JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Chronic Disease Prediction Using Probability Weighted Adaboost

Mohammed Imran, Dr.S. Shahar Banu
Research Scholar, Associate Professor
B.S Abdur Rahman Institute of Science and Technology, Vandalur, Chennai-600048

Abstract

Chronic disease prediction is one of the key research in healthcare. Chronic diseases growing as burden to healthcare, because these patients require therapy throughout life. A person can only survive without kidneys for an average time of 18 days. So, it is important to have effective methodology for early prediction of chronic kidney disease. In healthcare field, the accurate prediction plays the major role in finding the risk and level of the disease in the patient. The exact prediction will give lot of benefits like lifesaving, avoiding therapy for entire life and financial costs. This work proposes a Probability Weighted AdaBoost to predict chronic kidney disease based on various attributes. Re-Weight calculation of AdaBoost has been modified in proposed method based on incorrectly classified instances of False Chronic kidney disease, False Non-Chronic kidney disease to compare the performance of a proposed algorithm, existing algorithms like Naïve Bayes, Decision Tree, AdaBoost and Support Vector Machine, are used. Results show that the proposed method outperforms all the prevailing algorithms.

Keywords: Chronic kidney Diseases, Prediction Models, Accuracy, Disease Classification, AdaBoost, Probability Weighted AdaBoost**.**

I. Introduction

Now a days, diagnostics will be done manually All data must be collected manually, and all data must be provided prior to treatment with big data analytics. Because the patient history and treatment details are computerized, a patient can be admitted at any time without prior history.. The recent developments in big data made enormous opportunities for health and medical domains big data deals with four, which are volume of knowledge, the speed of knowledge, sort of data or veracity of knowledge and it's necessary to preprocess the info**.** A constant infection is a human ailment that is a determined or in any case enduring in its belongings. Basic constant sicknesses include joint pain, asthma, malignancy, diabetes and some popular illnesses. Persistent condition may have times of reduction and backslide where the sickness incidentally

disappears or accordingly returns. Accordingly, thinking about singular wellbeing, prescient displaying framework should be created to examine and forestall persistent sicknesses in healthcare management.

Today, "predictive data analytics" is regarded as an efficient and cost-effective way to identify the likelihood of future outcomes supported by historical healthcare big data preprocessing, integration involves data ETL-based algorithms and tools supported current big data platform are predominantly wont to assemble data analysis widely well-planned and sufficient in my work medical big data is primarily utilized in therapeutic data, checking, and timely threatening. Healthcare management is the process of organizing, managing, analyzing public healthcare systems in hospitals networks. Medical care utilization remains a complex problem. A Clinical Information Systems (CISs) have produced opportunities for evocative advances both in patient care and workflow but there's still an extended thanks to perfection. Healthcare providers are still fronting challenges of knowledge exchange, administration, and incorporation because of lack of functionality among these systems. Medical analytics research has reached new frontiers with the advent of more refined intelligent big data analytic techniques.

This research paper is structured in order as Section 2 carries information about Review of Literature, Section 3 carries information about the proposed algorithm of the study, exploratory data analysis and research methodology, Section 4 carries information about result and discussion of this study, and Section 5 is about the conclusion and future research work.

II. Literature Review

Lately, utilizing the enormous information method alongside expanding recurrence to anticipate the chance of infection. Numerous calculations and tool compartments have been made and concentrated by specialists. These have featured the enormous capability of this exploration field. In this segment, a couple of significant works that are firmly identified with the proposed issue are introduced.

In Yang Guo et al. [1] enchanted information from clinical data bases is crucial recalling a definitive goal to create productive accommodating confirmation. Preprocessing was appropriate to renew the concept of information. Classifier was devoted to the changed dataset to develop the Naïve Bayes model [2] eventually frail was accustomed do reenactment, and also the precision of the following model was 72.3%.

Gopalakrishna Palem [3] proposed solutions, for instance, mixed up investigations and vulnerable drug adherence present troubles to particular prosperity and security [3]. These troubles are presently being moderated, if not completely obliterated, with huge data assessment using tweaked drug frameworks, follow-up alerts and consistent investigation checking. Maker uncovers the demonstration of such insightful assessment in clinical consideration section, tending to the thoughts of electronic prosperity records, the huge use of sparks,

trademark language taking care of techniques used in ace decision systems, and so on, presenting positive use-case circumstances significant for each. An extensive part of these thoughts explained in this paper are correct now in powerful use in the business and the maker can be gone after more information on how your affiliation can benefit by them and start changing them.

