https://doi.org/10.36785/BUITEMS.JAES.278

Performance Analysis of Student Healthcare Dataset using Classification Algorithm

[a]Muhammad Saqib Javed, [b]Muhammad Aslam
saqibjaved@vu.edu.pk, maslam@uet.edu.pk

a: Virtual University of Pakistan
b: University of Engineering and Technology Lahore

***Abstract*-- Nowadays health is considered as a backbone in terms of performance based on Internet of things (IoT devices), which turned out to be important in diagnosing health level of person with the type of disease a person is suffering with plus its severity level. Basically, IoT sensors operate on medical devices produce large volume of dynamic data. The fluctuation in health data, which forced to use data mining tools and techniques for extracting useful data. Therefore, for applying data mining techniques, heterogeneous data needs to be preprocessed. Therefore, by refining the collection of data, health parametric data mining yields better results with associated benefits. The decision tree is proposed in order to consolidate the health attributes of the students to decide the metrics of health scale. This could lead to evaluate the level of performance of the student in class. After mining the student's health data it is passed to K-Fold cross validation check, so that to determine the accuracy, error rate, precision and recall. The proposed method is considered as an enhanced diagnosis method with fixed patterns for decision tree to make precise decisions. By considering a case study of student's health prediction based on certain attributes with its levels, the diagnostic such as pattern based using K-NN and decision tree algorithm are tested on trained dataset using WEKA tool. At the end, the comparison of different algorithms will be reflected to generalize the introduction of optimized classification algorithm.**

***Keywords*- Support Vector Machine, K- nearest neighbor classifier, Internet of Things (IoT), Student Health Care Monitoring Framework (SHCMF), Weighted nearest neighbor classifier (W-NNC).**

Date Received 10-07-2019
Date Accepted 25-12-2019
Date Published 31-12-12019

I. INTRODUCTION

CONSIDERING health monitoring paradigm, the multi agent-based modeling is a progressively widespread method for analyzing, envisioning and designing multifaceted vibrant systems in public health. There are different agent-based systems already developed, which promises vision into population-based health outcomes. There exists Human Activity Recognition System (HARS), which operates on the basis of Internet of Things (IoT) for monitoring major tasks performed by human in due time [1].

There exists huge amount of information with medical organization and association, where the health data not being efficiently utilized. The previously developed health care system is considered as "Data Rich" with "Knowledge Poor" [2]. The successful analysis method is absent in these systems leads to failure of pure diagnostic of health-related issues. These issues need to be addressed based on data mining algorithm using various knowledge abstraction techniques.

Based on some vigilant consideration of the upper bounds and challenges of already developed and deployed models, it is required to recognize their full potential. We deliver a framework for student health care monitoring, extracted data from different mining algorithms and based on the comparison; conclude the one best suited for the student health prediction.

The health care services are agreed with knowledge-oriented agents or components in order to attain identical value based on decision making for patient treatment. This multifaceted integration of human highlighted composite activities is progressively dependent on intelligent information based on technology [3].

Healthcare requires integrated solution of the problem based on flexibility to deal with different diverse environments. The quality services ensured in the proposed system design will ensure timely response, in case of disturbance based on normal health curve. Human health relies on certain events; one can face in daily life. The proposed system ensures secure health monitoring and effective decision making to call emergency before event occurs. It will also use to improve the work performance of patient being part of any organization by curing the disease timely. The system complexities can only be dealt with methods in such environments to promote system integration and adaptation.

The manipulation of pre-processed data needs to be addressed to trace out the best possible attributes to work out for efficient and relevant results. Formerly implemented framework, computes the criticality of student disease by predicting the likely disease along with its associated levels for measuring health

parameters, which was collected from different IoT and medical devices. The projected methodology is based on the loop holes of the existing system and in order to work on it, a proper technique is devised for effective decision making, in order to call care taker, doctor.

