A COMPARATIVE STUDY ON CHRONIC KIDNEY CANCER PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHMS

[1]Swarnalatha Prathipati,[2]Sagiraju Snehith varma,[3]Alumalla Jayanth rakesh,[4]Mulla Shyam Chandrashekar reddy,[5]Javeed ali,[6]Pinesetty Ruthvik varma sree

[1]Assistant Professor,[2,3,4,5,6]Student

[1,2,3,4,5,6] Department of Computer Science and Engineering, GST, GITAM University, Visakhapatnam, AP, India
sprathip2@gitam.edu,snehithvarma555@gmail.com ,jayanthrakesh499@gmail.com ,chandumula41@gmail.com, javeed07ali@gmail.com,ruthvikvarma1236@gmail.com.

---***---

Abstract - Early detection and appropriate treatment can halt or postpone the progression of this chronic condition until the point at which kidney transplantation or dialysis are the only options left for saving the patient's life. Consequently, it is crucial to have reliable tools for the purpose of early chronic renal disease detection.In this paper we will be using supervised machine learning models to predict chronic kidney disease.We integrated the influencing factors of Chronic kidney cancer ,collected a sample dataset , conducted data analysis and optimized the dataset .A kidney cancer prediction model was created using five supervised machine learning algorithms: Xgboost, gradient boosting, support vector machine, logistic regression, and random forest classifier.The predictive performance of these five algorithms were compared and discussed..The models are compared using a variety of metrics, including accuracy, precision, recall, F1 score, etc. The results of the experiment demonstrate that the random forest algorithm is the most accurate method for predicting kidney cancer.

Key Words: Chronic kidney disease, Classification metrics, Logistic regression, Xgboost , Random forest classifier, Support vector machine, Gradient boosting

1.Introduction

Renal cell carcinoma, another name for chronic kidney cancer, is a cancerous tumor that develops from the cells lining the tiny tubes inside the kidney. Early on, the disease frequently exhibits no symptoms, making a diagnosis challenging. As a result, there is a requirement for trustworthy and precise predictive models that can identify the disease's presence at an early stage.

Decision trees, random forests, support vector machines, and artificial neural networks are examples of supervised machine learning algorithms that have become effective tools for predicting chronic kidney cancer in recent years. In order to analyze large datasets and find patterns that can be used to identify people who are at risk of developing the disease, these algorithms make use of a variety of statistical and mathematical techniques.

Compared to conventional statistical methods, the use of supervised machine learning algorithms for the prediction of chronic kidney cancer has several benefits. First of all, these algorithms are capable of analyzing vast amounts of data and finding patterns that might not be obvious to human observers. Second, they can work with many different input data types, such as continuous, categorical, and binary variables. Finally, real-time accurate predictions made by these algorithms can be helpful in clinical settings.

The purpose of this study is to compare how well various supervised machine learning algorithms predict chronic kidney cancer. A sizable dataset of patient data, medical histories, and laboratory results will be used in the study. Different supervised machine learning algorithms will be applied to the training set in order to create predictive models. The dataset will be split into a training set and a testing set. Each algorithm's performance will be assessed on the testing set using a variety of metrics, including accuracy, precision, recall, and F1 score.

In conclusion, this research paper aims to provide insights into the potential of supervised machine learning algorithms in predicting chronic kidney cancer. The findings of this study can potentially aid in the development of an automated prediction system for chronic kidney cancer, which could lead to earlier diagnosis and improved patient outcomes.

2.Related work

Not just in the fields of mathematics and engineering, but also in the field of medicine, machine learning algorithms have become an extremely important instrument. Predictions are made frequently using machine learning algorithms so that speedy decisions can be made. Algorithms are mostly used to enhance performance and accuracy.

The decision tree algorithm was found to be the best performing supervised machine learning algorithm out of the four used by the researchers in this work.[1].

