Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018

Comparison of Machine Learning Algorithms in Data classification

Ch Anwar ul Hassan, Muhammad Sufyan Khan, Munam Ali Shah
Department of Computer Science.
COMSATS Institute of Information technology Islamabad, Pakistan
anwarchaudary@gmail.com mshah@comsats.edu.pk

***Abstract*—Data Mining is used to extract the valuable information from raw data. The task of data mining is to utilize the historical data to discover hidden patterns that helpful for future decisions. To analyze the data machine learning classifiers are used. Various data mining approaches and machine learning classifiers are applied for prediction of diseases. Where can supports, in timely treatment. The aim of this work is to compare the performance of ML classifier. These ML classifiers are Logistic Regression, Decision Tree, Niven Bayes, k-Nearest Neighbors, Support Vector Machine and Random Forests classifiers on two datasets on the basis of its accuracy, precision and f measure. The experimental results reveal that it's found that the Random Forests performance is better than the other classifiers. It gives 83% accuracy in heart data sets and 85% accuracy in hepatitis disease prediction**

Keywords-Data Mining, Machine Learning, Data Analysis

I. INTRODUCTION

Nowadays, extracting valuable information from the raw data is essential to take the effective business decision. The need of processing and exploring the useful information obtained from raw data has arisen in many fields of life; business, medicines, science, and engineering.

Today's intelligence technologies analyze the data, explore the information and then convert the information into knowledge. At that point, Data Mining (DM), Machine Learning (ML) play a vital role to accurately extract the information from the huge amount of data. Several DM methods exist for prediction these are; "Classification, Clustering, Association rules, Summarizations and Regression" as shown in Fig 1.

Data mining is used to finding out previously unknown, even though potentially useful, hidden patterns from the extensive amount of data. DM is effectively analyzed a large amount of data, complex data that contain multiple variable and nonlinear relations. DM is mainly used to predict outcomes or behaviors according to the future perspective also explore the relationship and associations that are currently not understood.

ML is more effective to explore knowledge, validate the data and their behavior. When data is available, split it into training and test datasets and trained to explore where it stand in future. ML steps are shown in Fig 2. In [1], [2], Author uses DM and ML concepts to extract the valuable information regarding kidney order using kidney disease data sets and provide treatment earlier to the patients. ML and DM are used in almost every practical field, in the medical sector, it is used for diseases predictions [3]. So by applying the ML, we not only analyze the current data also predict the future.

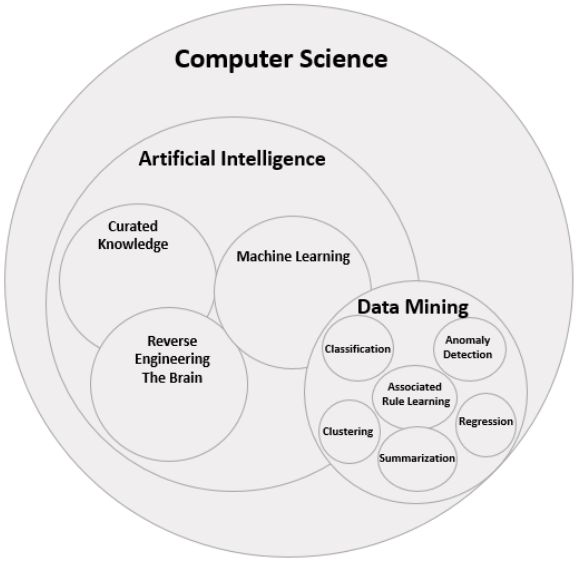

Figure 1. Machine Learning & Data Mining

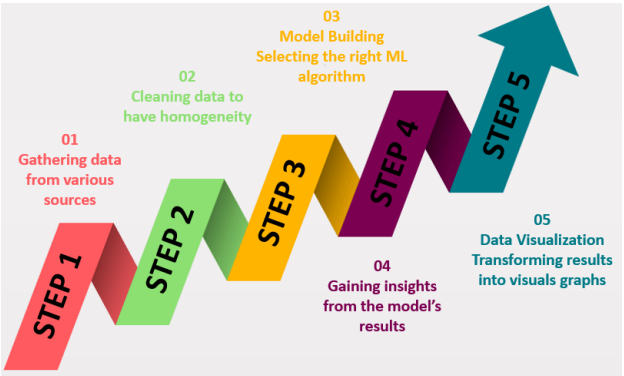

Figure 2. The Machine Learning Process

In this paper, we discuss the performance of ML algorithms and DM approaches on two different datasets. Then compared the performance of ML classifiers in terms of accuracy, precision, recall and F measure on both datasets and elaborate which techniques effectively analyze the data. Consequences describe that in order to evaluate the historical data, careful decide which classifier is applied to achieve an effective result for prediction and future decisions.

