Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease

[1]Prof. Vishal Shinde, [2]Mr. Shubham Narayan Bhoir, [3]Mr.Zaid Irfan Khan, [4]Mr. Mrunal Umakant Nagmoti.
[1]Asst.Professor,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India. [1]mailme.vishalshinde@gmail.com,
[2]bhoirshubham999@gmail.com, [3]zaid041999@gmail.com, [4]mrunalnagmoti@gmail.com

Abstract – In today's life everyone has been trying to be conscious about health due to workload and busy schedule this gives attention to the health when a it shows symptoms but chronic kidney disease does not shows particular symptoms it is difficult detect and predict, to prevent this best solution prediction and analysis this problem Chronic Kidney Disease (CKD) has the type of chronic disease which means that a it happened calmly over period of specific time and persisted for a lengthy time there later. It is harmful at its end level and will be only an improving by kidney replacing and regular dialysis which is an artificial filtering mechanisms. It is very important a to identify chronic kidney disease at the early stage so that necessary treatments can be the provided to prevent and cure a kidney disease. The main focusing in this paper on classification techniques, that is logistic regression, tree-based decision tree and random forest determined. Different type measures used to of compare betwixt algorithms for a dataset gather from a standard uci repositories.

Keywords- disease prediction, logistic regression ,decision tree, Random forest

I.INTRODUCTION

As the beginning of civilization, man wrote records and documents chronic kidney disease (CKD) has a dangerous and malicious health situation globally that is a major causes for harmful health outcomes, especially in different countries where income area from low-to-middle where millions people die constantly due to absence of a limited treatment. As per given levels in any CKD the fatality is related to the stage it had been without being cured. The high-risk factors of CKD are day by day increasing frequency of diabetic patient, hypertension, mellitus and family the past of kidney failure.

If Kidney disease is left undetected and therefore not treated, then it can take path to hypertension and in many cases to kidney get failure. This had obtained a standard datasets from mainly uci machine repositories for chronic Kidney Disease. Chronic kidney disease if predicted before time and accurate, can benefit gets patients in so many ways. It is expansion and increasing the expectation of a successful treatment while also adding years to the person's life span. This paper focus as aims to predict CKD next to using some about the choose machine learning algorithms with also used feature selection methods. The objectives are to gathers the combination of disparate types of feature and then used it as take input to the machine learning different algorithms. The algorithms have been implementing on basis of selected features and then comparing their performances. paper is organized in different sections. Section survey discusses research works similar to this paper. Sections system architecture are been given brief on the method followed in the research and describes the concepts used.

II. AIMS AND OBJECTIVE

a) Aim

Aim of performance analysis of machine learning classifier used for the predicting chronic kidney to the disease. It will help out the patients to recognize their health but also assist the doctors in medicine suggestion well in advance. Because in early stage it is diffcult to find the disease but with dataset obtained from uci (california university at Irvine city) repository.it can predict the disease easily. Self closing the processing of predicting diseases prove convenient and time-redeem for the practitioners in the areas of medical diagnosis sector.

b) Objective

- Identify and Early prediction of disease.
- Accurate prediction using classification algorithm using that algorithm predict disease.
- to predict patient amidst chronic kidney disease make use of less numbers attribute while maintained the higher accuracy.

III. LITERATURE SURVEY

Paper 1: Chronic Kidney Disease Using Machine Learning Techniques :

This chronic kidney disease (CKD) has harmful disease effecting many people worldwide. Idiomatic including chronic KD has often unaware that the medical tests they

 DOI : 10.35291/2454-9150.2022.0172

undergo can be provide useful information about CKD for other purposes and this information may not be used impressively to address disease diagnosis. The essential problem of that disease is that it is very hard to admit till it reaches advanced stage. In this research paper are doing the forecast chronic kidney disease {CKD) using the machine learning techniques. In this paper, are using machine learning different algorithms like naïve bayes classification, decision tree, logistic regression, support vector machine(SVM) and random forest in this project In research paper that this identify the chronic kidney disease (CKD) using the perfect suitable method and got 99.5% as most authentic result using random forest method.