In Aqueel Ahmed and Shaikh Abdul Hannan [4] discussed coronary ailment assumption. Helpful finding is important yet jumbled undertaking that have to be compelled to be fulfil unequivocally and advantageously. The skilful data assessment instruments are wont to detach obliging info from the monster extent of clinical data there's [2]. a huge data opens inside the clinical benefits frameworks. All things considered, there's an errand of plausible examination gadgets to seek out canvassed associations and models in data. Learning openness and data processing have sighted different applications in business and consistent region space. one among the applications is ailment finding where data processing gadgets are showing practical results. This appraisal paper proposed to get the guts disorders through data processing**,** Support Vector Machine (SVM), Genetic Algorithm, horrendous set speculation, association rules and Neural Networks [2].

Durairaj et al. [5] in dispute Neural Networks to form assumptions on helpful data. Neural Networks are called the Universal markers [2]. DM or fundamentally diabetes is an affliction made owing to the expansion level glucose in human blood. Various standard strategies, brooding about physical and designed tests, are usable for diagnosing diabetes. the unreal Neural Networks (ANNs) based framework can acceptably applied for hypertension risk estimate [2].

Creators in Vapnik [13] considered medical data from more than, 99000 individual experiences distinctive with more than 59000 explicit patients The information was gathered from very 75 million experiences related with 1.7 crore patients. It was assembled throughout a period of ten years, from 1999 to 2008, and contains different credits that contrast with the periods of affirmation and release of diabetic patients [2]. These manuscripts contain information concerning totally unique focus tests and frameworks, assurance, and drugs that were managed inside the length of the recuperating office stay.

III. Proposed Methodology

The principle point of this exploration paper is to investigate and anticipate CHKD information. The goal of this paper is to give a compelling procedure to anticipate the CHKD by utilizing likelihood weighted AdaBoost. Preprocessing includes taking care of and eliminating exceptions, missing worth supplant, information decrease, information change and highlight determination. The proposed philosophy of this examination is appeared in Figure 1.

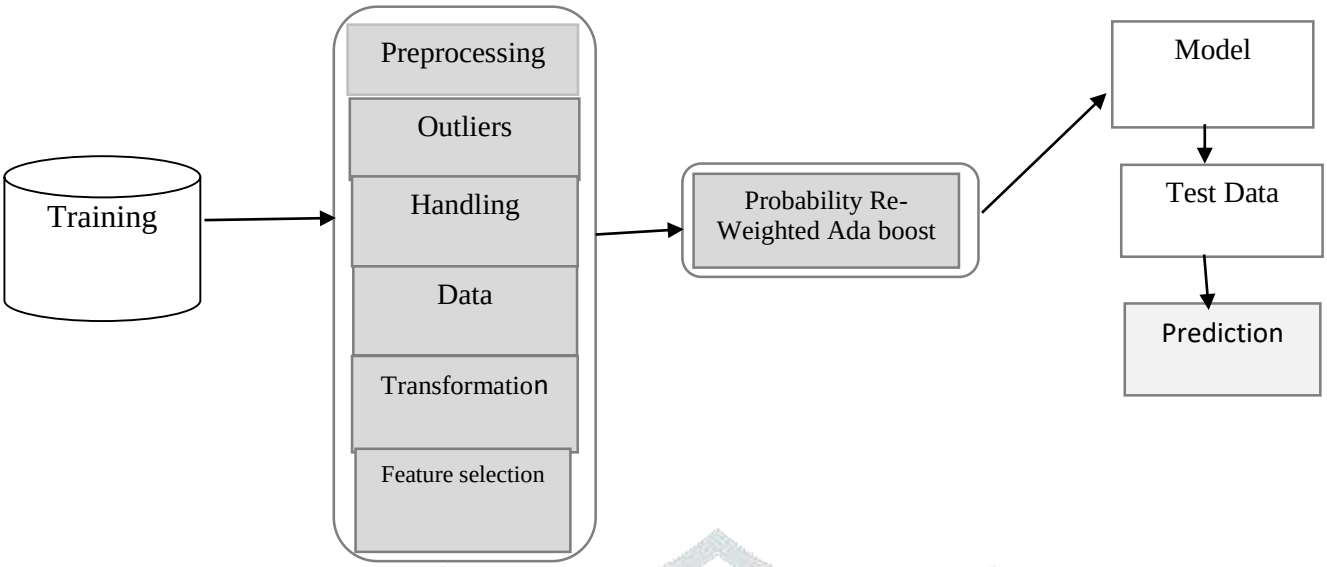

Figure 1: Proposed Methodology

1. Dataset

Chronic illness forecast (CHKD) Dataset has been considered during this examination work. The dataset has been taken UCI AI storehouse. it absolutely was gathered from Apollo Health service, India in 2015 assumed control over a two-month time span. It comprises of 400 perceptions of patients. the knowledge incorporates records of 250 CHKD patients with and records of 150 without CHKD.

