Comparative Analysis of Different Machine Learning Classifiers for the Prediction of Chronic Diseases

Kirti Gupta, Pardeep Kumar, and Shuchita Upadhyaya

Kurukshetra University, Kurukshetra, kirtigupta811@gmail.com, mittalkuk@gmail.com, shuchita_bhasin@yahoo.com

Abstract.

Chronic Diseases are the most dangerous diseases for humans and have significant effects on human life. Chronic Diseases like heart disease & Diabetes are the main causes of death. Precise diagnosis of these diseases on time is very significant for maintaining a healthy life. A comparative study of different machine learning classifiers for chronic disease prediction viz Heart Disease & Diabetes Disease is done in this paper. This paper forms the basis of understanding the difficulty of the domain and the amount of efficiency achieved by the various methods recently.

Keywords. «Machine Learning (ML), Prediction, Classification, Support Vector Machine (SVM), Decision Tree (DT), Artificial Neural Network (ANN)».

1. INTRODUCTION

The key demand of healthcare organizations tends to provide quality treatment at an affordable rate. The exact diagnosis of patients is required on time for delivering quality services to them. Unwanted and insufficient outcomes may produce poor quality clinical diagnosis and care. Machine learning techniques help in fast decision-making & reducing cost by making use of historical clinical data. Healthcare helps in maintaining healthy life by making use of proper preventive treatment. Automated techniques can be used to optimize the cost, facility, speed, precision, and reliability of this decision-making process.

Chronic diseases are prime reasons for death everywhere. Heart disease adversely affects the health of an enormous population [1]. The principal Cause of Heart Disease is the unhealthy & fast life style of human beings.

Diabetes is a chronic disease that occurs because of awkwardness in the discharge of insulin. Due to this, sugar level of blood remains unsettled. Diabetes is a vital cardiovascular disorder that can adversely affect the whole-body system. Early detection of diabetes helps in maintaining a healthy life [2]. Patients can have weight loss, less vision, infection, frequent urination, etc. There are 3 categories of diabetes viz. Type I, Type II & Gestational diabetes [3]. The State in which human body declines to deliver insulin is termed as type I. It mainly occurs during childhood or adolescence [3]. The state in which human body becomes resistant to insulin is termed as type II. It occurs when insulin is inside, but the human body is not using it correctly. It's more common in adults. Long-term complications that occur because of diabetes are kidney issues, coronary illness, stroke, and tumor. Pregnant women are distressed by Gestational diabetes. Diagnosis of Gestational Diabetes on time is crucial for the safety of both the mother and infant. The main aims of healthcare organizations are to minimize cost, increase the correctness of the result and be more patient-centric [4].

In this paper, some recent researches related to heart and diabetes disease are discussed in section II. Datasets used are discussed in section III. Experimental results are deliberated in section IV & Conclusion is delineated in section V.

2. Literature Review

Early detection & diagnosis of heart diseases help in maintaining a healthy life. Deepika & Seema [5] performed a comparative analysis on "Naive Bayes, Decision tree, Support Vector Machine (SVM) classifiers" to predict heart & diabetes diseases. It was observed for heart disease prediction that SVM gave the highest accuracy rate of 95.5%. For diabetes prediction, the accuracy of naive bayes was 73.5% which was the highest among all compared algorithms. Nikhar & Karandikar [6] compared Naive Bayes algorithm & decision tree algorithm to find out which of them is more suitable for heart disease prediction. From observation, they originate that the DT algorithm is better than NB. Feshki & Shinjani [7] used Particle Swarm Optimization & Neural Network Feed Forward BackPropagation method for anticipating affected & non-affected patients. Their main focus was on reducing the cost by using feature selection with variable ranking. They used 8 features namely Gender, Age, Blood Pressure, Cholesterol, FBS, Exercise Features (Old peak, Slope, Ex_ang) in their study. Thomas et al. [8] used ANN & ID3 algorithm in their study. Firstly, they applied KNN algorithm to classify age then they used ID3 algorithm for heart disease prediction. They found Smoking & history of heart disease as two important factors for heart disease prediction. Shah et al. [9] increased the accuracy of the SVM technique for heart disease prediction. They used RBF based SVM on three datasets cleveland, Switzerland, and Hungarian of UCI repository. The proposed technique gave 82.2%, 85.8%, and 91.3% accuracy for <<Cleveland, Hungarian, and Switzerland datasets respectively>>. Mohan et al. [10] proposed a hybrid method HRFLM for the prediction of cardiovascular disease. They joined the features of Random Forest & linear method in HRFLM. They used 13 features of UCI machine learning heart disease dataset. They performed their prediction on the R studio. The proposed model gave an 88.7% accuracy for heart disease prediction.

