal values for different CKD related attributes.

5. Feature Selection

In this step we select subset of relevant attributes from the total give attributes. This stage helps in reducing the dimensionality and making the model simpler and easy to use, thus leading to short training time and high accuracy.

To obtain highly dependent features for CKD prediction we have used Correlation and dependence method. The term correlation can be defined as mutual relationship between two. In this those attributes are chosen which highly influence the occurrence of Chronic Kidney Disease. By using the correlation it is found that 5 attributed were highly correlated to the occurrence of CKD from the total of 25 attributes.

The 5 attributes selected from a total of 25 attributes are:

1. specific gravity
2. diabetes mellitus_N
3. albumin
4. packed cell volume
5. red blood cells_N

6. Classification Algorithms

we have used K- Nearest Neigbour and Naïve Bayes as our classification algorithm

6.1. K-Nearest Neigbour

For finding a class for a new data point KNN scans through all the previous experiences known as data points and looks up the closest experience to find a solution [17].This algorithm is inspired by human reasoning. The data of previous data points is maintained and the class of a new data point is determined by the majority of nearest data points. This algorithm is fast and easy

6.2. Naive Bayes

Naïve Bayes are probabilistic classifiers, which are based on Bayes Theorem [18]. In Naïve Bayes each value is marked independent of the other values and features. Each value contributes independently to the probability. The higher the probabilistic value the higher are the chances of data point belonging to that class or category. Naïve Bayes algorithm uses the concept of Maximum Likelihood for prediction. This algorithm is fast and can be used for making real time predictions such as sentiment analysis.

7. Results

This study is carried to predict weather a patient is suffering from Chronic Kidney Disease or not. This Prediction model is created in Python programming language. In our classification model we have used K-Nearest Neigbour and Naïve Bayes as our classification algorithms; both the classification algorithms were applied to the same data set collected from UCI Repository.

Filtered dataset of 400 people with 5 CKD related attributes from a total of 25 attributes is used. The Filtered attributed are - Specific gravity, diabetes mellitus_N, albumin, packed cell volume and red blood cells_N.

Table 2. Predictive accuracy of classification algorithms

Algorithm	Accuracy
Naïve Bayes Classifier	96.25%
K-Nearest Neigbour	100%

Table 2 represent the prediction accuracy of both Naïve Bayes and K-Nearest Neigbour algorithms. Both the prediction accuracies are compared. Naïve Bayes performed with an accuracy of 96.25% and KNN performed with an accuracy of 100%.

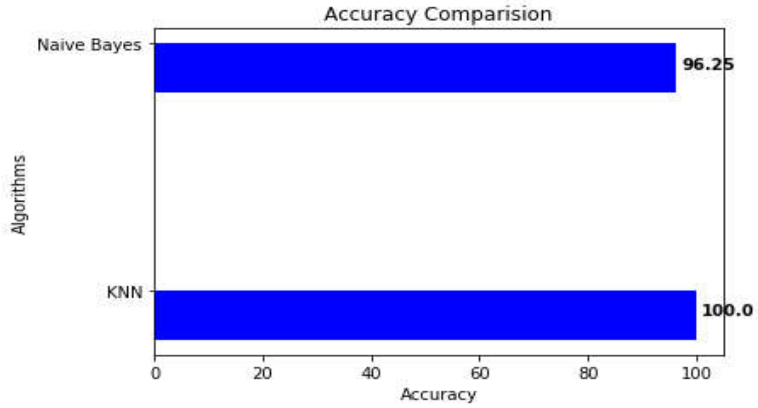

Figure 3. Graphical representation for accuracy comparison of KNN and Naïve bayes classification algorithm

Figure 3 is the graphical representation of both the prediction algorithms, KNN with 100% accuracy and Naïve Bayes with an accuracy of 96.25%. In the above figure x-axis represents the accuracy value and the y-axis represents the algorithm used. The above results highlight that accuracy of KNN algorithm is 3.75% higher than Naïve Bayes classification algorithm. The experimental results show that Chronic Kidney Disease can be better predicted by using K-Nearest Neigbour algorithm with 100% accuracy. The advantage of this research is that it will help Doctors to easily predict CKD with high accuracy and precision in less time period.

8. Conclusion

Various Researches have been made in the field of Chronic Kidney Disease prediction using Data Mining, Machine learning and different classification algorithms.This work is focused on predicting CKD status of a patient with high accuracy. Early and Accurate detection of CKD can helpful in preventing further deterioration of patient's health. In this research we have used 5 CKD related attributes from a total of 25 attributes and two classification algorithms KNN and Naïve Bayes for predicting CKD status of a patient. Same data set of 400 people was given to both the classification algorithms and results were obtained. We have compared the results of both the algorithms on the basis of accuracy. KNN classifier predicted chronic kidney disease with an accuracy of 100%, whereas Naïve Bayes Classifier predicted with an accuracy of 96.25%.Thus KNN performing better than Naïve Bayes with high accuracy.In conclusion this study helps doctors to predict the disease more accurately and in no time and the patient's to undergo minimal test as compared to large number of test required for CKD prediction.