The kidney function test (KFT) dataset was gathered by Vijayarani and Dhayanand [2] from medical labs, research facilities, and hospitals. There are 584 occurrences and 6 attributes in the dataset. The algorithms SVM and ANN were employed. The accuracy of ANN was the highest at 87.7%.

Xiao et al. [3] used the data from 551 patients and 9 machine learning algorithms. In accordance with their analysis of accuracy, ROC curve, precision, and recall, the linear model offered the highest level of accuracy.

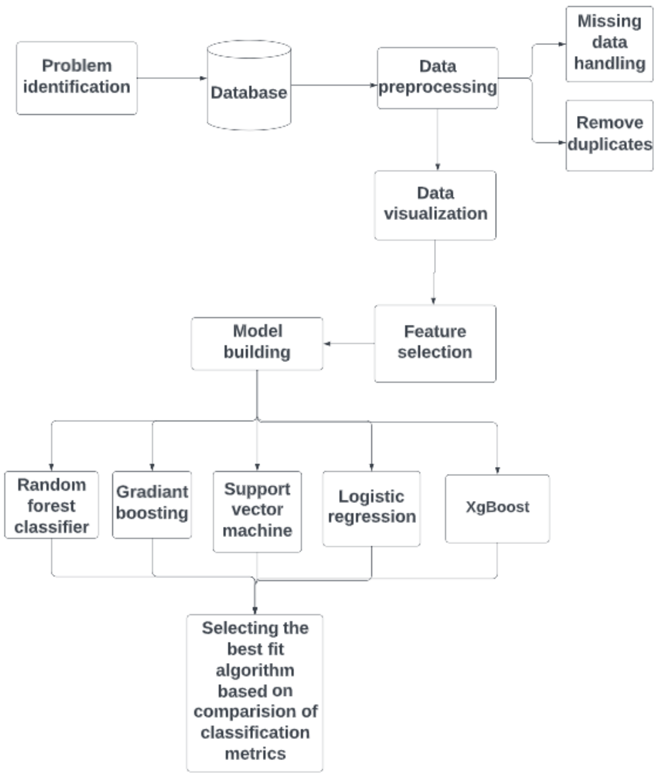
[4] In Reshma et al, the feature selection method was used on the CKD Dataset. The features were chosen using the ACO method. ACO is the abbreviation for the feature selection meta heuristic algorithm. It is a member of the Wrapper technique group. In their dataset, a total of 24 attributes were available. The feature selection technique was applied, and twelve features were used to build the model. The model was developed using the Support Vector Machine Classifiers method.

3.Methodology

The project's whole work flow is provided by the system methodology, which aids in a better understanding of the project. This project's dataset was gathered from an online data source. The gathered data is then preprocessed to make it ready for any machine learning algorithm to use. The missing values are added to the dataset through additional processing. The features that are not very useful in determining the outcome can be eliminated using feature selection.

The preprocessed dataset is used to implement the Xgboost , logistic regression, support vector machine, gradient boosting, and random forest classifier methods. To determine the best fit algorithm, the classification metrics of each method are compared.

fig1: The following figure illustrates the methodology deployed to find the best algorithm to detect kidney cancer



4. Proposed work Implementation

4.1 Dataset collection

We obtained the information from an online database, and it contains the records of 830 patients, and each one has 26 characteristics.the tools that we have used for this research is google collab as a model building tool and python as a programming language

4.2 Data Preprocessing

Data preprocessing is important because it helps to clean, transform, and prepare raw data for analysis. This step includes removing errors, filling missing values, handling outliers, and scaling variables to ensure accurate and reliable results. By performing data preprocessing, researchers can improve the quality of their data, increase the accuracy of their analyses, and obtain meaningful insights from their data.

The dataset we collected includes a few null values. Using the following methods, we have removed the null values.

A. Random value imputation

Random value imputation is a method for handling missing data by replacing the missing values with randomly generated values from the variable's distribution. This technique can be used for numerical or continuous variables, and the randomly generated values can be selected from a uniform or normal distribution.