The paper exhibits a system referred to as: Student Health Care Monitoring Framework (SHCMF) built on classification of data using data mining technique, which is basically categorized for providing low cost-effective health care management with respect to offered resources. The student's health parameters had been scanned and based on health attribute, the decision will be made for care taker to call or alarm, in order to communicate the sensitive information of student at study place. The data extracted from classification technique will be mandatory ingredient for decision making and for identifying the level of students studying in class. It can be optimized, if detailed study will be conducted for students in mobility.

Comparison discussed in detail based on classifiers performance levels, such as accuracy, precision, recall, and F1-score in order to choose the best approach for classifying the chosen data set.

II. LITERATURE REVIEW

In recent past, the required knowledge was extracted from targeted application using data mining techniques in healthcare industry.

Divya Tomar *et al* [4] "refer mining for revelation of modern inclinations in healthcare organization, which was helpful in determining other fields of interest. The author explored the expediency of numerous Data Mining techniques, such as regression, classification included in health care paradigm".

Illhoi Yoo *et al* [5] said "data mining contributes in gaining both novel and profound insights to biomedical datasets".

Wu, *et al* [6] proposed that "amalgamation of clinical decision support in relation with automatic patient record, this could reduce medical diagnostic errors, results in enhanced patient safety, decreases unsolicited practice variation along with improved patient outcome".

D. Rajeswara Rao *et al* [7] suggested "appropriate random forest classification for predicting lung cancer disease. Author extracted usefulness of data from different datasets and based on classification and comparison derived the results based on the formulization of one technique, which is random forest, it stands appropriate for diagnosing cancer diseases".

N. Keshan *et al* [8] gives opinion for "diagnostic of Various Diseases for achieving accuracy and precision. The author used classification algorithms: such as k-nearest neighbor, Naive bays, SVM and Neural network. It gives variant results based on variant data; classification algorithm is not fixed for addressing same kind of problems".

Jayanthi Ranjan [9] gives his opinion with "data mining techniques to identify the alternative measures of relief. Curing the disease in simplest way is the art, which was based on some extras associated attributes". Milan Kumari, Sunila Godara [10] gives "comparison of classification techniques as based on, Accuracy, Specificity, Error Rate, Sensitivity, True Positive Rate, and False Positive Rate. The previous research proved that SVM is best in applying and diagnosing disease of critical nature like cardiovascular disease. In this paper, the results showed the accuracy of an algorithm based on provided dataset as best of the selection of the best attributes for dataset".

There exist three different classification algorithms in health data monitoring domain. These are SVM, Naïve Bayes, and Random Forest. Among NN, KNN, Decision tree works better for health diagnostic considering waterborne disease caused by (washing, bathing, drinking infected water). These are the attributes of Decision Making. Prediction of Various Techniques. Diagnostic of Various Diseases for achieving accuracy and precision: Gives variant results based on variant data, classification algorithm is not fixed for addressing same kind of problems. Liyakahunsia [11] gives "reviews about breast cancer data classification operated by IoT device. The classifiers used for this operation are J.48, KNN, Multilayer perceptron, Neural networks etc."

SVM technique is mostly used for extracting physiological data, specifically in medical domain. Support vector machine technique is normally proposed for decision making tasks and anomaly detection in health care services. It is used in diagnostic of chronic diseases like diabetes and heart disease.

The conceptual framework of student interactive healthcare system, which is basically IoT based consists of three phases. "Phase 1 comprise of student's health data, mined from medical devices and sensors. The extracted data is communicated to cloud by the gateway based on Local Processing Unit (LPU)" [12].

Bayesian classifier is functioning depends upon fixed set of events that is why considered as bad estimator.

Weighted nearest neighbor classifier is one of the methods which operate on the weights assignment with training patterns as specified by the class.

K- nearest neighbor loop hole is basically the cost of training and collecting training samples for computation of test data.

Decision Tree is node based for handling numerical data of decision making for placing data categorically. It becomes unstable because it is difficult to implement such complex algorithm.

