Prediction of Tuberculosis disease using Data Mining Algorithms

Saksham Maggo, Anmol Gupta, Sahil Jamwal, Prachi Setia, Sonia Rathee
{maggosaksham16@gmail.com , anmolgupta.ag111@gmail.com,
sahil.jamwal78625@gmail.com}

Department of Computer Science, Maharaja Surajmal Institute of Technology, New Delhi

Abstract. The expectation and finding of Tuberculosis survivability have been a difficult research issue. Since the early dates of the related research, much headway has been recorded in a few related fields. For example, the biomedical advancements have better logical prognostic elements are being estimated and recorded; the low-cost computer technology and the hardware gives better quality information and the data which is gathered has been analyzed by using the different analytics methods. Tuberculosis is one of the main illnesses for all individuals in developed nations including India. It is the most widely recognized reason for death in individual. The high occurrence of Tuberculosis in all individuals has expanded essentially in the most recent years. In this paper we have talked about different data mining approaches that have been used for Tuberculosis diagnosis .and anticipation. Here, we exploited those techniques which gives better prediction results of the Tuberculosis survivability.

Keywords: Data Mining, Tuberculosis, PLS-DA, MLR, K-NN, Mycobacterium, K-means,Apriori,LDA.

1. Introduction

Data mining is a process in which patterns in large datasets are discovered and analysed in order to get some new information that may otherwise difficult to find. It is an interdisciplinary subfield of computer science and statistics which aims to extract information from a data set and transform the information for further use. It is the analysis step of the "knowledge discovery in databases". It is looking for hidden, valid & potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected previously unknown relationships amongst the data [1]. Though data mining is a new technology, it is frequently used in many industries including banking sector, insurance companies, telecommunication sector and retail industry. These industries may combine data mining technique with other tools like pattern recognition and statistics. Data mining algorithms such as classification, clustering etc, can be used to find the patterns for deciding the future drift in businesses. It also helps them to discover the needs and demands of their customers so that correct marketing decisions can be made which will ultimately help in growth and development of industry. Its Implementation

ICIDSSD 2020, February 27-28, New Delhi, India

DOI 10.4108/eai.27-2-2020.2303126

Process includes Data understanding, Data preparation, Data transformation, Modeling, Evaluation and Deployment.

Tuberculosis, popularly known as TB, is an infection which is caused by a bacteria called Mycobacterium and it usually spreads from one person to another through air. It mainly influence the lungs, but it may affect other organs and tissues as well mean it can attack almost any part of the body [2]. There are certain age groups which are more vulnerable to young adults which are working in developing countries, health care officers in medical institutes, people with weak immune system and those who are having HIV or are frequent smoker [3].The person suffering from active TB may release the bacteria in air by breathing, speaking, spitting, coughing, sneezing, etc. and if someone breathes in these germs then there is a chance that he will also get infected. In this way TB get transmitted. Tuberculosis infection can result in joint destruction and blood in urine, cardiac tamponade in heart, spinal pain in bones and meningitis in brain.

1.1 Tuberculosis types

Tuberculosis can be diagnosed in two ways:

Active TB Latent TB

1.1.1. Active Tuberculosis: In Active Tuberculosis the bacteria is active and symptoms of Tuberculosis are usually visible in patient. If the lungs get affected by this kind of bacteria then it can create a hole in the lung. It is caused when body resistance power (immune power) is minimum. Usually older persons and children have weak immunity [2].

1.1.2. Latent Tuberculosis: In Latent Tuberculosis the bacteria is usually in sleeping state and is not transmissible. But still it can become active later on. Usually the symptoms of Tuberculosis are not visible in-patient body and thus it is difficult to diagnose.

1.2 Symptoms

While no symptoms can be observed in Latent Tuberculosis but a person with Active Tuberculosis may have any of the following symptoms such as persistent cough (for more than 3 weeks), coughing up blood, High Fever, Chest Pain, Weight Loss, Night Sweats, constant fatigue, loss of appetite [4].