B. Mode imputation

Mode imputation is a technique for handling missing data by replacing the missing values with the most commonly occurring value in the variable. This method is most appropriate for categorical or nominal variables or when the variable has a discrete numerical range.

We have used random value imputation for numerical columns and mode imputation for categorical columns in the dataset to remove the null values.

Label encoding is important because it converts categorical variables into numerical variables, which is necessary for many machine learning algorithms. This process assigns a unique numerical value to each category of the variable, which allows the algorithm to recognize and process the data. Label encoding is simple and efficient, and it can significantly improve the accuracy and performance of machine learning models.

C. Label encoding

For label encoding,the label encoder class is used.The LabelEncoder class in scikit-learn is a preprocessing tool that allows for the conversion of categorical variables into numerical variables. It operates by giving each category in the variable a different integer value.The LabelEncoder class can be instantiated and fitted to the categorical variable using the fit() method, which identifies the unique categories in the data. Then, the transform() method can be used to transform the categorical variable into a numerical variable.

4.3 Feature Selection

The process of choosing a subset of pertinent features from a bigger set of features to be used in machine learning models is known as feature selection. This is done to reduce the complexity of the model, improve accuracy, and reduce computation time.A critical stage in the machine learning pipeline for effective and efficient analysis is feature selection.lasso regularization is used for feature selection in this research.

Lasso regularization

Lasso regularization is a machine learning technique that reduces the less significant characteristics to zero by adding a penalty term to the cost function in order to prevent overfitting.It results in sparse models where only the most important features are retained, improving the model's interpretability and performance. The amount of regularization is controlled by a hyperparameter, lambda, which can be tuned using cross-validation.

From the group of 26 features, a total of 18 features are chosen after feature selection.the following 18 characteristics: Age (years), Blood Pressure (mm/Hg), Specific Gravity, Albumin, Sugar, Blood Urea (mgs/dL), Blood Glucose Random (mgs/dL), Blood Creatinine (mgs/dL), Serum Creatinine (mgs/dL), Sodium (mEq/L), Potassium (mEq/L), Packed Cell Volume, Red Blood Cells (millions/cmm), Anemia, Diabetes Mellitus, Appetite, Pedal Edema, and Blood Pressure are more pertinent

4.4 Algorithms

Five supervised algorithms were used in this study.The modified dataset from the previous stages is then divided into training and testing dataset. the algorithms get implemented on the training dataset.

A. Logistic regression

A statistical method known as logistic regression is used in machine learning to solve issues involving binary categorization, such as predicting whether or not a consumer would purchase a product based on one or more input variables.
Using a logistic function, the logistic regression model calculates the likelihood of the binary result, which has a range of 0 to 1. The projected probabilities are then categorized into one of the two classes by the model using a threshold value.

B. Support vector machine

The supervised ML algorithm SVM can be applied to both classification and regression applications. SVM seeks to identify a decision boundary that divides the data into distinct groups during classification, while during regression it seeks to identify a function that forecasts a continuous output variable.

C. Gradient boosting

A well-liked machine learning algorithm for classification and regression tasks is gradient boosting. It is an ensemble learning technique that combines several weak learners to produce a strong learner (such as decision trees). The gradient boosting algorithm builds decision trees onto the model iteratively. Each tree is trained to forecast the residuals (i.e., the discrepancies between the actual values and the predicted values) of the preceding trees. In other words, every new tree is trained to fix the mistakes made by the older trees.

D. Random forest classifier

A common ensemble learning approach for classification problems, where the objective is to predict the class of a data point based on its attributes, is the Random Forest Classifier. The accuracy and resilience of the model are increased by the random forest approach, which combines numerous decision trees

E. Xgboost

XGBoost (short for eXtreme Gradient Boosting), is frequently used to tackle classification and regression issues. A strong classifier or regressor is produced by combining a number of weak decision trees using this effective ensemble learning technique.The gradient boosting framework serves as the foundation for XGBoost, which trains its models using this method. It functions by improving the prediction accuracy of the model by repeatedly training decision trees on the residuals of the prior trees.