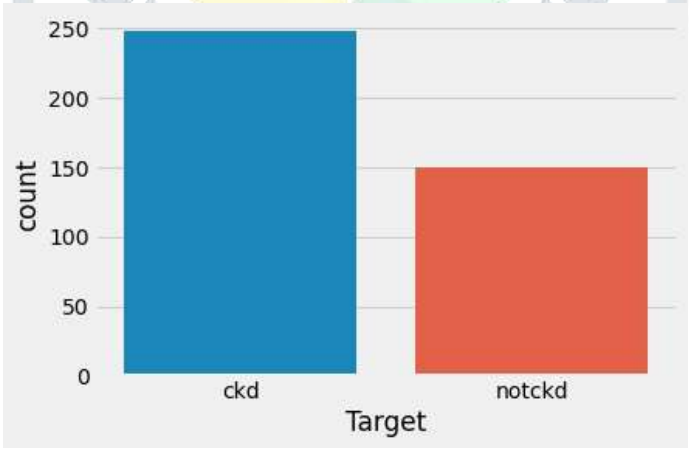

Figure 2: Class Label Distribution

Exploratory Data Analysis: This methodology is utilized to examine informational index and sum up their principle attributes. It augments understanding into an informational index, uncover unique design, selection basic factors, recognize exceptions and irregularities, test essential presumptions and decide ideal factor settings.

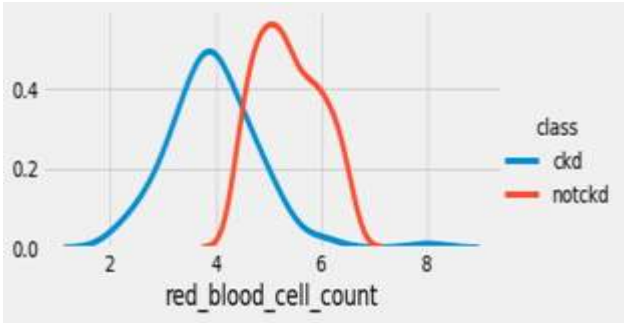

Figure 3: Red Blood Cell Count

Blood tally of non-infection patients is somewhat higher than the blood check of ailing individuals. Fig. 3. addresses red cell blood tally of both sickness and non-illness patients. CKD addresses illness patients and Not CKD addresses non-sickness patients.

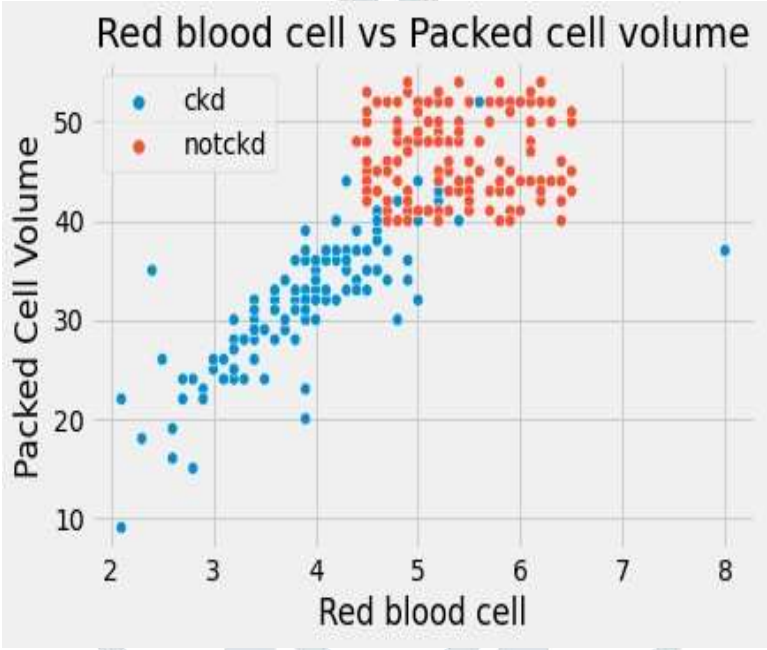

Figure 4: Red Blood Cell vs. Packed Cell Volume

Red platelet and hemoglobin have high connection. Infected individuals are having under 5 red platelet and under 40 stuffed cell volume level. Fig. 4. portrays the connection between Red Blood Cell versus Pressed Cell Volume. Red platelet and hemoglobin have high connection. Infected individuals are having under 5 red platelet and underneath 12.5 hemoglobin level. Figure. 5. portrays the connection between Red Blood Cell versus Hemoglobin.