For diabetes prediction, Various researchers proposed various machine learning models in their study. Nirmala et al. [11] proposed an Amalgam KNN model- a combination of KNN and k-mean clustering for diabetes prediction. They used 10-fold cross-validation with different k values using WEKA Software tool. The proposed hybrid algorithm gave higher performance than simple KNN & k-mean clustering algorithms. Sanakal & Jayakumari [12] compared fuzzy C-Mean clustering with the SVM algorithm for diabetes prediction. The accuracy of SVM & FCM was 59.5% & 94.3% respectively. They found fuzzy c-mean clustering as a better tool for the diabetes prediction. Vijayan et al. [13] used decision stump as a base classifier in the AdaBoost algorithm for diabetes prediction. They used a dataset of UCI repository as well as a dataset of Kerala in their study. The proposed algorithm gave an accuracy of 80.72% for the prediction of diabetes. The proposed algorithm gave better results than SVM, Naive Bayes & Decision Tree classifiers. Santhanam & Padmavathi [14] proposed a diabetes prediction model in which they used SVM as a base classifier. For reducing instances & for removing noisy data, they used K-mean clustering. For attribute selection, they used a Genetic algorithm. The proposed model took advantage of both supervised (SVM) & unsupervised algorithm (K-mean clustering). Anand et al. [15] developed a GUI-based diabetes prediction model based on the present lifestyle of people. For conducting the research, questionnaires were prepared with the help of doctors. The attributes included in their questionnaire were Eating Habit (Roadside Eating, Junk food), Sleeping Duration, Sugar Intake, Exercise Duration, Blood Pressure, BMI, Heredity Diabetes, Gender, Belly Size, etc.

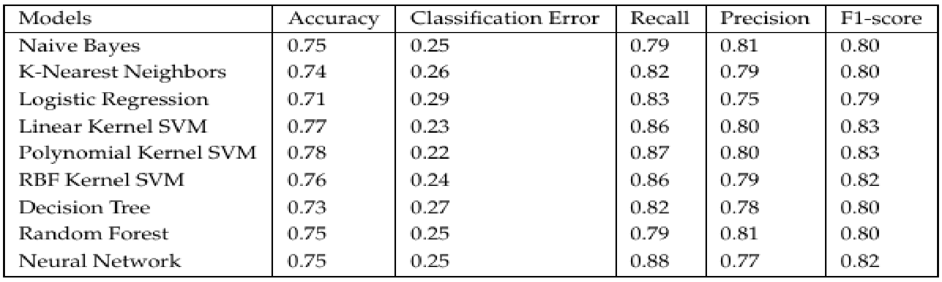
3. DATASETS AND METHODOLOGIES

For Comparing different classifiers on heart disease, two heart disease datasets are used in this paper. One is available on Kaggle website consisting of 1190 records having 12 features of patients from US, UK, Switzerland and Hungary & other is available on “UCI machine learning repository"[16] consisting of 303 records having 14 attributes from Cleveland clinic foundation. For diabetic prediction, "Pima Indian diabetes dataset" from Kaggle website is used in this paper. The dataset contains 768 records of people. "For the diabetes disease prediction dataset of Tabriz, Iran is also used by researchers"[16].