Rest of the paper will be structured like that section II contains literature review, methodology describes in Section III, then we conclude experiments and results in section IV and finally, Section V covers conclusion.

II. RELATED WORK

Many researchers have done a lot of work on data analysis, survivability analysis through Machine Learning (ML) and Data Mining (DM) approaches. Several studies reported that these techniques are significant for future prediction such as in the field of medical diagnosis. In these studies, authors applied multiple approaches to specified problems and achieve highly classification accuracies e.g. in the healthcare industry, these techniques are used for disease prediction.

In [1], [2] author applied DT, LL, NB, SVM, KNN, PCA, ICA classifier respectively to analyze the kidney disease data. Early detect and treatment the diseases prevent getting from worst, that is not only difficult to cure also impossible to provide treatment. Breast cancer affects many women, so researchers work on different classifiers such that DT, SMO, BF Tree and IBK to analyze the breast cancer data and examine the performance of these techniques in order to accurately predict the breast cancer using DT [4] and Weka software [5]. RBF Network, Rep Tree, and Simple Logistic DM techniques are used to predict and resolve the survivability of breast cancer patient [6]. Simple Logistic is used for dimension reduction and proposed RBF Network and Rep Tree model used for fast diagnose the other diseases.

The prediction of heart disease and patient survivability is a critical research problem for few decades. Globally, the heart diseases are one of the major cause of deaths. About 80% of deaths in low and middle-income countries are happened due to heart diseases[7]. Researchers use multiple DM techniques to develop a prediction model for the survival of heart disease patients. K-mean, C4.5 techniques are used in [8]. NB, J48 DT and Bagging algorithm, CART, ID3 (Iterative Dichotomized 3) and Decision Table, Logistics Classification, Multilayer Perception and SMO these three algorithms are respectively used in [7], [9], [10] to predict about heart diseases patients and their survivability. However, with the advancement of the technologies, these measurements are not sufficient for the prediction of diseases.

Artificial Neural Networks (ANN) are used in[11], [12] for kidney dialysis and prevent patient form kidney diseases. ANN is used to develop a tool to classify the patient health condition leading to "End Stage of Kidney Disease"[12]. In [11], author also applied DT and LR to analyze the kidney dialysis data and examine the performance of these classification techniques. From the experimental results author observed ANN perform better compared to DT and LR algorithms. Prediction system is developed to control kidney failure and heart disease, using A-Prior and K-Mean Algorithms [13]. They analyzed the statistics by ML and evaluate using calibration and Receiver Operating Characteristic plots.

DM classification methods and ML algorithms are used for the prediction of diseases. Diseases like diabetes, asthma, high/low blood pressure can be predicted by analyzing data using multi-rank novel method [14]. Practitioners also work on Liver, eye and sickle cell diseases by applying BN, DT classifiers and Weka tool respectively to discover their symptoms [15]–[17]. Supervised ML Taxonomy is shown in Fig 3.

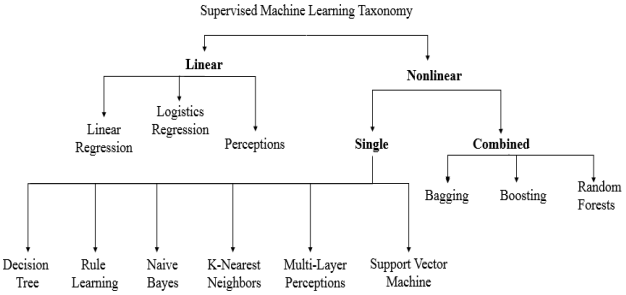

Figure 3. Supervised Machine Learning Taxonomy

ML algorithms NB and SVM are used to predict the kidney disease[18]. The author examined that SVM perform better than NB and in [19] author uses SVM and KNN in order to introduce decision support system to predict kidney diseases. Many of the researchers to predict the kidney disease. However, still it is exploring to find better accuracy to predict kidney disorder problem in this regard NB, ANN, Fuzzy logic and Weka tool are also used to analyze the kidney disease datasets respectively in [20]–[22]. In their results, KNN performs better than SVM. LR and ANN algorithm is also implemented to predict Long-Term Kidney Transplantation Outcome)[23]. The comparison has been performed based on the specificity and sensitivity of LR and ANN in kidney transplant prognosis.