Naïve Bayes algorithm for health monitoring based on certain parameters is used to improve the classification performance by removing the unrelated options. It has short computational time. Large number of records to manage needs amendment.

Health related dataset domain entails dynamic attributes, which must be exploited to measure student health's along with its status.

The attribute values in association with ("heart rate, respiration rate, blood pressure and body temperature"). The mentioned parameters values are retrieved using various health sensors for defining and declaring student's health status. The sensors deployed on human body for calculating student's health level are: blood oxygen sensor, blood pressure sensor, heart sensor.

Polat et al. [13] analyzed "disease and neuro fuzzy inference for student's diabetic data classification". Deng and kasabov [14] achieves 78.4% classification accuracy in the measurement of quality dataset of health. Yu et al. [15] proposed "a hybrid technique based on combined Quantum Particle Swarm Optimization (QPSO) and Weighted Least Square (WLS) Support Vector Machine, in order to diagnose diabetes".

Smith et al. [16] "proposed a neural network algorithm to build associative models. It comprises of 575 randomly selected data,

used for training and 191 test cases showed an accuracy of 76%". Author applied "C4.5 algorithm for testing immune system accuracy and found classification accuracy as 72.1%". Sahan et al. [17] "used Attribute Weighted Artificial Immune System along with 10- fold cross validation method based on classification accuracy of 75.87%".

Performance forecasting of students depends solely upon the attribute for performing certain task [18]. For achieving more optimal performance, author consider Support Vector Machine with K-Nearest Neighbor algorithm with algorithm for operation, this turns out to be the best technique [19]. For ensuring better performance comparison, Author applied both SVM and KNN on same dataset to predict students' performance level.

III Methodology

The methodology basically revolves around patient monitoring and centralized control center for diagnostic, treatment in timely manner of patient. Consider the health dataset of alleged students, numbered as 182, in order to generate relevant health issues.

The diagnosis scheme based on pattern based is applied on health data set using various classification algorithms, which are then computed for having results for measuring accuracy, sensitivity and response time.

The developed system described two significant concepts first; the contextual information related to various student health issues second, decision making based on student health care dataset for numerous critical circumstances.

There exist numerous tools for the purpose of data mining. In current paper, data mining and the analysis of various algorithms performance is focused by using tool called WEKA.

The complete architecture leads us to the health application for providing patients with centralized solution to have nearest doctor's search on same platform. Communication revolves around patient and doctor and this evolves based on learning to deal with same kind of situation many times.

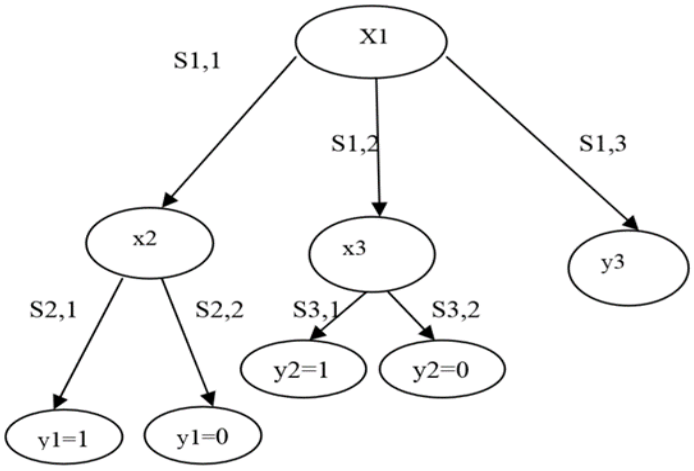

Figure 1. Decision Tree for Health care

The mentioned above decision tree gives clear indication about the diagnostic of disease as X indicates the root node, whereas y indicates the decision leaf's. The S1,2 and S1,3 indicates the level of tree and its decision based on attributes input to it. This is the universal tree formalization and can be mapped to any decision support system or health care system. The indication in the leaf node as Y1=1 or Y1=0 shows that either disease exists based on attributes level or not. 1 means exists and needs treatment either to call care taker or doctor, whereas 0 means normal state or condition of an individual under consideration.