1.3 Facts about Tuberculosis

Though both sexes and all age groups are equally vulnerable to TB but according to a report, in 2018, men with age more than 14 years were observed to be at highest risk with 57% share in total TB cases. On the other hand, Women and children with age less than 15 years only accounted for 32% and 11% respectively. In the Global Tuberculosis report published by WHO in November 2019 most TB cases reported in 2018 were from the WHO Regions of South-East Asia which is 44% followed by 24% in Africa .It was observed that countries with low incomes have more cases of TB as compared to high income countries .The countries where the influence of TB is seen more are Pakistan , Nigeria , Bangladesh and South Africa .

2. Methodology

Six data mining algorithms are discussed below. In these algorithms, using the data instances we predict the group membership. It is classification of data in certain class. It is based on the building of training set. A model is built after the classification algorithm is trained by the training set. These techniques are used in predicting and diagnosing the Tuberculosis disease.

2.1 Linear Discriminant Analysis

Linear discriminant Reduction analysis is a technique used to convert the data having multiple dimensions to a lower dimensional space i.e. only two or three dimensions and hence pre-process our data for further use. Dimension reduction can be achieved by removing dependent or redundant features from original data [6] keeping in mind that minimum or no loss of information will be there so that our data will only consist of relevant features. We have to assume that the data is gaussian and each variable is deviated from its mean value equally [7]. Reducing the dimensions will help in classification of data points in a more precise way as after applying this technique the classes produced will be such that the data point belonging to same class will have similar properties and data points from different classes will be significantly different from each other [8]. There are two types of LDA techniques: class dependent and class Independent. In class dependent technique, for each class separate lower dimension class will be generated and it is used to minimize the ratio of within class variance to between class variance. In class independent technique, a common lower class will be generated for all the classes and it is used to minimize the ratio of within class variance to overall class variance. This technique is commonly used in chemistry, bioinformatics, data mining, biometrics, machine learning and medicine [6].

2.1.1 Steps in Algorithm:

Step 1: Calculate the within class scattered matrix.

$$Sw = \sum_{i=1}^{k} Si \tag{1}$$

where k is total number of classes and

$$Si = \sum_{x \in Di}^{n} (x - mi)(x - mi)^{\mathrm{T}} \tag{2}$$

$$mi = \frac{1}{ni} \sum_{xi \in Di}^{n} xk \tag{3}$$

Step 2: Calculate between class scattered matrix.

$$Sb = \sum_{i=1}^{c} Ni(mi - m)(mi - m)^{\mathrm{T}} \tag{4}$$

where *Ni* is size of ith class, *mi* is sample mean and *m* is overall mean.

Step 3: For each scatter matrix calculate eigen vector and corresponding eigen value.
Step 4: Among the eigen vectors select k eigen vector that have the highest eigen values.
Step 5: Make a matrix of eigen vectors in which every column represents an eigen vector.

Step 6: Project the data on a lower dimension space to obtain new features using the equation

$$Y=X \times W \tag{5}$$

where W is eigen-vector matrix, X is n×d matrix representing samples and Y is the transformed matrix.

2.2 k-NN Algorithm

The k-Nearest-Neighbors (k-NN) is a simple, effective, powerful and a non-parametric classification method [9]. Among all machine learning algorithms, it is one of the simplest methods [11].Classification, regression and pattern recognition are the major fields in which k-NN is used widely [13] .It is used when there is little or no prior knowledge about the distribution of the data [12]. In k-NN, the rough or imprecise value of k is chosen and classification success is highly dependent on this k value chosen. So, the k-NN method is influenced by k. It first decides an appropriate value of k, then chooses the k nearest neighbors and then based on the majority it assigns the new data point to the selected category.