From last decade's researcher done a lot of work on diseases prediction, through ML techniques to predict the e.g. kidney, heart diseases, breast cancer, asthma, eyes problems, sickle cell and liver disease as described in Table 1.

TABLE I. CLASSIFIERS & DISEASE DISCUSSED IN LITERATURE REVIEW

References	Diseases	Classifiers	Year
[1]	Kidney	DT, NB, SVM,LL	2018
[2]	Kidney	KNN, NB, SVM,PCA, ICA	2017
[3]	Multiple	Map Reduce	2016
[4]	Breast Cancer	Classification(DT)	2014
[5]	Breast Cancer	SMO, IBK, BF	2007
[6]	Breast Cancer	Rep Tree, RBF,SL,DS	2017
[7]	Heart	J48DT, NB	2014
[8]	Heart	K-means, C4.5	2014
[9]	Heart	CART, ID3, Decision Table	2013
[10]	Heart	Weka Tool	2013
[11]	Kidney	DT, ANN, LR	2014
[12]	Kidney	ANN	2013
[13]	Heart & Kidney	Apriori,K-means	2014

References	Diseases	Classifiers	Year
		clustering	
[14]	Asthma	Multi ranks	2016
[15]	Liver	Bayesian Network	2014
[16]	Eye	Classification(DT)	2014
[17]	Sickle cell	Weka tool	2014
[18]	Kidney	SVM, NB, GFR	2015
[19]	Kidney	KNN, SVM	2015
[20]	Kidney	NB, ANN	2016
[21]	Kidney	Fuzzy Logic	2014
[22]	Kidney	Weka Tool	2015
[23]	Kidney	ANN, LR	2013

III. METHODOLOGY

The proposed methodology using six classification techniques; Logistic Regression (LR), Decision Tree (DT) ,Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Random Forest (RF) to predict about the heart and hepatitis diseases as the proposed methodology shown in Fig 4. These classifiers are used to improve the prediction. We applied above mention classifiers on heart and hepatitis diseases datasets to predict that how many patients are affected with heart problem and faces hepatitis problem. The performance of these classifier are evaluate on the bases of accuracy, precision, recall and F measure.

The datasets of heart and hepatitis are taken from UCI repository, the classifier taking it as input for disease prediction. These classifiers are implemented in Python language. Python is a powerful interpreter language and a reliable platform for research.

The accuracy of prediction increased by comparing the results of these six classifier using evaluation parameters. The experimental result describes which classifier is best between them.

A. Evaluation Parameters

Some evaluation parameters in data mining are accuracy, precision, recall, and F measure. Where TP- True Positive, TN- True Negative, FP- False Positive and FN- False Negative [19].

- Accuracy is defined as the number of accurately classified instance divided by a total number of instance in the dataset as in (1).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \dots (1)$$

- Precision is the average probability of relevant retrieval as described in (2).

$$Precision = \frac{TP}{TP + FP} \dots (2)$$

- The recall is defined as the average probability of complete retrieval as defined in (3).

$$Recall = \frac{TP}{TP + FN} \dots (3)$$

- F- Measure is the calculated by using both precision and recall as shown in (4).
- $F\ Measure = \frac{2*(Precision*Recall)}{Precision*Recall} \dots (4)$

B. Data Sets

To perform the research, datasets of heart disease and hepatitis are used. Heart disease dataset contains 14 attributes and 303 instances. Whereas hepatitis disease contains 20 attributes and 155 instance [24]. These datasets are taken from UCL repository. It's an online repository that contains 412 diverse datasets. UCI provides data for ML to perform analysis in a different direction. The UCI database is known for its extensiveness in data, its completeness, and accuracy.