A. Disease Classification using Decision Tree Scenario

The health of a student is classified based on the performance metric of student in class room environment by keeping certain health parameters constant based on the values. J48, C45 applied on the dataset of problem. J48 is the most frequently used techniques for data analysis. It classifies the records of medical field, based on decision at tree levels. A decision tree specifies the sequence of attributes based on nodes.. Let's have the representation of different nodes of decision tree, X'={X1,X2,.....Xk}, shows branches where health level represented by "X" with symptom as "S" and leafs represent decisions Y'={Y1,Y2.....Yk}, directly associated with binary values as XY={0,1} . An exemplary decision tree was presented in the fig1.

B. Data Mining of Data Flow

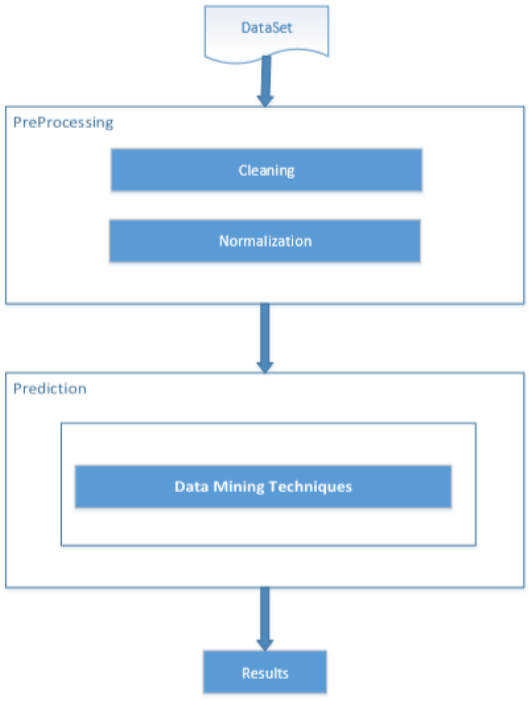

Figure 2. Flow of Data Mining

Fig 2 shows the process of cleaning and normalization of RAW data as received from the samples input. The prediction

technique is then applied for making it refined, so that to predict results correctly.

C. *Knowledge Extraction Steps*

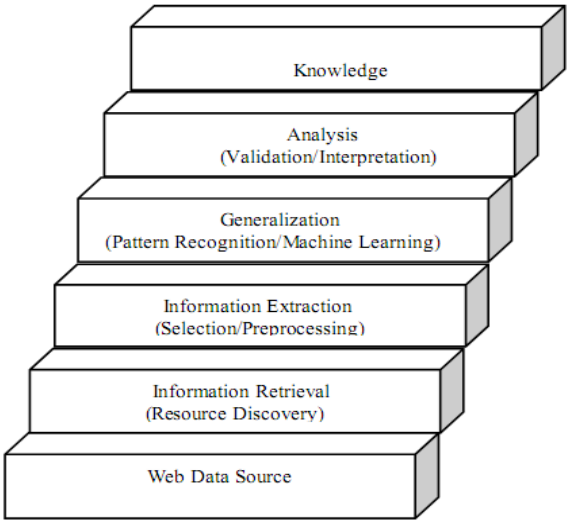

Figure 3. Knowledge Mining Steps

Figure 3 gives clear picture of the knowledge extracted from the dataset based on different information retrieval algorithms.