Paper 2:Diagnosis of CKD using effective classification and feature selection techniques:

The huge quantity of data collected by the healthcare section can be very productive in analyzing, diagnosing and making decisions if it properly mined. Encrypted information extracted from very enormous quantity data can provide assistance and solutions to address critical health care situations. ckd is a baleful disease that can prevented with proper predictability and appropriate safety precautions. Excavation of data collected from preceding diagnosed patients unlock up new chapter of medical development. However, certain strategies should be used to achieve a better outcome. In this manuscript ability to separate Support Vector Machine, Decision Tree, Naïve Bayes and K-Nearest algorithm, frinspecting the CKD information and data collected at UCI site, was examine to predict kidney disease. Data set was analyzed according to accuracy, Root mean squared errors, Mean Absolute Error and collect Performance Curve. In the current study, Decision Tree shows auspicious results when used with the WEKA data mining tools. measurement algorithm provides distinguished improvements in the classification of appropriate numerical symbols. fifteen demonstrated that the magic number of select attributes of given database leading to the highest percentage accuracy **Paper3:Prediction performance of individual and ensemble learners for CKD:**

Automated diagnostic procedures prove helpful and save time for doctors in the field of medical diagnosis. Accurate prognosis for any disease not only comfort patients to know their health but also helps physicians to prescribe medication earlier.

In today's life-style, pre-existing knowledge of health and proper concern can extend the life span of a patient. In this paper, the prognosis for CKD has made using individual student and combination. The experiments were performed on the CKD database collect from the UCI archive. Forest (RF), bags, AdaBoost respectively used for forecasting. For weka tool this used the open source, for all the testing. Results are tested using accuracy, precision, recall, roc and f-measure performance measurements.

The results suggested that a student-based decision-making tree (J48) and a random forest from the classifier of the series respectively worked better than other class dividers.

Using precision, accuracy recall, f-measure performance measurements gets accurately Results are tested

The results suggested that a student-based decision-making tree (J48) and a random forest from the classifier of the series respectively worked better than other class dividers. Its disadvantage that in this if it will not provided proper data it will not provides the proper solution and not give proper prediction .in this paper method used that consisted naïve bayes minimal sequential optimization is used. this used an unlock source, weka tool, for all testing. Results are tested using accuracy, precision, recall, F-measure and ROC performance measurements. The results suggested that a student-based decision-making tree (J48) and a random forest from the classifier of the series respectively worked better than other class dividers.

IV. EXISTING SYSTEM

The cosmology and mechanical learning of the chronic disease (ckd) as a flexible WEKA flexible tool. The Ontology and machine learning different methods had been used in the existing system.

Therefore, it shows kidney disease to helps the error detection tool and helps physicians to adequately identify patients with sever kidney disease to those with different major symptoms and causes of kidney diseases. Another machine learning process is the method of Coronary arteries diseases called N2 Genetic optimizer agent (another genetic preparation of it generated) introduced in this a methods.

These results are be aggressive and are similar to the best results in the field represented and shown.

Limitation of exiting system :

1. based on machine coronary artery disease examined through datasets, test sizes, highlights, areas of information accumulations and execution measurements
2. applied ML are the basic different methods that have implemented broken down in this methodology.
3. Chronic kidney Failure Detection is an anticipated the Constant kidney break down the identification from heart sound appropriate as pile of machine learning classifiers.
4. The plan of action used to for feature extractions to the models and see comprises filtering segmentation.
5. Kidney Failure Detection is a frequently anticipated.