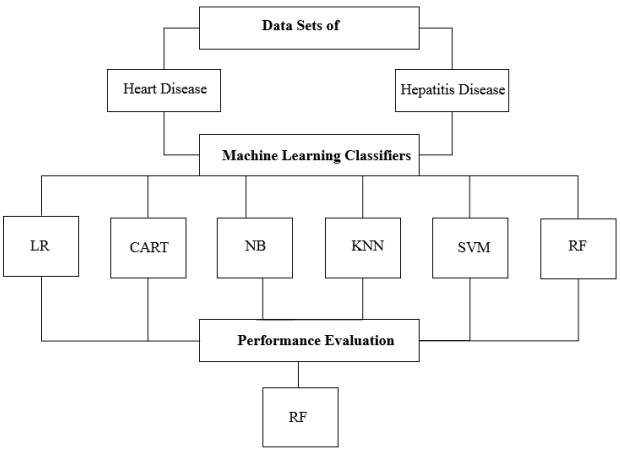

Figure 4. Proposed Methodology

C. Machine Learning Classifiers:

In this research, six classification methods are implemented in python Enthought Canopy. These models are used to improve prediction. These classifiers are compared to find out best that predicts the chance of heart and hepatitis diseases in patients. In next section, we briefly describe these classification techniques/ classifiers.

1) Logistics Regression (LR): is the predictive analysis to conduct on discrete (binary) values based on a specified set of independent variables. LR describes the data and clarifies the relationship between one (binary) dependent variable and independent variables. It predicts the event occurrence probability by fitting the data into a logit function. Therefore, it is also called logit regression [25].

Input values x are linearly combined using coefficient values b, to calculate an output value p. The output values as predictable lies between 0 and 1. Input data associated with coefficient b (constant value) learned from training data. Where p is the output, $b0$ is an intercept term and $b1$ is the coefficient of input value x as shown in (5).

$$P = \frac{e^{(b0+b1*x)}}{1 + e^{(b0+b1*x)}} \dots (5)$$

2) Decision Tree (DT): is a decision supporting tool that represented by a tree-like a graph or model of decision that represent the possible outcomes. It is an approach, that contains conditional statements in an algorithm. DT commonly used in research operations or for classification problems. It help to identify a strategy, particularly for decision analysis to reach the goal more effectively[25].

DT a one of the popular classifier in machine learning. In DT we split the data into at least two homogeneous sets. Then perform decision analysis on that data. This is done on the bases of independent variables or significant attributes to make distinctive groups. Typically, this classifier is referred as "decision trees", however on some platforms or stages like R they are denoted by another term called Classification and Regression Tree (CART).

3) Naïve Bayes (NB): is probabilistic classifiers based on Bayes theorem with naïve independence assumption between the predictors or features. NB classifier assumes, that the existence of a particular feature is not related to the existence of any other feature in a class [25]. For example; apples is consideres a fruit, if it is red and round. Even if these are features related to each other or depend upon the presence of other features. NB classifier consider all these features to contribute independently to probability identifying that the fruit is an apple.

NB model is easy to construct and particularly valuable for large data sets and a Bayes theorem gives a way to calculate posterior probability *P(c|x)* from likelihood (predictor probability) *P(x|c),* class prior probability *P(c)* and predictor prior probability *P(x)* as shown in (6).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \dots (6)$$

4) K-Nearest Neighbours (KNN): is used for regression and classification problems. KNN is commonly used for classification problems. KNN classifier store all the existing cases then classified new cases by the majority votes of its neighbors. The case is assigned to that class, which is most common to its k nearest neighbors, measured by distance function. These distance functions are Euclidean *Eu*, Manhattan *Ma* and Minkowski *Mi* calculated using (7), (8) and (9) respectively.

$$Eu = \sqrt{\sum_{k=1}^{n} (pk - qk)^2} \dots (7)$$

$$Ma = \sum_{k=1}^{n} |pk - qk| \dots (8)$$

$$Mi = \left(\sum_{k=1}^{n} |pk - qk|^r \right)^{1/r} \dots (9)$$

Where as *r* is the parameter, *n* is the number of attributes or dimensions. *pk* and *qk* are respectively, the kth element of objects p and q [25].