2.2.1 Steps in Algorithm:

Let us take a simplest 2D dataset. Take two categories of points belonging to different classes (let's say category1 has green colour points and category 2 has red colour points) distributed in the 2D space. Each point represents a datapoint. Here the green colour datapoint represents a positive datapoint and a red colour datapoint represents a negative datapoint.

Let $D=\{(Z_i,S_i) \mid Z_i \in R^2 , S_i \in (0,1)$ where 0 stands for positive data point and 1 stands for negative datapoint$\}$.

Step 1: Select the number of neighbors i.e. the value of k where k can be 1, 2, 3……. Common default value taken is k=5.Avoid choosing the even value of k. The value of k is chosen by various means, but the simplest among is to choose distinct values of k and to execute algorithm on these values and decide that value of k that gives best result [9].

Step 2: According to the distance calculated by the Euclidian equation, k nearest neighbors are chosen for the new datapoint Zq, Let these k nearest neighbors are named as Z_1, Z_2,Z_3,Z_4 and Z_5. Euclidian distance between two points P_1 (X_1, Y1) and P_2 (X_2, Y_2) is

$$\sqrt{((X_2-X_1)^2 + (Y_2-Y_1)^2)} \tag{6}$$

Step 3: Out of the k neighbors, compute the number of data points in both categories i.e. for each datapoint Z_1, Z_2,Z_3,Z_4 andZ_5, find out the value of S_1, S_2, S_3,S_4 and S_5. Let's say value of S_1, S_2, S_5 comes out to be 0 belonging to category 1 and that of S_3, S_4 comes out to be 1 belonging to category 2.

Step 4: New data point Zq is allocated to one of those categories where you find out most neighbors. So, determining using majority, the new datapoint belong to category of green colour datapoints i.e. category 1.

Step 5: Model is ready.

2.3 K-means Algorithm

K-means algorithm is a partitioned clustering algorithm that works by dividing various data points into set of similar data sets having Kcentroids (i.e. each cluster is associated to a centroid). Here K denotes the number of centroid points (i.e. the arithmetic sum of all the data points belonging to a particular cluster) and each point is associated with the closest centroid [14].It is an iterative approach that keeps on segregating data points into clusters until the centroids are found to be stable (no change in their value is observed). Homogeneity of data points is observed when there is less variation within the clusters. The assignment of data points to a cluster is based on the minimum sum of square distance value between the data point and the assumed centroid and after assigning the new data points again the value of centroid is computed and the process is repeated until there is no change in centroid value of all the datasets [15]. Expectation Maximization Technique is followed by the K Means algorithm, assigning data points to closest cluster corresponds to Expectation step and computing centroid of each cluster corresponds to Maximization step. K means is quite popular application for analysing academic performance, market segmentation, image segmentation, diagnostics system, image compression, pattern detection [16].

2.3.1 Steps in Algorithm:

Task: Given a D - dimensional space having a set X of n points and an integer value K. Group the points into C={C_1,C_2,.....C_K} clusters such that

$$\text{Cost(C)} = \sum_{i=1}^{k} \sum_{x \epsilon Si} |x - ci|^2 \quad (7)$$

is minimum and c_i represents the i^{th} centroid point(i.e. mean of the points in a Cluster(C_i))and every data point in $X \epsilon S_i$and $S_i \cap S_j = \emptyset$.But this computation is very costly and is a NP-Hard Problem. So, by using Lloyd's Algorithm which is a type of approximation Algorithm this job can be easily done in 4 simple steps:

Step 1: Initialization Phase: Randomly pick K points from set D and take them as centroids C_1,C_2….C_k.

Step 2: Assignment Phase: For each point x_i (i=1, 2, 3…. n) in D, Select the nearest centroid point C_j(j=1, 2, 3…. K) by computing the value of distance between the data point and the centroid. Add x_i to the nearest Set S_j.

Step 3: Recompute centroid and Update: Recalculate the value of C_j again using equation (8):

$$Cj = \frac{1}{|Sj|} \sum_{xi \in Sj} xi \quad (8)$$

Here $|S_j|$ represents the no of points in Set S_j.