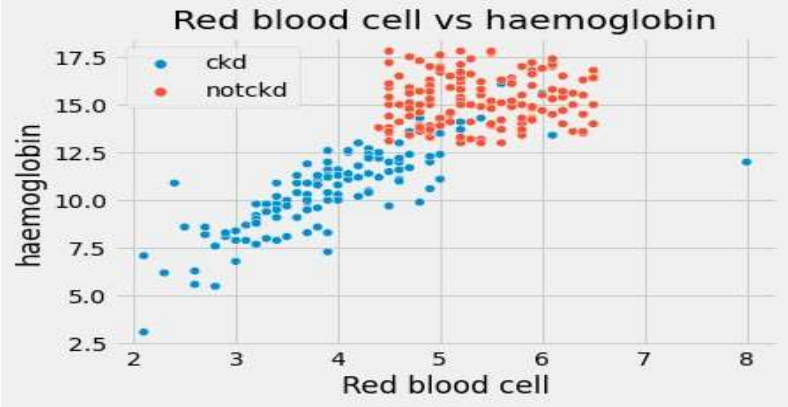

Figure 5: Red Blood Cell vs. Hemoglobin

Outliers: Outliers are out of reach esteems which are situated far away from focal qualities. Every exception should be identified in the CHKD dataset. If there should be an occurrence of potassium and sodium, three limit information focuses are available which are unsuitable. Potassium level with 39 and 47 should be taken care of utilizing exception taking care of system. For sodium, one limit information point was identified, which is 4.5. This likewise should be taken care of utilizing exception dealing with instrument. Z-Score exception recognition instrument utilized in this paper.

Missing Values, in actuality, datasets, missing qualities are shared issue. When all is said in done, all understanding record and subtleties contains not many missing qualities. In any case, CHKD dataset is having around 96% of its factors have missing qualities. There are different rates of missing qualities for all factor. These missing qualities are dealt with utilizing KNN Imputation.

Information Reduction: It is utilized to decrease the quantity of highlights or occasions while keeping a decent logical outcome. For this reason, highlight choice and highlights affiliations or connection have been concentrated to eliminate excess data. Pearson's connection is utilized to consider the connection between factors.

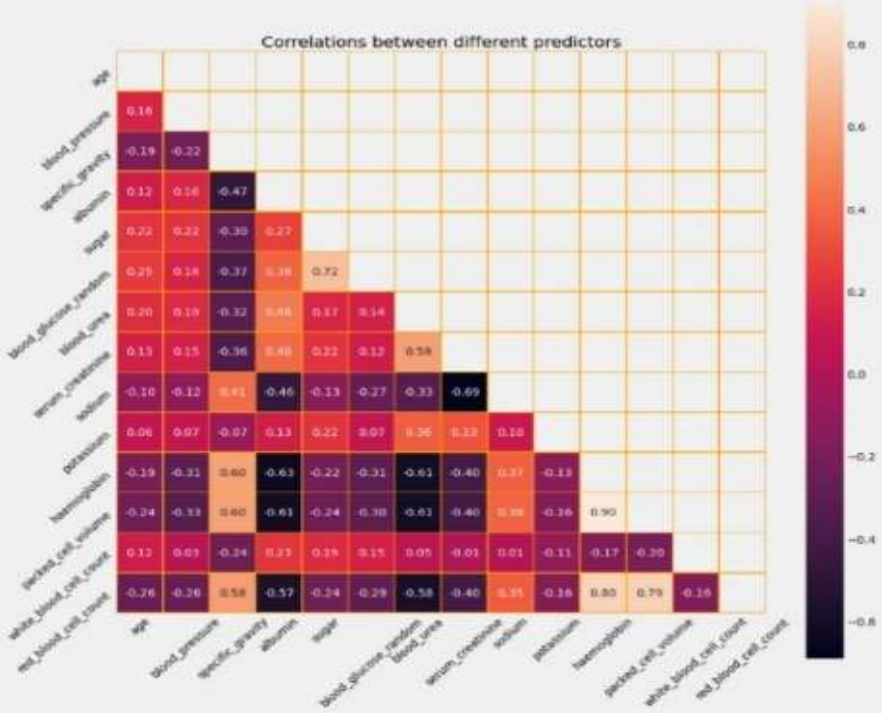

Figure 6: Correlation

Data transformation: In data transformation, data is switch into convenient forms for mining purposes. During this paper, z-score normalization has been applied on fraction data types.