Table 1: Datasets Description

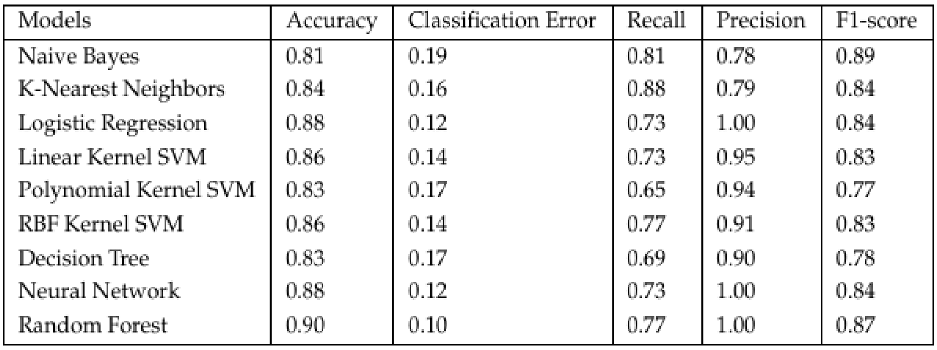
Dataset	No. of Records	Attributes	Dataset Source
HD Dataset1	303	14	UCI ML repository
HD Dataset 2	1190	12	Kaggle website
Diabetes dataset	768	9	Kaggle website

Table 2: Results on Diabetes Dataset



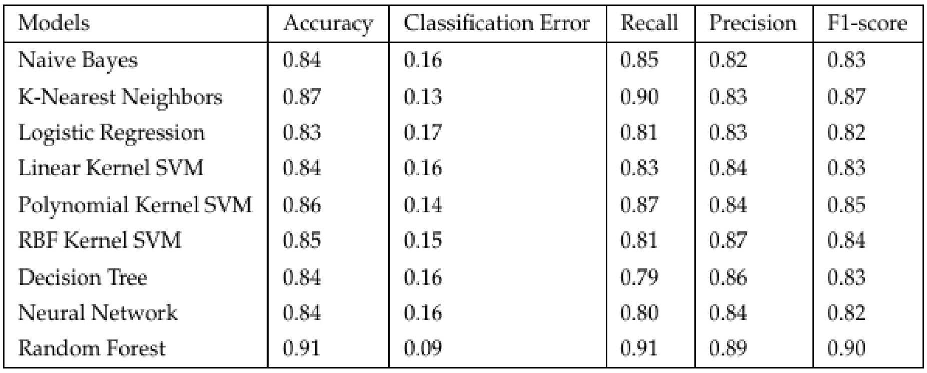
Models	Accuracy	Classification Error	Recall	Precision	F1-score
Naive Bayes	0.75	0.25	0.79	0.81	0.80
K-Nearest Neighbors	0.74	0.26	0.82	0.79	0.80
Logistic Regression	0.71	0.29	0.83	0.75	0.79
Linear Kernel SVM	0.77	0.23	0.86	0.80	0.83
Polynomial Kernel SVM	0.78	0.22	0.87	0.80	0.83
RBF Kernel SVM	0.76	0.24	0.86	0.79	0.82
Decision Tree	0.73	0.27	0.82	0.78	0.80
Random Forest	0.75	0.25	0.79	0.81	0.80
Neural Network	0.75	0.25	0.88	0.77	0.82

Table 3: Results on Heart Disease Dataset 1



Models	Accuracy	Classification Error	Recall	Precision	F1-score
Naive Bayes	0.81	0.19	0.81	0.78	0.89
K-Nearest Neighbors	0.84	0.16	0.88	0.79	0.84
Logistic Regression	0.88	0.12	0.73	1.00	0.84
Linear Kernel SVM	0.86	0.14	0.73	0.95	0.83
Polynomial Kernel SVM	0.83	0.17	0.65	0.94	0.77
RBF Kernel SVM	0.86	0.14	0.77	0.91	0.83
Decision Tree	0.83	0.17	0.69	0.90	0.78
Neural Network	0.88	0.12	0.73	1.00	0.84
Random Forest	0.90	0.10	0.77	1.00	0.87

Table 4: Results on Heart Disease Dataset 2



Models	Accuracy	Classification Error	Recall	Precision	F1-score
Naive Bayes	0.84	0.16	0.85	0.82	0.83
K-Nearest Neighbors	0.87	0.13	0.90	0.83	0.87
Logistic Regression	0.83	0.17	0.81	0.83	0.82
Linear Kernel SVM	0.84	0.16	0.83	0.84	0.83
Polynomial Kernel SVM	0.86	0.14	0.87	0.84	0.85
RBF Kernel SVM	0.85	0.15	0.81	0.87	0.84
Decision Tree	0.84	0.16	0.79	0.86	0.83
Neural Network	0.84	0.16	0.80	0.84	0.82
Random Forest	0.91	0.09	0.91	0.89	0.90