 DOI : 10.35291/2454-9150.2022.0172

V. COMPARATIVE STUDY

Sr. No	Papers Name	Authors/ Publication	Method	Advantage	Disadvantage
1.	Prediction of chronic kidney disease (CKD) Using Machine Learning Technique	Mrs PrasunaKotturu, Mr. VVss Sasank, G Supriyaa, ch Sai Manoj.	Decision tree, Naïve Bayes classification,Logistic Regression, Random Forest.	it can identify disease easily,get 99.3% accurate result .	Major problem is this disease is hard to recognize till it reaches to advanced stage.
2	Diagnosis of CKD using effective classification & feature selection technique	Nusrat Tazin , Shahed AnzaruSahab	SVM, Decision tree,Naïve Bayes and K-Nearest neighbor algorithm	Massive amt of data collected by healthcare section can be adequate for analysis.	Problem of this is it hard to regonize till it reaches advanced stage, time consuming.
3.	Prediction performance of individual amd ensemble learners for CKD	Dili Singh Sisodia, Akansha Verma	Naïve Bayes, minimal sequential optimization (SMO)	Result suggest that algo decision tree based particular learners and random forest algo ressemble classifier respectively perform give better than other classifier	if not provide proper data it will not provide proper prediction
4.	Prediction of CKD stages using the data mining algorithms	El-Housaainy, A Radya, Ayman S.Anwa	Probabilistic Neural Networks(PNN),(SVM) ,Multilayer Perceptron(MLP),Radial function algo	The PNN algorithm provides better classificatio-n and prediction performance	Early detection & characterization are taken to be critical factors to manage and control of CKD Performance severity stage.

*Table no. 01 – **Comparative analysis***

VI. PROBLEM STATEMENT

- Sometimes model also get failure due the lot of load frequently increased.
- As given dataset has not correct and arranged improperly then it cause problem.
- At a time model cannot run multiple model, so efficiency decrease.
- Chronic kidney Failure Detection is an anticipated the Constant kidney malfunction detection coming from heart sounds make use of pile of a machine learning classifiers.
- The perfect accuracy of the classifiers was estimated using the confusion matrix.

VII. PROPOSED SYSTEM

This project worked on heart disease data acquired from uci set(university of the california at irvine city) repository, the data set consist of attributes such as age, sex, cp, Blood pressure, Blood urea, Sugar, Potassium, ca, and target with 401 instances has taken. At first stages of, dataset is first cleansed and processed using preprocessing techniques like Data Integration, Data transformation, Data depletion and Data cleaning using pandas tool.

The proposed framework a total of 401 patient records were anticipate. Data visualization techniques helps the data scientist to understand the feasibility of the dataset.

Advantage proposed system:

1.The accuracy of the classifiers was estimated help using the confusion matrix and correlation matrix or features selections.

2. The classifier which bags up the highest accuracy could be determined as the best classifier.

ALGORITHMS: logistic regression algo ,decision tree algo , random forest algo.

VIII. ALGORITHM

The Algorithm for predicting the chronic KD :

Step.1: Start
#importing libraries
from admins import views as admins
from sklearn.ensemble import RandomForestClassifier from sklearn.treeimportDecisionTreeClassifier
from sklearn.svmimport SVC
from sklearn.linear_modelimportLogisticRegression from
Step.2: processing dataset
import files from section google colab #Only use for Google Colab uploaded ← files.upload() #Only use for Google Colab
df ← pd.read_csv("kidney_disease.csv")
Step.3: feature selection process
- feature selection method implement through univariate selection and correlation matrix and feature importances
Step.4: classification algo implement
1. random forest
clf←Model(model←RandomForestClassifier(),X←X,y←y)
clf.crossValScore(cv←10)
clf.crossValScore(cv←10)
clf.accuracy()
clf.confusionMatrix()
clf.classificationReport()

 DOI : 10.35291/2454-9150.2022.0172

2 . Decision tree

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

dt←Model(model←DecisionTreeClassifier(),X←X,y←y)
dt.crossValScore(cv←10)
dt.accuracy()
dt.confusionMatrix()
dt.classificationReport()
3. logistic regression
lr←LogisticRegression()
pipeline←make_pipeline(QuantileTransformer(output_distribution←'normal'), lr)