References

[1] Sinha, Parul, and Poonam Sinha. "Comparative study of chronic kidney disease prediction using KNN and SVM." International Journal of Engineering Research and Technology 4, no. 12 (2015): 608-12.

[2] Yildirim, Pinar. "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction." In Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, vol. 2, pp. 193-198. IEEE, 2017.

[3] Gunarathne, W. H. S. D., K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi. "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)." In Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on, pp. 291-296. IEEE, 2017.

[4] Dulhare, Uma N., and Mohammad Ayesha. "Extraction of action rules for chronic kidney disease using Naïve bayes classifier." In Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on, pp. 1-5. IEEE, 2016.

[5] Ramya, S., and N. Radha. "Diagnosis of chronic kidney disease using machine learning algorithms." International Journal of Innovative Research in Computer and Communication Engineering 4, no. 1 (2016): 812-820. IJIRCCE,2016.

[6] Vijayarani, S., and S. Dhayanand. "Data mining classification algorithms for kidney disease prediction." International Journal on Cybernetics and Informatics (IJCI) (2015).

[7] Wickramasinghe, M. P. N. M., D. M. Perera, and K. A. D. C. P. Kahandawaarachchi. "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms." In Life Sciences Conference (LSC), 2017 IEEE, pp. 300-303. IEEE, 2017.

[8] Shaikhina, Torgyn, Dave Lowe, Sunil Daga, David Briggs, Robert Higgins, and Natasha Khovanova. "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation." Biomedical Signal Processing and Control (2017).

[9] Radha, N., and S. Ramya. "Performance Analysis of Machine Learning Algorithms for Predicting Chronic Kidney Disease." (2015).

[10] https://en.wikipedia.org/wiki/Data_mining.

[11] https://en.wikipedia.org/wiki/Feature_selection.

[12] https://machinelearningmastery.com/supervised-and-unsupervisedmachine-learning-algorithms/

[13] https://www.medicalnewstoday.com/articles/172179.php

[14] https://www.medicalnewstoday.com/articles/172179.php

[15] https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease

[16] https://archive.ics.uci.edu/ml/index.php

[17] https://www.kdnuggets.com/2017/09/rapidminer-k-nearest-neighbors-laziest-machine-learning-technique.html

[18] http://www.statsoft.com/textbook/naive-bayes-classifier

[19] Sheng, Gehao, Huijuan Hou, Xiuchen Jiang, and Yufeng Chen. "A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model." IEEE Transactions on Smart Grid 9, no. 2 (2018).

[20]Choi, Tsan-Ming, Hing Kai Chan, and Xiaohang Yue. "Recent development in big data analytics for business operations and risk management." IEEE transactions on cybernetics 47, no. 1 (2017): 81-92.

[21]Serpen, Alexander Arman. "Diagnosis Rule Extraction from Patient Data for Chronic Kidney Disease Using Machine Learning." International Journal of Biomedical and Clinical Engineering (IJBCE) 5, no. 2 (2016): 64-72.

[22] Mao, Yi, Yixin Chen, Gregory Hackmann, Minmin Chen, Chenyang Lu, Marin Kollef, and Thomas C. Bailey. "Medical data mining for early deterioration warning in general hospital wards." In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pp. 1042-1049. IEEE, 2011.

[23] Ghotra, Baljinder, Shane McIntosh, and Ahmed E. Hassan. "A large-scale study of the impact of feature selection techniques on defect classification models." In Mining Software Repositories (MSR), 2017 IEEE/ACM 14th International Conference on, pp. 146-157. IEEE, 2017.

[24] Ma, Lei, Manchun Li, Yu Gao, Tan Chen, Xiaoxue Ma, and Lean Qu. "A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation." IEEE Geoscience and Remote Sensing Letters 14, no. 3 (2017): 409-413.

[25] Sin, Katrina, and Loganathan Muthu. "APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW." ICTACT Journal on soft computing 5, no. 4 (2015).

[26] Dutt, Ashish, Maizatul Akmar Ismail, and Tutut Herawan. "A systematic review on educational data mining." IEEE Access 5 (2017): 15991-16005.

[27] Dean, Jeff, David Patterson, and Cliff Young. "A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution." IEEE Micro 38, no. 2 (2018): 21-29.

[28] Thirumalai, Chandrasegar, Anudeep Duba, and Rajasekhar Reddy. "Decision making system using machine learning and Pearson for heart attack." In Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of, vol. 2, pp. 206-210. IEEE, 2017.

[29] Madni, H. A., Anwar, Z., & Shah, M. A. (2017, September). Data mining techniques and applications—A decade review. In Automation and Computing (ICAC), 2017 23rd International Conference on (pp. 1-7). IEEE.2017.

[30] https://en.wikipedia.org/wiki/Precision_and_recall