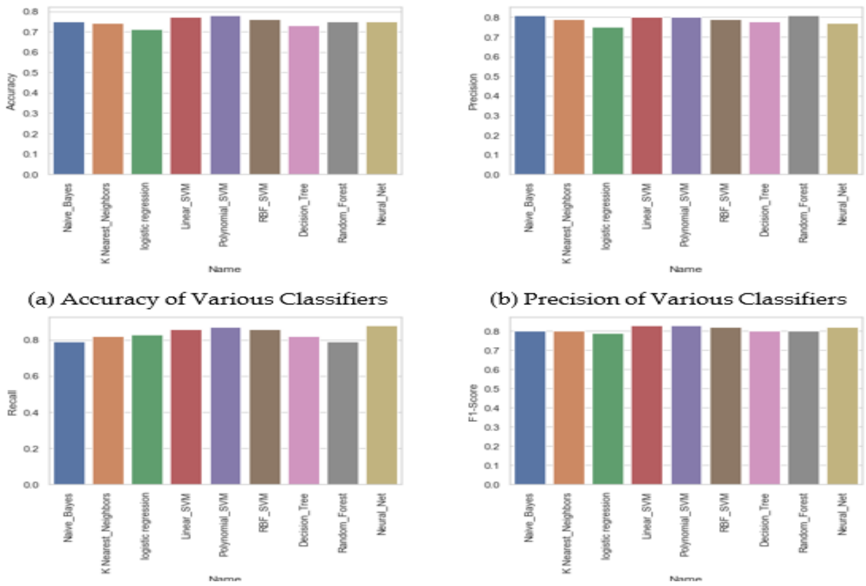
4. EXPERIMENTAL RESULTS AND ANALYSIS

Seven different Classifiers namely "Naive Bayes, KNN, Logistic regression, SVM, Decision tree, Random Forest, Neural Network” are applied on the all mentioned datasets for prediction of diabetes & heart disease. For K-NN Classifier, different values of k are used in order to find the best possible output of the model. Similarly, for Finding the best performance of SVM, different kernels like linear, polynomial & RBF are applied on the datasets. From Observations, it can be concluded that SVM is best among all other specified classifiers for the prediction of diabetes. From Observations, it can also be deduced that Random Forest is best among all other specified classifiers for heart disease prediction.

5. Conclusion & Future Scope

Diagnoses & prediction of chronic disease are the toughest challenges in the pharmaceutical field. It is based on the thorough investigation of clinical data of victims. Because of advancement in machine learning and IT, Feature Selection becomes easy. This also decreases the complexity of the system & increases its accuracy. It is recommended that Feature Selection & Feature Ranking should be done based on Cost & Time so that one should get accurate & precise result in a short time & with minimum cost possible. Researchers have tried several different models and the one with the highest accuracy is picked for the Pre- diction of Chronic Diseases. It has been observed that not only the choice of machine learning technique but also the dataset selected for prediction affects the efficiency of the model.

As prediction deals with uncertainty, machine learning techniques can be combined with fuzzy logic to better deal with uncertainty. Both soft & statistical data can be used in combination to effectively deal with uncertainties in the medical field.