5) Support Vectors Machine (SVM): is a classification and regression algorithm. In SVM, every data item is plotted in n-dimensional space, a number of dimensions are equivalent to the number of features or attributes. Where *n* represents the number of attributes. The value of each attribute being the value of certain coordinate. Once plotting all the data items then performed classification by drawing a line or by finding the optimal hyperplane that separates two classes completely.

For example, if we have two features of individual like Hair and Height length. First, we plot these two features in two-dimensional space. Where every point has two co-ordinates (these co-ordinates are also known as Support Vectors[25].

6) Random Forests (RF): are ensemble learning technique for regression, classification and for other tasks. That operate by making a multitude of DT at training stint and outputting that is the mean prediction (regression) or mode of classes (classification) of the distinct trees. Random decision forests groups of decision trees (so known as "Forest") overfitting practice of their training set.

Every tree in the forest contributes for a classification. To classify new case based on its attributes. We identify the tree "votes" for that class. So, the forest indicates the classification of the case that is taking the most votes [25].

Every tree is planted and grownup as follows:

- If the number of objects N in training set, the sample of N objects is taken randomly with replacement. This sample act as a training set for growing tree.
- If there is input variable M. A number m<M is stated that at each node, randomly selection of m variable out of input variable M. So, the best splitting on this *m* is used to split the node. The value of m (node splitter) is constant during growing the forest.
- Each tree is growning up to the largest magnitude possible. So, there is no trimming.

IV. Experimental Results

The experiments are conducted for the prediction of heart and hepatitis diseases by applying various machine learning classifiers. From the experiment results, we identify that Random Forest classifier perform better as compared to other five ML classifiers in the prediction of these diseases.

The below Fig 5, 6, 7 and 8 showing performance of various evaluation parameters in the prediction of heart disease. The experimental results show the comparison of LR, DT, NB, KNN, SVM and RF classifiers and evaluate

the performance on the bases of accuracy, precision, recall and F measure. In all classifiers RF perform better, its accuracy is 83% after that SVM perform betters it holds 82% accuracy.