D. *Architecture of the Proposed System*

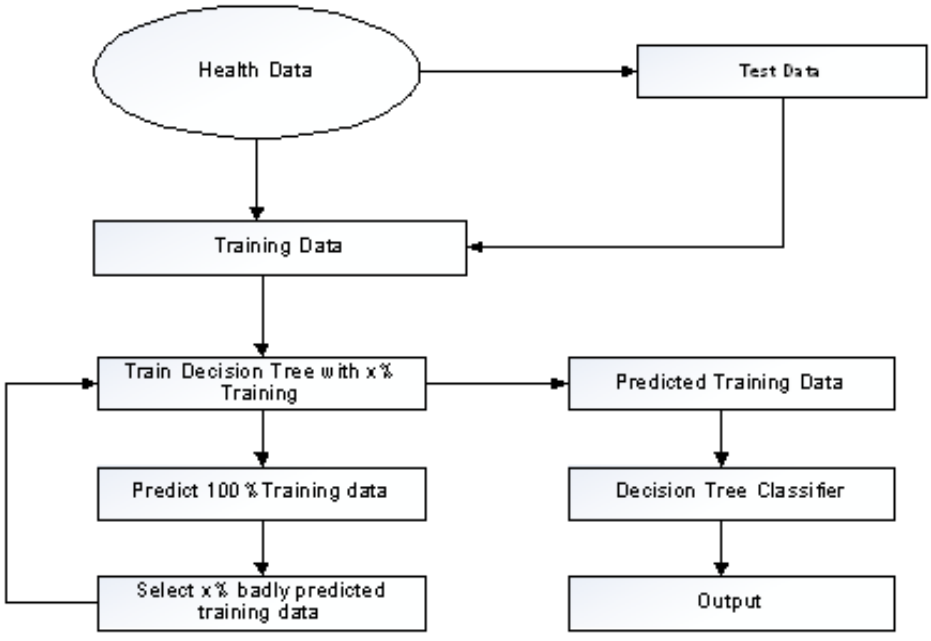

Figure 4. Proposed Health Diagnostic System Flow

Figure 4 gives the proposed layout of the working of the system, right from the Health data as input from different IoT devices toward the training data sample and trained samples tested over test data for validation of same. Decision tree classifier works to classify the data and based on the values of attributes; the system analyzes the need to amend/update the values for having more refined results. Train the data samples then retrain to achieve the level of evolution and learning, which is basically the proposed system.