Modeling: Naïve Bayes, Decision Tree, SVM, AdaBoost has been accustomed compare the performance of the classifiers.

2. NAIVE BAYES

The Naive Bayesian system is a smoothing method that aims to expand the classifier: the fashion of naming tricky examples, as a spotlight path of appreciation, somewhere where the category script disappears from a few limited collections. There is no doubt that there is no longer an introverted solution for parallel classifiers, but many estimators who are quite harmless and enthusiastic about and standard Bayes classifiers assume that the reckoning of decided on element is free of the consideration of some other element, given the sophistication adaptable. As an example, an natural product will be see over as an apple at the off hazard that it's far red, round, and round 10 cm in quantification. A guileless Bayes classifier provide for all of these features to enrich freely to the probability that this herbal product is an apple, paying little thoughts to any capacity relativity among the tone, sphericity, and expand the main feature. Certainly, a form of probability representation, innocent Bayesian classifiers are often adequately organized through controlled achievement of attitudes. In several useful functions, the boundary scoring for duplicate Bayesian prototypes is ahead of the method for a series of intense leads; At best, some performance can be achieved with the credulous Bayesian system without Bayesian probability or application of any Bayesian technique. In any case, gullible Bayesian classifiers have admirably worked on a myriad of wonderful and humble events due to their bleak design and irrefutable false assumptions. In 2004, a Bayesian Difficulty Survey confirmed that there was a reliable hypothetical explanation for the adequacy of the susceptible Bayesian Algorithm, the obviously infeasible**.** classifiers. The full correlation

with other feature calculations in 2006 has shown that different theories and concepts are beyond the Bayesian order**,** such as trees or inconsistent forests to help. One advantage of good-faith Bayes is that few people prepare the knowledge to evaluate the key to characterization boundaries [7].

3. DECISION TREE

A selection tree could also be a simple representation for sorting models. For this particle, all the main characteristics of the received information have a limited discrete space and there is a single objective component called "characteristic". Each primitive of the grouped planets is called a category. The selection tree or feature tree can be a tree in which each internal node (not leaf) is understood as information content. The center pie segment named after the information highlight is marked with all possible intent component estimates or contains a subordinate center curve request for equivalent information. Each leaf in the tree is said to have a cross category trend mapping, which means that the tutorial index through the tree falls into the selected category or into a certain probability transition (if the selection tree is built fabulously strong, it is almost specifically tilted) Class Fragment).

A tree is created by dividing the set of sources that form the center of the tree's ideas into subsets that include standing youth. The farewell ceremony is full of enthusiasm for some farewell rules related to the functional sequence Yang Guo, et al. [8]. This interaction is prepared for each specific subset in a highly repetitive manner called recursive allocation. When the subset at the hub has roughly similar estimates of the target variable, or when the goodbye does not increase the desired price, the recursion ends. The Hierarchical Acquisition Chain of Selection Trees (TDIDT) [3] is an example of insatiable computing, and is by far the most popular method of ingesting selection trees from data [10].

In information prospecting, choice trees are often represent additionally because the mixture of fractional and computational procedures to help the description, order and hypothesis of a given arrangement of knowledge [11].

4.SUPPORT VECTOR MACHINE

A assist vector gadget builds a hyperplane or set of hyper planes all through a high-or boundless spatial space, which may be applied for grouping, relapse, or distinctive assignments like exceptions revelation [12]. allegedly, an straight section is achieved with the aid of using the hyperplane that has the maximum crucial distance to the nearest getting ready facts factor of any class (pretended realistic edge), In view of ordinary the larger the sting, the decrease the hypothesis mistake of the classifier.

5. ADABOOST

AdaBoost Ensemble have the advantage of integrating weak classifiers into one best strong classifier. Re-Weights of the classifiers are calculated at each iteration. This is one of the widely used classification algorithm.

Probability Weighted AdaBoost: Standard AdaBoost suffers from overhead issue. In order to solve this issue, this paper uses probability-based re-weighting strategy to get the good classification accuracy without overhead problem.

Based on the overall error, the weight has been given to strong classifier in AdaBoost. If two classifiers have the similar error ratio then same weights allocated to that classifier. In proposed methodology, re-weight given to the classifier considers both positive and negative class labels based on their probability values.