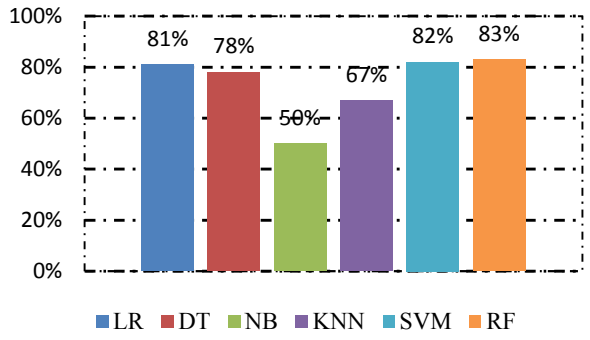

Figure 5. Heart Disease Accuracy

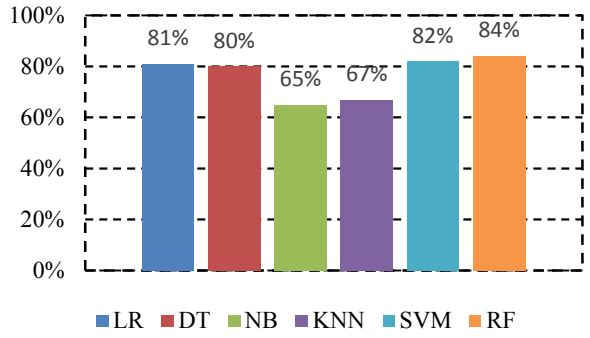

Figure 6. Heart Disease Precision

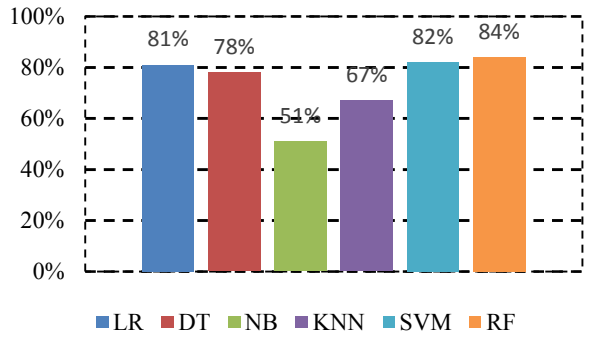

Figure 7. Heart Disease Recall

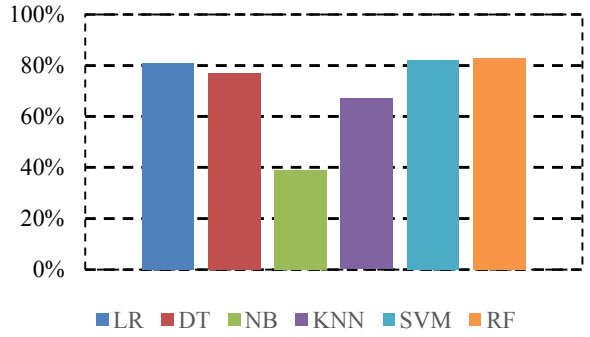

Figure 8. Heart Disease F Measure

The Fig 9, 10, 11 and 12 presenting the prediction of hepatitis diseases through accuracy, precision, recall and F measure evaluation parameters. The experimental outcomes reveal the comparison these six classifiers; LR, DT, NB, KNN, SVM and RF and evaluate that RF also performs better in hepatitis prediction, its accuracy is 85% after that LR it holds 82% accuracy.

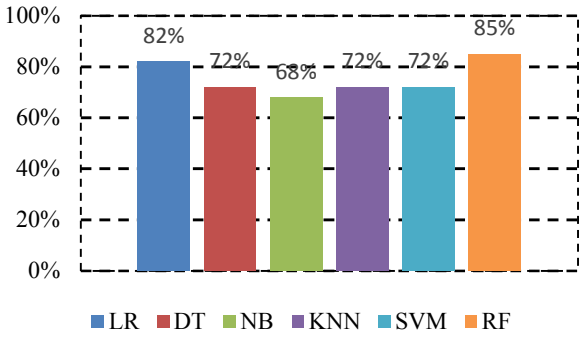

Figure 9. Hepatitis Disease Accuracy

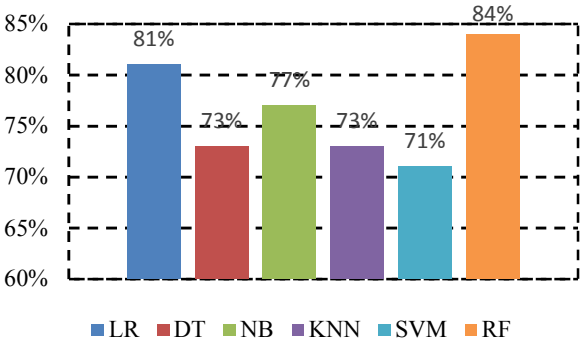

Figure 10. Hepatitis Disease Precision

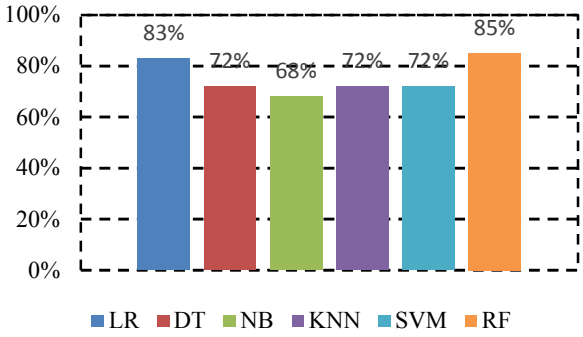

Figure 11. Hepatitis Disease Recall

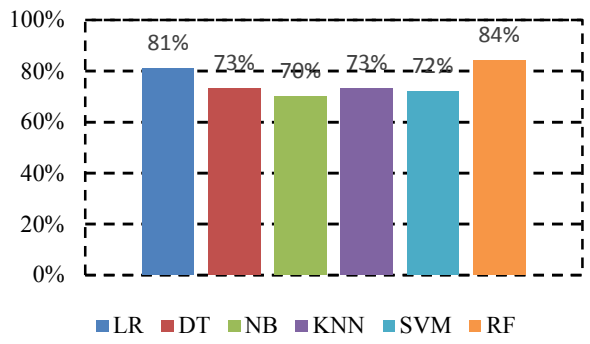

Figure 12. Hepatitis Disease F Measure

V. CONCLUSION

The necessity of extracting the valuable information from raw data has arisen in many fields of life like medical area, business areas etc. As we have already seen the applications of ML and DM in medical area for diseases prediction and in business area for effective business decisions. In this paper, we performed the comparison analysis of ML classifiers for the prediction of heart and hepatitis diseases. Heart and hepatitis diseases is predicted using six different classifiers. Experimental result shows that different classifiers behave differently on the same dataset. From the analysis, we observed that out of six classifiers RF performed better than the all others on both datasets for diseases prediction.