E. Proposed Classification Algorithm

Proposed Decision Tree for Students Health Care System

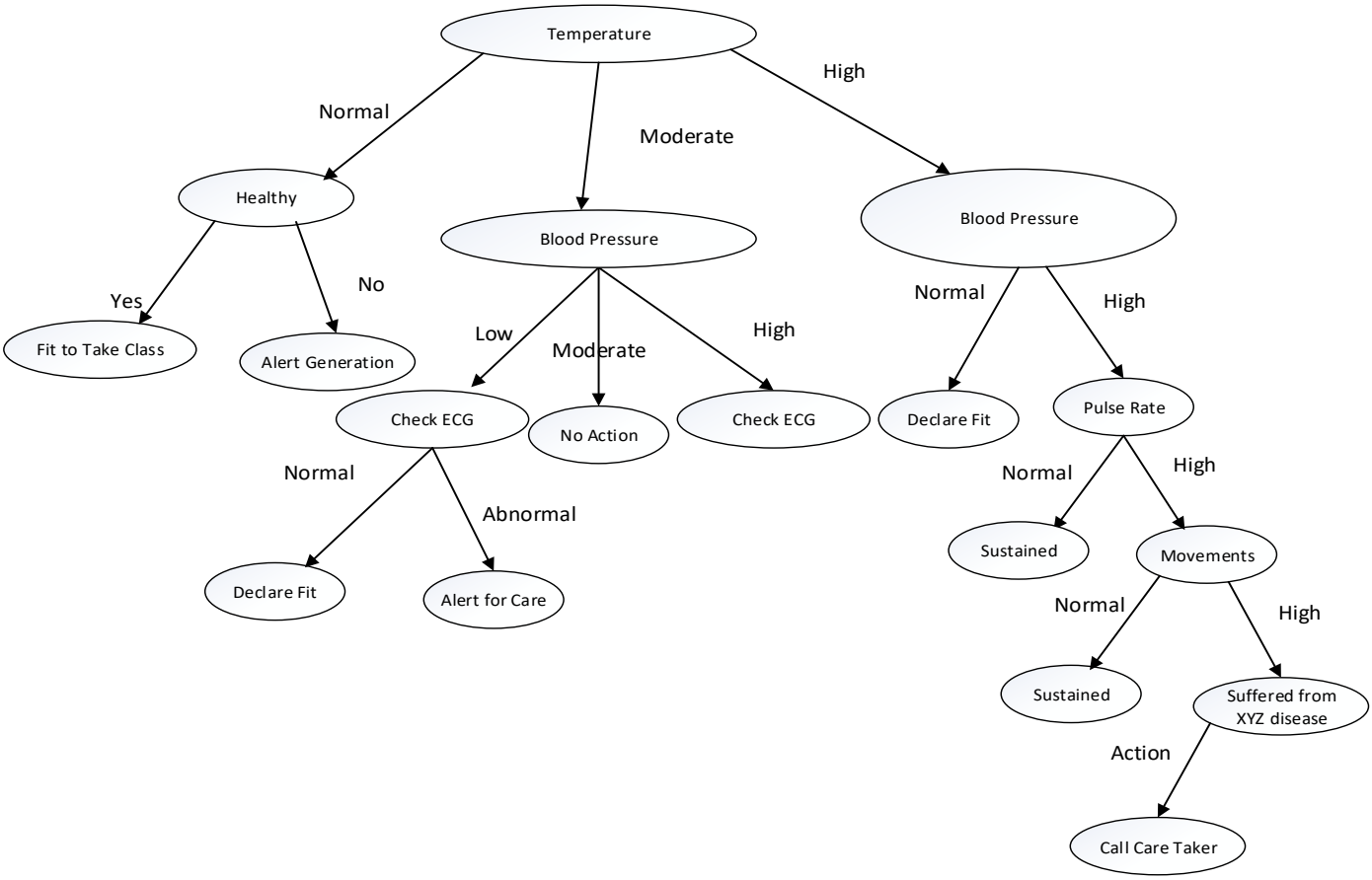

Figure 5. A Comprehensive Decision Tree for Students Health Monitoring

F. Decision Tree Description:

In data mining, decision tree structures are considered as proper way to organize different classification schemes. Classification based on decision tree is performed by routing from the root node traversing till reaching leaf node of the tree in binary format as states represented as “0” or ”1”.

Input: The partition of Data as D, data is basically a composed of training tuples and their related class labels with attribute list based on attribute selection method.

Output: The output of the formalized set of nodes is a complete decision tree.

G. Attributes of Decision Tree

The Attributes for the decision tree of student’s health care system are as under:

- Body Temperature Monitoring
- Blood Pressure Monitoring
- ECG Monitoring
- Heart beat rate monitoring
- Pulse rate monitoring
- Excessive Body Movement Monitoring
- Stress Monitoring = HBR + BP

H. Functions

- If Body Temperature is normal Then declare Normal.
- If Body Temperature is somewhat high check Blood Pressure, if Blood pressure is moderate Then declare Normal.
- If Body Temperature is high grade check blood Pressure, if Blood pressure is high check pulse rate, if pulse rate is high declare Alarm (unhealthy).
- If Blood pressure is high and pulse rate also high and movements are high declared suffered disease.
- If ECG normal Then declare normal.
- If ECG value raised check for BP, HBR, if raised then declare emergency.

I. Decision Tree Pseudo code

1) Choose root node and all nodes for association.
2) Select the appropriate attribute of the dataset and place at root node of the decision tree.
3) Choose proper percentage for train and test set. The sets and subset should contain data with the relevant value of considered attribute.
4) At the end repeat both steps 1, 2 for each node until you traverse tree till its leaf nodes for all the branches of the tree.

IV. RESULT AND DISCUSSION

The summary of the result given by WEKA is as under:

Correctly Classified Instances 206 (77.5799 %)
Incorrectly Classified Instances 63 (23.4201 %)

A. Classification Based on Parameter

The classification of data sample based on input provided by IoT devices for student’s health prediction is mapped on the parameter as mentioned in Fig 7.