Step 4: Repeat Steps-2 and 3 until convergence is hit i.e. centroids does not change from their old values.

2.4 Apriori Algorithm

Various mining algorithms of association rules has been proposed till now. One of the most popular is Apriori algorithm. It is the most classical and important algorithm for mining frequent item sets. Frequent item sets from large database are extracted and then getting the association rule [18]. This theory is based on the basic principles that all the non-empty subset of frequent item set must be frequent and the supersets of infrequent item set are infrequent item sets [19].It is more efficient than the level wise generation of frequent data sets because it uses less search space.It is a seminal algorithm which uses candidate generation ingeneratingfrequent item sets.It basically works in two steps:

Step 1(Join step): Using linking process, candidate itemset is generated.Self-join of frequent elements of previous level is done for determining the elements of the next level.

Step 2(Prune step): Only those frequent itemsets are kept which satisfies the threshold support by scanning the database[17]. Major terms used in Apriori Algorithm are:

Support (S): It is the popularity of an item set i.e. what is the frequencyof occurrence of an item.

Support(item) = No. of transactions in which the item has appeared /Total no. of transactions.

Confidence: Determines how often an item occurs in transactions with respect to another item already contained in those transactions [17].

$$confidence(M \rightarrow N) = \frac{Support(M\ U\ N)}{Support(N)} \tag{9}$$

Lift: It is given as:

$$lift(M \rightarrow N) = \frac{Support(M\ U\ N)}{Support(M) * Support(N)} \tag{10}$$

While taking into account the popularity of Y, the lift is the likelihood of the itemset Y being purchased when item X is purchased.

2.4.1 Steps in Algorithm:

The following are for steps in the Apriori Algorithm.

Step 1: The algorithm will count the occurrences of each item. Start with item sets containing just a single item i.e.each item is taken as a 1-itemsets candidate in the first step of the algorithm.

Step 2: Only those item sets above the threshold support are accepted and rest others are rejected.

Step 3: All the possible itemset configurations are generated using the itemsets from Step 1.

Step 4: Steps 1 & 2 are repeated until there are no more new item sets.

2.5 Multinomial Logistic Regression

Multinomial logistic regression is used to find the probability of a dependent variable with the help of one or more independent variables [22] and also predict how dependent variable is related to one or more given independent variables. It is used when we have more than two, nominal and unordered categories and observations are assumed to be independent [23]. It uses the concept of binomial logistic regression model in a generalised way which categorize the features into two classes i.e. 0/1.It calculates k-1

binary regression models for k categories and for each category there is a variable that is set to 1 and all others are set to 0[24].We can assume any category (generally kth category) as reference or baseline category. To ensure that the probability is between 0 & 1 it uses logit model [25].With independent variables it uses some other parameters that are determined according to the given problem using maximum likelihood estimation [26].Multiple logistic regression is widely used in machine learning, mathematical finance, psychology and medicine. It is also used in risk analysis [27].

2.5.1 Steps in Algorithm:

Step 1: Assign integer values to the features present in the database from 1 to k. These features will act as input.

Step 2: Choose a reference or baseline category among the given categories. Using this reference category all the other categories is determined separately.

Step 3: Let outcome k is selected as reference, applying logit model

$$\Pr(y_i=k)=1/1+(e^{w_1 x_i}+ e^{w_2 x_i}+\ldots\ldots+e^{w_{k-1} x_i}) \quad (11)$$

and

$$\Pr(y=j)=e^{w_j x_i}/1+(e^{w_1 x_i}+ e^{w_2 x_i}+\ldots\ldots+e^{w_{k-1} x_i}), \text{ for } j=1,2,\ldots,k-1; \quad (12)$$

where Pr (y_i=k) denotes the probability that ith observation will have kth outcome.

Step 4: Find the value of unknown coefficients w_k using maximum likelihood estimation [27].