Weight updation of misclassified records is based on PE of False Positive and False Negative.

PE_{FP} (Probability Error for False Positive) = False Positive Rate

PE_{FN} (Probability Error for False Negative) = False Negative Rate

Re-Weight Updating in proposed Adaboost for False Positive Records based on Probability Error for False Positive.

$RFP = \log(\varepsilon t/ (1- \varepsilon t)) * ((1-PE_{FP})/FP)$ ----------→ (1)

Here,

RFP->Re-weight for False Positive

PE_{FP}-> Probability Error for False Positive

Re-Weight Updation in proposed Adaboost for False Negative Records based on Probability Error for False Negative.

$RFN = \log(\varepsilon t/ (1- \varepsilon t)) * ((1-PE_{FN})/FN)$ ----------→ (2)

Here,

RFN->Re-weight for False Negative

PE_{FN}-> Probability Error for False Negative

IV. EXPERIMENTAL RESULTS:

To assess the exhibition of grouped things, the exactness and AUC measures are consolidated. Four cases are reflected as the meaning of classifier [2. TP (True Positive): The degree of evidence that can absolutely be classified in its class. TN (True Negative): the number of preliminaries that were excluded from this class [2]. FP (false positive): the degree of **e**vidence that was mistakenly excluded from this class. FN (false negative): the number of tests that were incorrectly assigned to this class [2]. The degree of adequacy of the characterization model results from the number of correct and unhealthy arrangements in each possible evaluation of the factors to be grouped. Rightness will be determined by utilizing beneath recipe.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{----------→} \quad (3)$$

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)

F1-Score= 2 * Precision * Recall / (Precision + Recall)

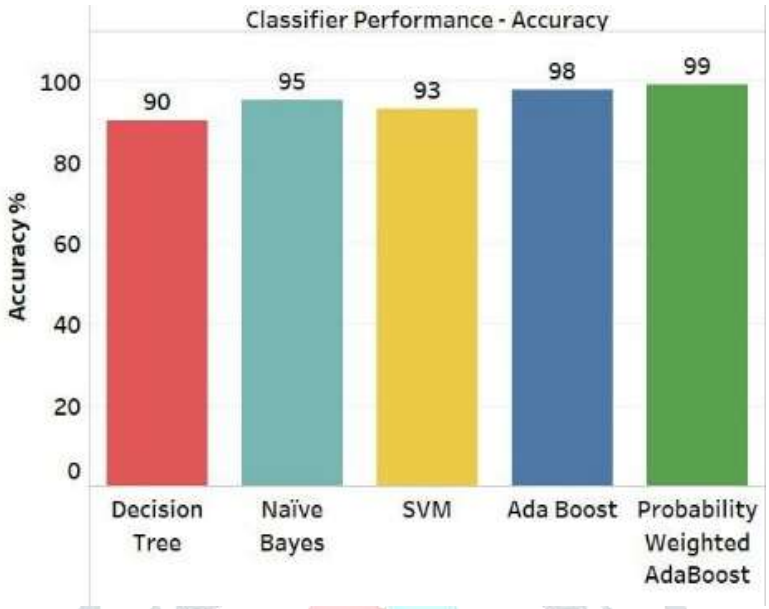

Figure 7: Accuracy

Figure 7 shows the performance of classification algorithms on the dataset based on accuracy. Accuracy of Decision Tree, SVM, Naïve Bayes, AdaBoost, and Probability Weighted AdaBoost is above 90% in all cases. The above results show that proposed algorithm outperforms the other algorithms.

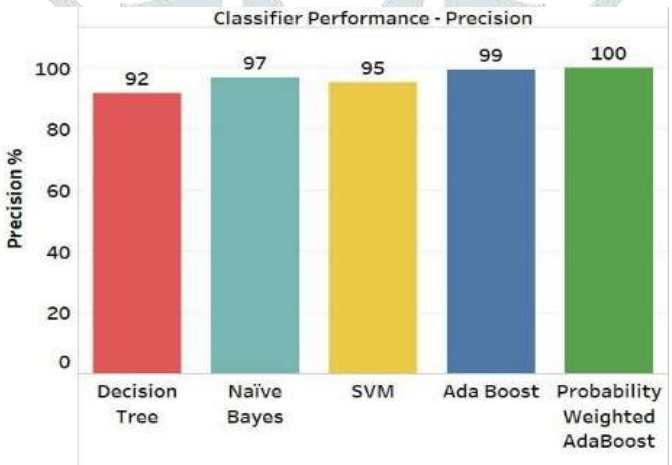

Figure 8: Precision

According to Figure 8, we will clearly see that [2]. PWA gives 100% precision and other classifications gives but 95% except AdaBoost. 100% precision is extremely rare to urge, but this algorithm achieves that result.