Accuracy	$\frac{TP + TN}{P + N}$
Percision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1-Score	$\frac{2TP}{2TP + FP + FN}$

Figure 6. Classification Parameters [4]

B. Classification Accuracy using Decision Tree Using J.48

Using Weka tool, the results derived from the parametric health dataset trained and tested on J.48 classification algorithm is given as total samples tested as 269, where 65% training data and 35% test data with accuracy of 77.5799 %. The result shows better performance as compared to Naïve Bayes, K-NN which gives percentage accuracy of 64% on same dataset.

A. Result Analysis

The results derived from WEKA tool on student's dataset based on parameters appointed is given comprehensibly as:

TABLE 1 WEKA Results on Students Dataset

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Precision/Recall Curve Area**
Class A (Healthy)	0.773	0.250	0.864	0.773	0.816	Tested Negative
Class B (Not Healthy)	0.750	0.227	0.617	0.750	0.677	Tested Positive
Weighted Avg	0.766	0.242	0.783	0.766	0.771	TN + TP

B. Classification Accuracy using Decision Tree Using J.48 Val

TABLE 2 Overall Performance of Proposed Algorithm

Dataset	**Samples**	**Training Data**	**Testing Data**	**Attributes**	**No of Classes**	**Using J.48 Val**
Temperature. BP	269	175	94	8	2	77.5799 %

C. Analysis of Classification Algorithm

The analysis derived on the basis of current results clearly states that Naïve Bayes gives F1 score of 0.6005 with accuracy of 80.11% on same dataset at which different other classification algorithms was tested. The decision tree gives optimum results with 84.33% accuracy and F1 score as 0.6208, hence chosen as the best with that health dataset of student. These results lead to timely decision making in preventing student's health and in progressing the judgment of one's performance.

TABLE 3 Comparison of Classification Algorithms

Classification Algorithms	**Accuracy**	**F1 Score**
Naïve Bayes	80.11%	0.6005
K-Nearest Neighbors	82.56%	0.4924
Decision Tree	84.33%	0.6208
Stochastic Gradient Descent	82.20%	0.5780

The comparative analysis clearly states the facts driven by the decision tree algorithm as compared to Naïve Bayes, K-Nearest Neighbor and Stochastic Gradient accuracy and f1 score varies by some margin.

The result shows varying difference on health dataset of student when trained and tested on different classification algorithm. Decision tree gives better result in terms of accuracy and F1 Score as 84.33% accurate result with F1 score as 0.6208. Similarly, for Naïve Bayes, this is the simplest of all classification algorithms gives values as: precision 80.11% whereas F1 score as 0.6005.

D. Confusion Matrix

The values derived from confusion matrix are given as the ratio of variance from Actual to the Predicted.

TABLE 4 Confusion Matrix

	Predicted		
		Positive	Negative
Actual	Positive	TP	TN
	Negative	FP	FN

TABLE 5 Confusion Matrix for Dataset

	Predicted		
		Positive	Negative
Actual	Positive	140	41
	Negative	22	66

IV. CONCLUSION

The potential health care prediction of student is achieved by the implementation of effective classification algorithm based on pure dataset. This paper investigates different classification algorithm and their results on student's health dataset for improving performance level of sample and scale. The result shows that Decision tree mapped as J.48 in WEKA gives better result in health care dataset of students with different attributes as compared to rest of classification algorithm. It gives 77.5799% of accuracy to the rest of classification algorithm. The attributes identified for measuring student's health is ideally mapped in decision tree paradigm to achieve enhanced results.

V. FUTURE WORK

It will be made better by introducing refinement in decision tree algorithm as introducing more and more attributes could reduce the performance as complexity increased in that case. Pre-processing of same dataset before applying any hybrid approach could leads to more optimized results. The results would be much better if weight assignment procedure is applied at attribute level then apply KNN for effective decision making

VI. REFERENCES

[1] Shirwalkar, N., Gursalkar, S., Tak, T., & Kalshetti, A. (2018). Human Heart Disease Prediction System Using Data Mining Techniques.