Step.5: all the three implement algo predict result.
Step.6: Exit

IX. MATHEMATICAL MODEL

$$F = \frac{\frac{\sum \quad n_j(\underline{X_j} - \underline{X})^2}{(k-1)}}{\frac{\sum \quad \sum \quad (X - \underline{X_j})^2}{(N-k)}}$$

Univariates selection depend on result has , at a univariate statistical test. The best study that are the selected on the tests basis. This selection are compared particular features to the variable, to recognize the either statistically significant affinity between target supervised variable. it called ANOVA. When analysis between feature and supervised variable is done , it ignores other features is called to be univariate; each feature has its individual test score finally all test scores will compared top scores will be selected .this formula used for analysis of variances:

On dataset, set of a correlation value between each pair of its attributes is arranged in form as matrix which called as correlation matrix . correlation is statistical term which refers to estimate of closeness the most popular method person correlation coefficient . this this formula have used for correlation the

$$\sigma = \sqrt{\frac{\sum \quad (\chi - \mu)^2}{N}}$$

Standard deviation ;

Where, here COV (X ,Y) =

$$\frac{\sum_{i=0}^{n} \quad (x_i - E(X))(y_i - E(Y))}{N-1}$$

Measured performance evaluation :

To measure the performance of this algorithms this will use different test functions of the tree. In these activities TP refers to positive and predictable conditions by algorithms. TN as negative conditions are predicted the negative FP are conditions that are predicted to have but were actually negative but actually get positive.

$$\frac{TP + TN}{TP + FP + FN + TN}$$

Accuracy: as a count of correctly predicted recognition in all noted and ratio of accurately predicted monitoring to the total monitoring.

Precision: Part of the predicted predictors are actually positive. It is therefore the count of positive things (TP) and the count of cases predicted elected positive value (TP + FP)

$$\frac{TP}{TP + FP}$$

Recall: recall is define that proportion actual positives that were expected that positive.it ratio of true measure (TP) to the actual (TP+FP)

$$\frac{TP}{TP + FN}$$

Classification algorithm :

Decision Tree: In this category of tree-shaped models, data is divided into smaller sets. The final result of the algorithm is a tree along with algorithm decisions:
• With training data D, it starts with one node N.
• N will be a leaf if all data in D falls under the same class.
• The example in 'D' is appropriately classified.
• Repeat this algorithm for each one subclass of 'D'
Random forest Algorithm:
A random algorithm creates a forest of 'decision tree' The accurize of the Forest is directly related to the count of trees in it.
• First its select 'p' randomly in absolute features of 'q' (p << q).
• In the selected 'p' feature, it should calculate the area, pointing 'd' using the method The best point of difference.
• Using this method it must split the nodes into daughters.
• Forest 'n' tress number created using the above steps 'nTime sets.
Logistics regression :
The central nervous structure system (cns) or heartbeat is a logistic activity, also known as sigmoid activity.
Another type of regression is Linear Regression that will predict continuous dependency variability.
Decreased mobility also predicts phase-dependent variables in other arms.
Obtain accuracy using the three algorithms mentioned below:
[98.10% 100% , 99.61%]

X. SYSTEM ARCHITECTURE

Description: 4 Step Process

 DOI : 10.35291/2454-9150.2022.0172

1. Uploading raw data and then preprocessing data

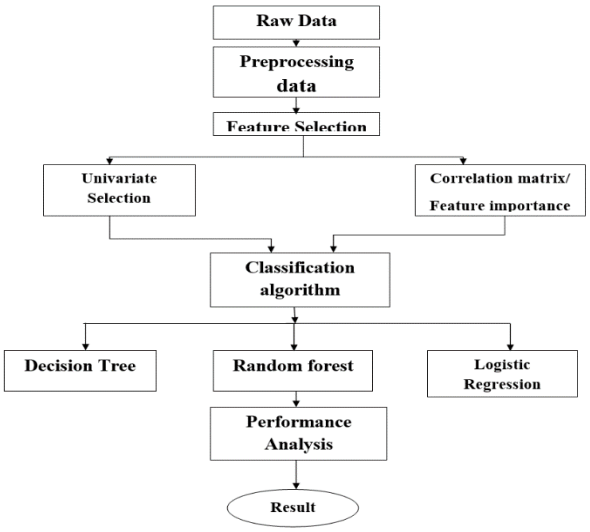

Fig.1: **System architecture diagram**

2. Feature selections in univariate selection and correlation matrix
3. Implement on that classifications algorithm.
4. Performance analysis and given the result.