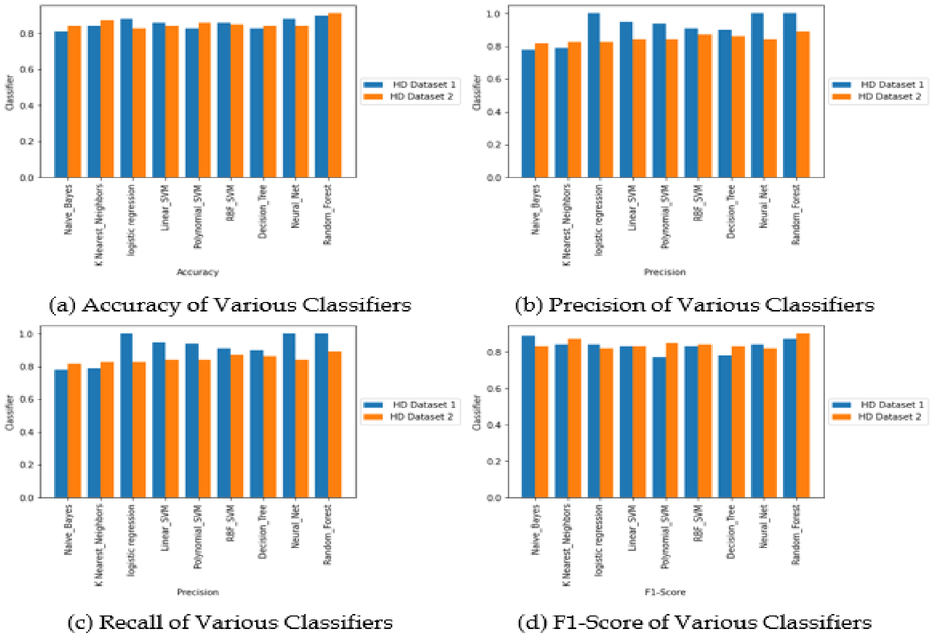
(a) Accuracy of Various Classifiers

(b) Precision of Various Classifiers

(c) Recall of Various Classifiers

(d) F1-Score of Various Classifiers

Figure 1. Performance of Various Classifiers on Diabetes Dataset



(a) Accuracy of Various Classifiers

(b) Precision of Various Classifiers

(c) Recall of Various Classifiers

(d) F1-Score of Various Classifiers

Figure 2. Performance of Various Classifiers on Heart Disease Datasets

6. REFERENCES

[1] T. Karayılan, O. Kılıç, ‘Prediction of heart disease using neural network’, In 2017 International Conference on Computer Science and Engineering (UBMK), pages 719–723. IEEE, 2017.

[2] H. Kaur, V. Kumari, ‘Predictive modelling and analytics for diabetes using a machine learning approach.’, Applied computing and informatics, 2020.

[3] C Kalaiselvi, G. Nasira, ‘A new approach for diagnosis of diabetes and prediction of cancer using anfis’, In 2014 World Congress on Computing and Communication Technologies, pages 188–190. IEEE, 2014.

[4] B Nithya. ‘Study on predictive analytics practices in health care system’, IJETTCS, 5:98–102, 2016.

[5] K. Deepika, S. Seema. ‘Predictive analytics to prevent and control chronic diseases’, In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pages 381–386. IEEE, 2016.

[6] S. Nikhar, A. Karandikar. ‘Prediction of heart disease using machine learning algorithms’, International Journal of Advanced Engineering, Management and Science, 2(6):239484, 2016.

[7] M. Feshki, O. Sojoodi Shijani, ‘Improving the heart disease diagnosis by evolutionary algorithm of pso and feed forward neural network’, In 2016 Artificial Intellience and Robotics (IRANOPEN), pages 48–53. IEEE, 2016.

[8] J. Thomas, R. Theresa Princy, ‘Human heart disease prediction system using data mining techniques’, In 2016 international conference on circuit, power and computing technologies (ICCPCT), pages 1–5. IEEE, 2016.

[9] S. Muhammad et al., ‘Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis’, Physica A: Statistical Mechanics and its Applications, 482:796–807, 2017.

[10] S. Mohan et al. ‘Effective heart disease prediction using hybrid machine learning techniques’, IEEE Access, 7:81542–81554, 2019.

[11] M. NirmalaDevi et al., ‘An amalgam knn to predict diabetes mellitus’, In 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), pages 691–695, 2013.

[12] R. Sanakal, T. Jayakumari, ‘Prognosis of diabetes using data mining approach-fuzzy c means clustering and support vector machine’, International Journal of Computer Trends and Technology, 11(2):94–98, 2014.

[13] V. Vijayan, C. Anjali, ‘Prediction and diagnosis of diabetes mellitus—a machine learning approach’, In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pages 122–127. IEEE, 2015.

[14] T. Santhanam, M.S. Padmavathi, ‘Application of k-means and genetic algorithms for dimension reduction by integrating svm for diabetes diagnosis’, Procedia Computer Science, 47:76 – 83, 2015. Graph Algorithms, High Performance Implementations and Its Applications (ICGHIA 2014).

[15] A. Anand, D. Shakti, ‘Prediction of diabetes based on personal lifestyle indicators’, In 2015 1st International Conference on Next Generation Computing Technologies (NGCT), pages 673–676, 2015.

[16] Weifeng Xu et al., ‘Risk prediction of type ii diabetes based on random forest model’, In 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), pages 382–386. IEEE, 2017.