2.6 PLS – DA (Partial Least Squares Discriminant Analysis)

It is a linear classification method which has properties of both partial least squares regression and discrimination power of classification technique [29]. Experimental group membership bring about the value of variable Y which is mapped into a linear space and it is basically based on PLS regression (PLS-R)

Since the final predictive model can be reduced to standard linear form therefore this algorithm can be treated as linear regression method.

$$Y^* =\beta_0+\beta_1 x_1+\beta_1 x_2+\ldots+\beta_n x_N \quad (13)$$

where Y* is the model prediction and β_0, $\beta_1 \ldots \beta_N$ is a vector of PLS coefficients and [30]. (PLS-DA target variables (Y1, Y2, Y3 …) from the values of several input variables (X1, X2, X3…). Initially PLS Regression was defined for prediction of continuous variables but now it can also be used in forecasting the values of discrete variables in the problems like Supervised Learning It is more often used in smaller datapoints(samples) when the number of independent variables are large. It is used in the inspection of multivariate dataset which can also be seen in derivation from NMR based metabolomics [31]. Applications include in various fields like banking sector, agriculture related studies, forensics and medical science etc.

2.6.1 Steps in Algorithm:

PLS-DA algorithm consists of 2 main steps:

Step1: Construction of PLS Component it is sometimes also referred as dimension reduction.
Step2: Construction of Prediction Model also called discriminate analysis

Table 1. Pros and cons of various data mining techniques.

	Algorithm	Advantages	Disadvantages
1.	**LDA Algorithm**	1. As features get reduced to a great extent, it reduces the cost of computation. 2. Unlike MLR, it will not lose stability when classes become well-separated [7].	1. We cannot use this technique when samples in data matrix are significantly lower than number of dimensions [6]. 2. LDA cannot differentiate between non-linearly separable classes [6].
2.	**k-NN Algorithm**	1. It is one of the simplest algorithms, easy to understand and implement [10]. 2. There is no prior need of calculating or approximating parameters and thus no guidance is required. 3. Since, the algorithm does not acquire any knowledge from the training dataset, so it is easier to add new data.	1. The cost of computation of distance between the new and every existing data point becomes difficult when there is large dataset and which in turn break down the performance of the algorithm. 2. It is less efficient as it does not acquire any knowledge from the training set which results in restricting it in many applications [10].
3.	**K-Means Algorithm**	1. Algorithm is quite effective unsupervised method and works well with large-scale data. 2. .Easy to implement as compared to other complex clustering algorithms as no distributional assumption of data is assumed. 3. K-Means is very effective in its job if the clusters are almost spherical in shape.	1. This algorithm suffers asshape of data deviatesfrom spherical to other shapes and hence performs poor clustering. 2. It does not let the data points far awayfrom each other sharethe same cluster evenif theybelong to thesame cluster. 3. Random initialization used introduces initialization sensitivity.

4. Apriori Algorithm	1. Among association rule learning algorithms, it is easily implementable and understandable [17]. 2. The derived rules are simpler to understand by the user. 3. It makes use of large item sets property.	1. For scanning the database, it takes large time [20]. 2. When the minimum threshold is low a large number of candidate sets containing regular item sets are generated, wasting a lot of time [21]. 3. Large numbers of in-frequent item sets are generated and thus increase the space complexity [20].
5.MLR Algorithm	1. It is very easy to implement and interpret and does not require independent variables to be in intervals [26]. 2. We can use this technique in the case where dependents have more than two classes [24]. 3. We can bound independent variables in case of MLR [26].	1. Being linear in nature, it cannot be uses for non-linear problems. 2. We cannot use this model when independent observations are related to each other. 3. We can only consider independent observations while working with MLR [28].
6.PLS-DAAlgorithm	1. It can handle more descriptor variables than compounds. 2. It provides more predictive accuracy and is having lower risk of chance correlation [32].	1. It is having very high risk of overlooking 'real' correlation. 2. It is sensitive to the relative scaling of descriptor variables [32].