4.5 Classification metrics

Machine learning metrics are used to track and evaluate the model's performance. It is necessary to compare the models using various metrics after applying the machine learning models on the dataset. We must utilize the classification metrics in this case because we are performing categorization.The classification metrics that we used in this study are accuracy,precision,f1-score,recall and error rate.

5.Result & Analysis

In this study the dataset was initially divided into numerical and categorical columns, and the null values in each column were removed.For removing null values in numerical and categorical columns, we used random value imputation and mode imputation, respectively.we have turned the non numeric columns into numeric columns using scikit-learn's label encoder class.the modified dataset is then divided into testing and training for the implementation of the five algorithms and are then compared using the classification metricsThe accuracy of the Random Forest method was 96.3%, with precision, recall, F1-score, and error rate coming in at 96.1%, 96.9%, 96.5%, and 3.6%, respectively.The F1-score, precision, recall, and errorate of the logistic regression method were 95.6%, 84.6%, 89.7%, and 10.0%, respectively, with an accuracy of 89.9%.The accuracy of the support vector machine method was 59.4%, and the precision, recall, F1-score, and error rate were, respectively, 89.1%, 25.3%, 39.5%, and 40.0%.The F1-score, precision, recall, and errorate of the gradient boosting algorithm were 95.9%, 91.5%, 93.7%, and 6.4%, respectively, with an accuracy of 93.5%.The F1-score, precision, recall, and errorate of the Xgboost algorithm were 96.8%, 93.8%, 95.3%, and 4.8%, respectively, with an accuracy of 95.1%.

Table1: comparison of performances of different algorithms based on classification metrics

	Random Forest	Suppor t Vector	Xgboost	Logisti c	Gradient Boostin g
Accuracy	96.3	59.4	95.1	89.9	93.5
F1 score	96.5	39.5	95.3	89.7	93.7
Recall	96.9	25.3	93.8	84.6	91.5
Precision	96.1	89.1	96.8	95.6	95.9
Error rate	3.6	40.0	4.8	10.0	6.4