XI. ADVANTAGES

1.It is impulsive identify the essential accent without any human supervision.

2.Helps solve complex problem and real-world problems with several constraint.

3.Provides a path towards adequate artificial General Intelligence some day in the future.

4.In less quantity of data, this can achieve more accuracy and prevent patient form death and identify different types of features

XII. DESIGN DETAILS

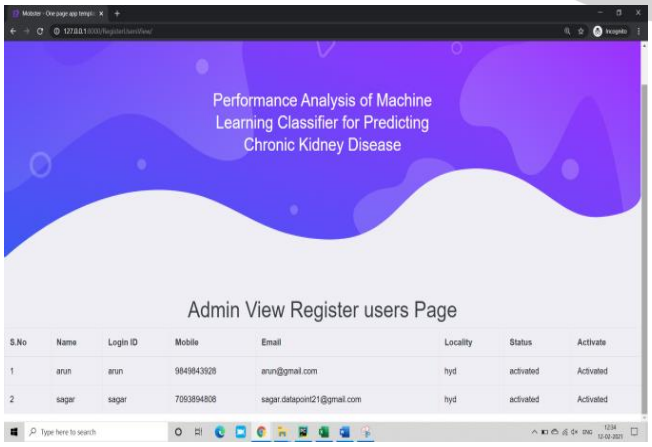

Fig 2: **Result**

The web application is created using python django framework. The user enters symptoms when user click on predict then the pre-trained model predict that if he has chronic disease or not.

XIII. CONCLUSION

Thus, we have tried to implement the paper "Rahul Gupta , Nidhi Kolo ,Niharika Mahor ,and N Tejashri","Performance Analysis of machine learning Classfier for predicting Chronic kidney disease", IEEE 2020 and according to the implementation this paper conclude that the proposed algorithms are random forest , decision tree and logistic regression have achieved an accuracy of 97.48, 94.16 and 99.24 respectively. Precision of 100, 95.12 and 98.82 and recall of 97.61, 96.29 and 100. Two feature selecting techniques are combined by leveraging the strength of each the techniques. On comparison find the Logistic Regression with highest accuracy and recall while Decision tree have the highest precision.

REFERENCE

[1] Rahul Gupta , Nidhi Kolo ,Niharika Mahor ,and N Tejashri,"Performance Analysis of machine learning Classfier for predicting Chronic kidney disease." International Conference For Emerging technology (INCET),IEEE 2020.

[2] D. S. Sisodia and A. Verma, "Prediction Performance of Individual and Ensemble learners for Chronic Kidney Disease," 2017, pp. 1027– 1031.

[3]A. V Kshirsagar et al., "A Simple Algorithm to Predict Incident Kidney Disease," ARCH Intern Med, vol. 168, no. 22, pp. 2466–2473, 2008.

[4] S. Kumar Sahu and, A. K. Shrivas "Classification of Chronic Kidney Disease using Feature Selection Techniques," IJCSE, vol. 6, no. 5, pp. 649–653, 2018.

[5] M. Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm," Int. J. Computer. Sci. Mob. Computer., vol. , no. 2, pp. 24–33, 2016.

[6] A. S. Anwar and E. H. A. Rady, "Prediction of kidney disease stages using data mining algorithms," Informatics Med. Unlocked, vol. 15, pp. 1–7, Jan. 2019, doi: 10.1016/j.imu.2019.100178.

[7] K. Chandel, V. Kunwar, S. A. sai, and A. Bansal, "Chronic kidney disease analysis using data mining classification techniques," 2016, pp. 300–305.

 DOI : 10.35291/2454-9150.2022.0172