3. Data Analysis

The analysis of the tuberculosis disease can be done by using the details of 500 patients. The data should pe placed in one place which contains multiple records. Each data has an enough information of single patient. Initially we get the symptoms of the patients and go for test which is required by the doctor. The test details of the patients have so many attributes but we select only 8 attributes. The Table 2 explains the names of 8 attributes with their data types(DT). N represents the Numerical and C represents the Categorical.

Table 2. List of attributes and data types

S.NO	Name	DT
1.	Age	N
2.	Chronic Cough	N
3.	Weight loss	C
4.	Sputum	C
5.	Wheezing	C
6.	Chest Pain	C
7.	TB type	C
8.	Blood cough	C

4. Result and Discussion

The classification of tuberculosis disease can be done by using the two classifiers i.e. performance and accuracy. Error rate and computation time used for calculating the performance. The accuracy depends on the sensitivity and specificity. The computation time of each classifier is considered. The frequency of correct and incorrect predictions can be display by using the classification matrix shown in table 3.

Table 3. Confusion matrix

Predicted	Classified as Healthy (0)	Classified as non-healthy(1)
Actually Healthy	TP	FN
Actually non-healthy	FP	TN

Table 3 explains the classification matrix for all the six models. The predicted value represents by row and actual values represent by columns (patients with tuberculosis is represented by 1, and patients without tuberculosis is represented by 0). We use the mathematical equations for measuring the accuracy, sensitivity, specificity and error rate.

Table 4. Various formulas used in analysis

S.NO	Formulas
1.	$Sensitivity = \frac{TP}{FN + TP}$
2.	$Accuracy = \frac{TP + FN}{TP + FP + TN + FN}$
3.	$Specificity = \frac{TN}{FN + TP}$
4.	Error Rate $= \frac{FP+FN}{TP+FP+TN+FN}$

The calculation of sensitivity, specificity, error rate and accuracy can be obtained by using the distinguished confusion matrix.

Step 1: By using computing time (<1200ms) we filtered those 6 algorithms. Those six can be reduced to five algorithms.

Step 2: The positive precision values are used for the filtration of the above algorithms. If the precision value is less than 0.05, then we get the four algorithms.

Step 3: By using the cross validation error rate which is low(<0.06) can filter the above remaining algorithms, then we get three algorithms

Step 4: By using the bootstrap validation error rate (<0.29) can filter the above remaining algorithms, then we get two algorithms.

Step 5: The highest accuracy and lowest computing time used for the remaining algorithms. We get the best algorithm i.e. PLS-DA.

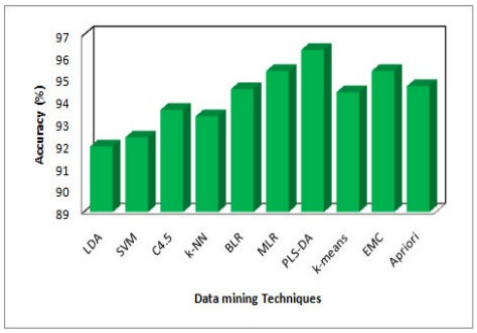

Figure 1. Predicted Accuracy

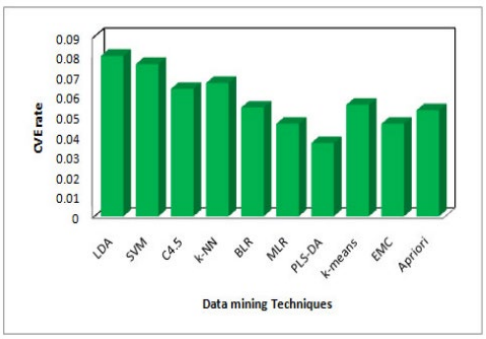

Figure 2. Computing Time

5. Conclusion

Tuberculosis, being a fatal disease though needed to be diagnosed in early stages for complete recovery but it is a difficult task leading to large number of deaths. Various Data Mining algorithms are used to unveil the undiscovered aspects of raw information but