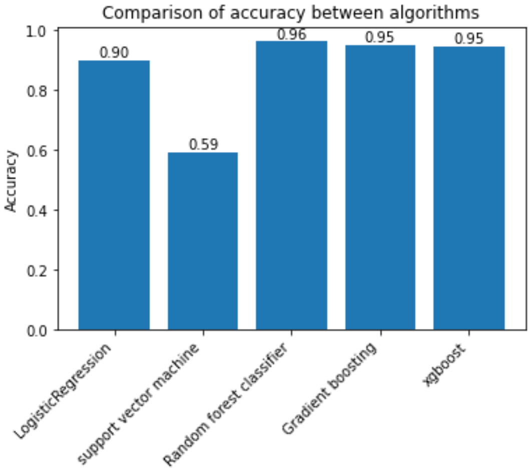


Fig 2:Comparison of accuracy between algorithms

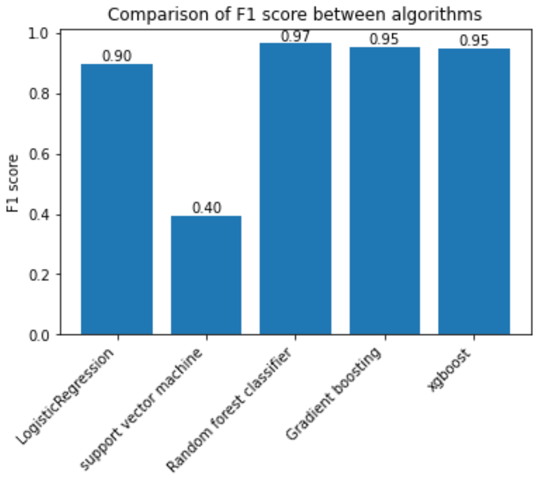


Fig 3:Comparison of F1 score between algorithms

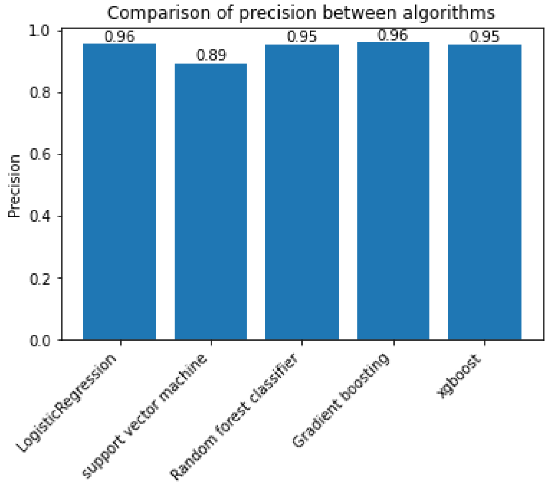


Fig 4:Comparison of precision between algorithms

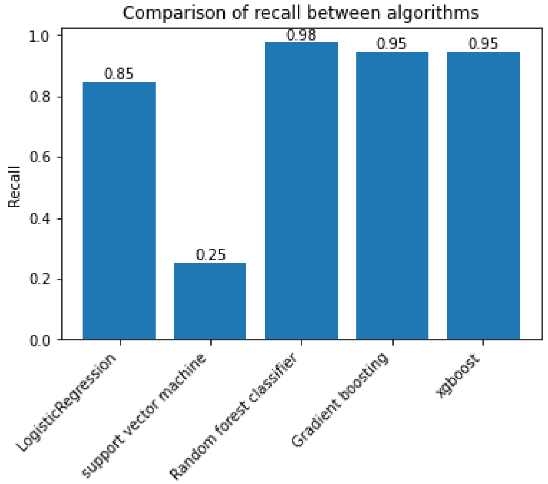


Fig 5:Comparison of recall between algorithms

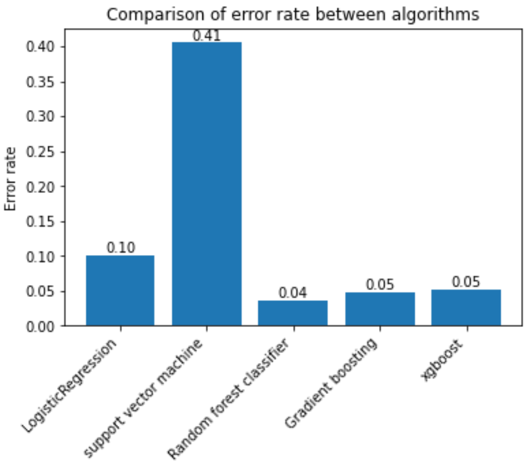


Fig 6: comparison of error rate between algorithms

6.Conclusion & Future scope

Random forest classifier, out of the five algorithms, gives the best performance.support vector machine being the least.The random forest and other machine ML algorithms can be compared similarly in further research. If the dataset is larger or when better data is available, a brief study can also be done. This research can be utilized to create a website where people can provide information from test results and estimate their risk of developing kidney cancer.A real time system can be developed which gives the best communication to the end user or the patient.

7.References

[1].“Implementation of machine learning algorithms to detect the prognosis rate of kidney disease”, *inocon.* ,Nov ,2020.

[2].S. Vijayarani1 and S. Dayananda, "Kidney disease prediction using SVM and ANN algorithms' ', *Int. J. Comput. Bus. Res.*, vol. 6, no. 2, pp. 1-12, Mar. 2015.

[3].J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, et al., "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression", *J. Transl. Med.*, vol. 17, pp. 119, Dec. 2019.

[4].J. M. Pereira, M. Basto and A. F. D. Silva, "The logistic lasso and ridge regression in predicting corporate failure", *Procedia Econ. Finance*, vol. 39, pp. 634-641, Jan. 2016.