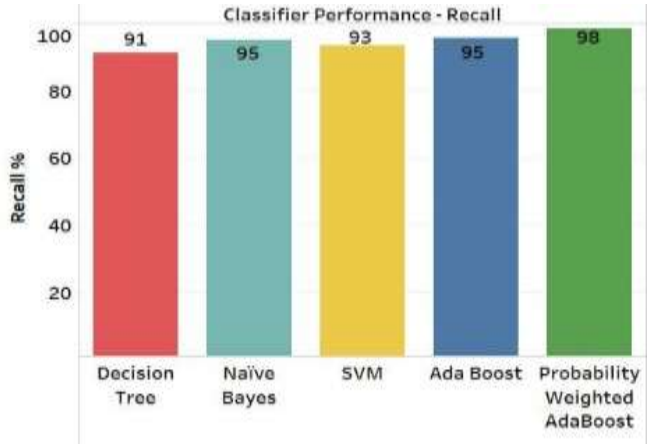

Figure 9: Recall

Figure 9 shows the performance of classification algorithms on the dataset based on recall. Recall of naïve bayes is good compared to Decision Tree and SVM. The above result indicate that proposed algorithm outperforms the other algorithms in recall parameter.

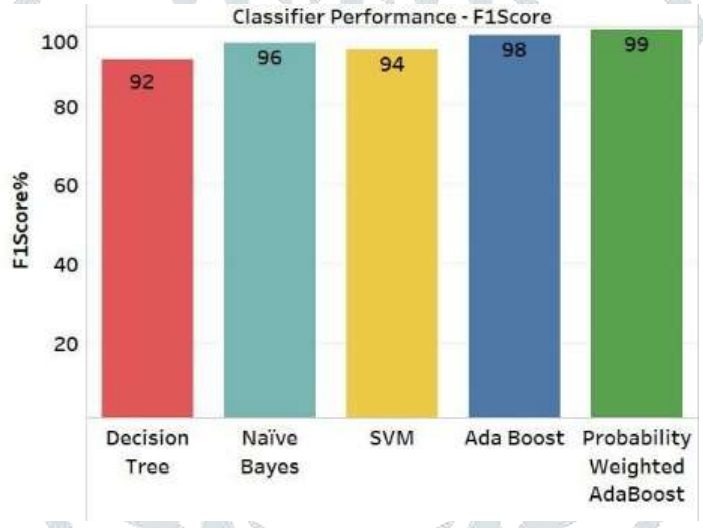

Figure 10: F1Score

According to Figure 10, we are able to clearly see that PWA gives 99% f1score and other classifications gives but 95% except AdaBoost which supplies 98%. 99% f1score indicates the effective classification performance of the proposed approach.

V. Conclusion:

The proposed research work assesses the capacity to tell apart CHKD with the help of Machine Learning Algorithms. The classifiers are talented, checked, and affirmed utilizing 10-crease cross-approval. Better was cultivated with the arranged calculation by F1-measure (99%), exactness (100%), review (98%) and precision (99%). This outcome is that the most extreme among going before examines. Since the data utilized during this examination is inconsequential, later on, the scientist focus to approve our outcome by utilizing greater dataset or partner the outcomes utilizing another dataset that contains similar highlights. Additionally, so as to assist in diminishing the predominance of CHKD, the analyst will gauge if a personal with CHKD hazard impacts like

diabetes, hypertension, and family background of kidney disappointment will have CHKD soon or not by utilizing suitable dataset.