REFERENCES

[1] M. V. Garg Rajni, "A Comparative Study of Different Classification Algorithms on Kidney Disease Prediction," Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. II, pp. 1–6, 2018.

[2] S. A. Kaur Guneet, "Predict Chronic Kidney Disease using Data Mining Algorithms in Hadoop," international J. Adv. Comput. Eng. Netw. , vol. 5, no. 6, pp. 1–5, 2017.

[3] P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu, "Analyzing Healthcare Big Data With Prediction for Future Health Condition," IEEE Access, vol. 4, pp. 9786–9799, 2016.

[4] M. Jahanvi Joshi RinalDoshiDr Jigar Patel, "Diagnosis And Prognosis Breast Cancer Using Classification Rules," Int. J. Eng. Res. Gen. Sci., vol. 2, no. 6.

[5] V. Chaurasia and S. Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques," Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ., vol. 3297, no. 1, 2007.

[6] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability." 29-Jun-2017.

[7] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Diseases." 09-Jan-2014.

[8] R. Chadha and S. Mayank, "Prediction of heart disease using data mining techniques," CSI Trans. ICT, vol. 4, no. 2–4, pp. 193–198, Dec. 2016.

[9] V. Chaurasia and S. Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques." 2013.

[10] S. Vijiyarani, S. Sudha, and M. P. Research Scholar, "An Efficient Classification Tree Technique for Heart Disease Prediction," 2013

[11] K. R. Lakshmi, Y. Nagesh, and M. Veerakrishna, "Performance Comparison Of Three Data Mining Techniques For Predicting Kidney Dialysis Survivability," Int. J. Adv. Eng. Technol., Vol. 7, No. 1, Pp. 242–254, 2014.

[12] T. Di Noia et al., "An end-stage kidney disease predictor based on an artificial neural networks ensemble," Expert Syst. Appl., vol. 40, no. 11, pp. 4438–4445, Sep. 2013.

[13] A. Chaudhary, P. Garg, and A. Case, "Detecting and Diagnosing a Disease by Patient Monitoring System," vol. 2, no. 6, pp. 493–499, 2014.

[14] M. Archana Bakare and A. Professor, "Prediction of Diseases using Big Data Analysis," Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ., vol. 3297, no. 4, 2007.

[15] S.Dhamodharan, "Liver Disease Prediction Using Bayesian Classification," in 4th National Conference on Advanced Computing, Applications & Technologies, 2014, pp. 1–3.

[16] Y. Agarwal and H. M. Pandey, "Performance evaluation of different techniques in the context of data mining- A case of an eye disease," in 2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence), 2014, pp. 72–76.

[17] A. V. Solanki, "Data Mining Techniques Using WEKA classification for Sickle Cell Disease."

[18] S. Vijayarani, S. Dhayanand, and M. P. Research Scholar, "Data Mining Classification Algorithms For Kidney Disease Prediction," Int. J. Cybern. Informatics, vol. 4, no. 4, 2015.

[19] P. Sinha and P. Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM."

[20] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," in 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016, pp. 300–305.

[21] S. Ahmed, M. Tanzir Kabir, N. Tanzeem Mahmood, and R. M. Rahman, "Diagnosis of kidney disease using the fuzzy expert system," in The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), 2014, pp. 1–8.

[22] L. Jena, N. K.-I. J. of E. R. in, and undefined 2015, "Distributed data mining classification algorithms for prediction of chronic-kidney-disease," ermt.net.

[23] G. Caocci, R. Baccoli, R. Littera, S. Orru, C. Carcassi, and G. La, "Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long-Term Kidney Transplantation Outcome," in Artificial Neural Networks - Architectures and Applications, InTech, 2013.

[24] "Index of /ml/machine-learning-databases." [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/. [Accessed: 06-May-2018].

[25] "Essentials of Machine Learning Algorithms (with Python and R Codes)." [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/. [Accessed: 06-May-2018].