[2] Tien J, Goldschmidt-clermont P (2009). Healthcare: a complex service system. *J Syst Sci Syst Eng*, 18(3):257-82.

[3] KARAMI, M., & SHAHMIRZADI, A. H. (2018). Applying Agent-based Technologies in Complex Healthcare Environment. *Iranian Journal of Public Health*, *47*(3), 458-459.

[4] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", (IJCSE) International Journal on Computer Science and Engineering, 2010.

[5] Divya Tomar and Sonali Agarwal(2013), "*A survey on Data Mining approaches for Healthcare".* International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266

[6] IIlhoi, Patricia Alafaireet, et at (2012), **"***Data Mining in Healthcare and Biomedicine: A Survey of the Literature".* Journal of medical sciences Volume 36, Issue 4, pp 2431-2448

[7] Pal, D., Jain, A., Saxena, A., & Agarwal, V. (2016). Comparing Various Classifier Techniques for Efficient Mining of Data. In *Proceedings of the International Congress on Information and Communication Technology* (pp. 191-202). Springer, Singapore.

[8] Keshan, N., Bichindaritz, I., Parimi, P. V., & Phoha, V. V. (2017, August). Temporal Analysis of Stress Classification Using QRS Complex of ECG Signals. In International Symposium on Sensor Networks, Systems and Security (pp. 35-44). Springer, Cham.

[9] Monali Dey and Siddharth Swarup Rautaray, "*Study and Analysis of Data mining Algorithms for Healthcare Decision Support System".* International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, pp 470-477

[10] Jayanthi Ranjan. "*Applications of data mining techniques in pharmaceutical industry"* . Journal of Theoretical and Applied Information Technology (20052007).,pp(61-67)

[11] Kumari, M., & Godara, S. (2011). Review of data mining classification models in cardiovascular disease diagnosis. *International Journal of Computer Science and Technology*, *2*(2), 304-305.

[12] Sareen, S., Sood, S. K., & Gupta, S. K. (2016). IoT-based cloud framework to control Ebola virus outbreak. *Journal of Ambient Intelligence and Humanized Computing*, 1-18.

[13] Verma, P., Sood, S. K., & Kalra, S. (2018). Cloud-centric IoT based student healthcare monitoring framework. *Journal of Ambient Intelligence and Humanized Computing*, *9*(5), 1293-1309.

[14] Polat, Kemal and Salih Gunes, "An expert system approach based on principal component analysis and adaptive neurofuzzy inference system to diagnosis of diabetes disease," Expert system with Applications, pp. 702-710, Elsevier, 2007.

[15] D. Deng and N. Kasabov, " On-line pattern analysis by evolving self-organizing maps", In Proceedings of the fifth biannual conference on artificial neural networks and expert systems (ANNES), 2001, pp. 46-51.

[16] Yue, et al. "An Intelligent Diagnosis to Type 2 Diabetes Based on QPSO Algorithm and WLSSVM," International Symposium on Intelligent Information Technology Application Workshops, IEEE Computer Society, 2008.

[17] Smith, J.W., J. E. Everhart, et al.- "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", Proceedings of the Symposium on Computer Applications and Medical Care (Washington, DC). R.A. Greenes. Los Angeles, CA, IEEE Computer Society Press, 1988, pp. 261-265.

[18] S.Sahan, K.Polat, H. Kodaz, and S. Gunes, "The medical applications of attribute weighted artificial immune system (awais): Diagnosis of heart and diabetes diseases", in ICARIS, 2005, p. 456-468

[19] Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., ... & Olatunji, S. O. (2017, April). Student performance prediction using Support Vector Machine and K-Nearest Neighbor. In *Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on* (pp. 1-4). IEEE".