each algorithm is perfect in its own way according to the need of situation. The main aim of this paper is to study six data mining algorithms for prediction of Tuberculosis which may act as supportive tool for diagnosis and prevention of TB as traditional tests are slow and expensive [33]. The Algorithms are studied for accuracy, performance, sensitivity and error rate using inputs such as age, chest pain, chronic cough, blood cough, weight loss, etc. and graphs have been plotted. So, after studying, PLS-DA comes out to be the best one in terms of all the above-mentioned parameters with accuracy of about 96%.Therefore, it is suggested to use PLS-DA algorithm for classification of Tuberculosis in order to get high accuracy and performance.

References:

[1] Bharati, M. &Ramageri,. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering. 1.

[2] Ilievska-Poposka B, Metodieva M, Zakoska M, Vragoterova C, Trajkov D. Latent Tuberculosis Infection - Diagnosis and Treatment. Open Access Maced J Med Sci. 2018; 6(4):651–655. Published 2018 Apr 14. doi:10.3889/oamjms.2018.161

[3] Nicole Fogel, Tuberculosis: A disease without boundaries, Tuberculosis, Volume 95, Issue 5, 2015,Pages 527-531,ISSN 1472-9792,https://doi.org/10.1016/j.tube.2015.05.017.

[4] K.R.Lakshmi, & Krishna, M. Veera & Kumar, s.Prem. (2013). Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability. International Journal of Modern Education and Computer Science. 5. 10.5815/ijmecs.2013.08.02.

[6] Tharwat, Alaa& Gaber, Tarek & Ibrahim, Abdelhameed&Hassanien, Aboul Ella. (2017). Linear discriminant analysis: A detailed tutorial. Ai Communications. 30. 169-190,. 10.3233/AIC-170729.

[7] Jason Brownlee," Linear discriminant analysis for machine learning",Machine learning mastery, April 2016.

[8] Yu Zhang and Dit-Yan Yeung, "Worst case Linear discriminant analysis ", Department of Computer Science and Engineering Hong Kong University of Science and Technology.

[9] Guo, Gongde& Wang, Hui & Bell, David & Bi, Yaxin& Greer, Kieran. (2003). KNN Model-Based Approach in Classification. Lect Notes Comput Sci. 2888. 986-996. 10.1007/978-3-540-39964-3_62.

[10] JingwenSun,WeixingDu,Niancai Shi "A Survey of kNN Algorithm " Information Engineering and Applied Computing (2018).

[11] Cheng, Debo& Zhang, Shichao& Deng, Zhenyun& Zhu, Yonghua&Zong, Ming. (2014). kNN Algorithm with Data-Driven k Value. 499-512. 10.1007/978-3-319-14717-8_39.

[12] SadeghBafandehImandoust And Mohammad Bolandraftar" Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical

Background " S B Imandoust et al. Int. Journal of Engineering Research and Applications www.ijera.com Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610.

[13] ShufengChen "K-Nearest Neighbor Algorithm Optimization in Text Categorization" IOP Conf. Series: Earth and Environmental Science 108 (2018) 052074 doi :10.1088/1755-1315/108/5/052074

[14] Jyoti Yadav, Monika Sharma. "A Review of K - mean Algorithm". International Journal of Engineering Trends and Technology (IJETT). V4 (7):2972-2976 Jul 2013. ISSN: 2231-5381. www.ijettjournal.org. published by seventh sense research group.

[15] John A. Hartigan , Clustering Algorithms, John Wiley & Sons New York , London , Sydney , Toronto,1975

[16] A. Mahendiran, N. Saravanan, N. Venkata Subramanian And N. Sairamm, Implementation Of K-Means Clustering In Cloud Computing Environment, Research Journal Of Applied Sciences, Engineering And Technology 4(10): 1391-1394, 2012.