VI. REFERENCES

1. Yang G, Guohua B,Yan H School of computing Blekinge Institute of Technology Karlskrona, Sweden, "Using Bayes Network for Prediction of Type-2 Diabetes" (2012)
2. Deepika K, Seema S Predictive analytics to prevent and control chronic diseases," 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp.:381-386, 2016 https://doi: 10.1109/ICATCCT.2016.7912028.
3. Gopalakrishna P the Practice of predictive analytics in healthcare 15 September 2014 pp 1-23.
4. Aqueel Ahmed, Shaikh Abdul Hannan Data Mining Techniques to Find Out Heart Diseases: An Overview. Inter J Inn Tech and Exploring Eng (2012) 1(4):1-3.
5. Durairaj M, Kalaiselvi G Prediction of Diabetes Using Soft Computing Techniques- A Survey". Inter J Inn Tech and Exploring *Eng.* (2015) 4(3):201-206.
6. Carter H Clinical sympathy: the important role of affectivity in clinical practice. J Europ Med Health Care and Phil (2019) 22:499–513.
7. Patil BM, Joshi RC, Toshniwal D Hybrid prediction model for type-2 diabetic patients. Expert systems with applications (2010) 37(12):8102-8108 https:// DOI:10.1016/j.eswa.2010.05.078.
8. Han J, Rodriguez JC, Beheshti M Discovering decision tree-based diabetes prediction model. In International Conference on Advanced Software Engineering and Its Applications (2008) https://doi.org/10.1007/978-3-642-10242-4_9.
9. Qin Y, Yu T, Peng F, Li-Li T, Yang MQ, Jing SL Design and development of a medical big data processing system based on hadoop. J Med systems (2015) 39(3):23.
10. Ivo DD, Methodological challenges and analytic opportunities for modelling and interpreting Big Healthcare Data. Gig science (2016) 5(1):12.
11. Shamim Hossain M, Ghulam M (2016) Healthcare Big Data Voice Pathology Assessment Framework. IEEE Access, 2016 4:7806–7815.https:// Doi: 10.1109/ACCESS..2626316.
12. Alejandro Rodríguez-González,AthenaVakali.Miguel,A,Mayer,Takashi ,Okumura Ernestina Menasalvas-Ruiz and Myra Spiliopoulou,. Introduction to the special issue on social data analytics in medicine and healthcare". J Sci and anal 2019 8:325-326
13. .V. Vapnik, "The Nature of Statistical Learning Theory." NY: Springer- Verlag. 1995 https://doi.org/10.1007/978-1-4757-3264-1
14. Peter Georgantopoulos, Jan M. Eberth, Bo Cai, Christopher Emrich, Gowtham Rao, CharlesL. Bennett, Kathlyn S. Haddock & James R. Hébert," Patient- and area-level predictors of prostate cancer among South Carolina veterans",Journal of cancer and causes control, vol:31. pp:209-220, January 2020.
15. hoi, E., Xu, Z., Li, Y., Dusenberry, M., Flores, G., Xue, E., & Dai, A. Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer. Proceedings of the AAAI

Conference on Artificial Intelligence, (2020). *34*(01), 606-613. https://doi.org/10.1609/aaai.v34i01.5400

16. Akshay Rajaram, Zachary Hickey, Nimesh Patel, Joseph Newbigging, Brent Wolfrom, Training medical students and residents in the use of electronic health records: a systematic review of the literature, *Journal of the American Medical Informatics Association*, Volume 27, Issue 1 January 2020, Pages 175–180, https://doi.org/10.1093/jamia/ocz178
17. Roosan, D.; Karim, M.; Chok, J. and Roosan, M Operationalizing Healthcare Big Data in the Electronic Health Records using a Heatmap Visualization Technique. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies* - . (2020). *Volume 5*, pages 361-368. DOI: 10.5220/0008912503610368
18. Z. Liu, X. Li, H. Peng, L. He and P. S. Yu, "Heterogeneous Similarity Graph Neural Network on Electronic Health Records," IEEE *International Conference on Big Data (Big Data)*, 2020, pp. 1196-1205, doi: 10.1109/ BigData50022.2020.9377795.
19. Le Meur, Nolwenn, and Fei Gao. "French Health big data experience: The evolution of the national health insurance information system." *European Journal of Public Health* 30.Supplement_5 (2020): ckaa165-1209.
20. Han, X., Zhang, S., Chen, Z., Adhikari, B. K., Zhang, Y., Zhang, J., ... & Wang, Y.). Cardiac biomarkers of heart failure in chronic kidney disease. *Clinica Chimica Acta*, (2020 *510*, 298-310.
21. Khanra, S., Dhir, A., Islam, A. N., & Mäntymäki, M. Big data analytics in healthcare: a systematic literature review. *Enterprise Information Systems*, (2020). *14*(7), 878-912.
22. Rehman, Arshia, Saeeda Naz, and Imran Razzak. "Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities." *Multimedia Systems* (2021): 1-33.
23. Sivaparthipan, C. B., N. Karthikeyan, and S. Karthik. "Designing statistical assessment healthcare information system for diabetics analysis using big data." Multimedia Tools and Applications 79.13 (2020): 8431-8444.
24. Rath, Mamata. "Big data and iot-allied challenges associated with healthcare applications in smart and automated systems." Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications. IGI Global, 2020. 1401-1414.