[17] Charanjeet Kaur "Association Rule Mining using Apriori Algorithm: A Survey " International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013.

[18] Mohammed Al-Maolegi[1], Bassam Arkok[2] "AN IMPROVED APRIORI ALGORITHM FOR ASSOCIATION RULES" International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, February 2014.

[19] Jiao Yabing "Research of an Improved Apriori Algorithm in Data Mining Association Rules " International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.

[20] Deepali Bhende[1] , Usha kosarker[2] , Mnisha Gedam[3] "Study of various Improved Apriori Algorithms"IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 55-58 www.iosrjournals.org.

[21] K.R.Lakshmi[1] ,M.Veera Krishna [2] ,S.Prem Kumar[3] "Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability" I.J.Modern Education and Computer Science, 2013, 8, 8-17 Published Online October 2013 in MECS (http://www.mecs-press.org/)I.J.Modern Education and Computer Science, 2013, 8, 8-17 Published Online October 2013 in MECS (http://www.mecs-press.org/)

[22] Peng, Chao-Ying Joanne and Nichols, Rebecca Naegle (2003) "Using Multinomial Logistic Models To Predict Adolescent BehavioralRisk,"Journal of Modern Applied Statistical Methods: Vol. 2 :Iss. 1 , Article 16.DOI: 10.22237/jmasm/1051748160

[23] Arı, Erkan. (2016). Using Multinomial Logistic Regression to Examine the Relationship Between Children's Work Status and Demographic Characteristics. Research Journal of Politics, Economics and Management. 4. 77-93.

[24] Anass BAYAGA," MULTINOMIAL LOGISTIC REGRESSION: USAGE AND APPLICATION IN RISK ANALYSIS ",School of Initial Teacher Education (SITE), Faculty of Education, University of Fort Hare, South Africa,2010.

[25] R.C. Neath, M.S. Johnson, " MULTINOMIAL LOGISTIC REGRESSION", in International Encyclopedia of Education (Third Edition), 2010.

[26] Madhu B, Ashok N C and S Balasubramanian, ” A Multinomial Logistic Regression Analysis to Study The Influence Of Residence And Socio-Economic Status On Breast Cancer Incidences In Southern Karnataka”, JSS University, Mysore – 570015 India Volume 2,Issue 5, May 2014.

[27] Fredua, Benjamin, "Multinomial Logistic Regression Analysis Of Varicella Vaccination - 2011 National Immunization Survey (NIS) –Teen Survey Data." Thesis, Georgia State University, 2015.

[28] El-Habil, Abdalla,” An Application on Multinomial Logistic Regression Model". Pak.j.stat.oper.res.. 8. 10.18187/pjsor.v8i2.234.,2012.

[29] Ballabio, Davide1 &Consonni, Viviana2. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. Analytical methods. 5. 3790-3798. 10.1039/c3ay40582f.

[30] Mendez, K.M., Reinke, S.N. & Broadhurst, D.I. Metabolomics (2019) 15: 150. https://doi.org/10.1007/s11306-019-1612-4

[31] Loong Chuen Lee, Choong-YeunLiong, and Abdul Aziz Jemain,Partial least squares-discriminant analysis (pls-da) for classification of high-dimensional (hd) data: a review ofcontemporary practice strategies and knowledge gaps, Analyst143(2018),3526–3539.

[32] Cramer, R.D. Partial Least Squares (PLS): Its strengths and limitations. Perspectives in Drug Discovery and Design 1, 269–278 (1993) doi:10.1007/BF02174528

[33] Aiyesha Sadiya, Anusha V Illur, Aekhata Nanda, Eshwar Rao, Vidyashree K P, Mansoor Ahmed “Differential Diagnosis of Tuberculosis and Pneumonia using Machine Learning” , International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-6S4, April 2019