WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Data-Driven Approach to Kidney Disease Prediction Using Machine Learning

JENIFER A[1], Dr. NANCY JASMINE GOLDENA[2]

[1] Reg.no: 23081205300112017, Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Affiliated to Manonmaniam Sundaranar University, Tirunelveli - **627007**jeniferrosy14 @gmail.com

[2]Associate Professor, Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Affiliated to Manonmaniam Sundaranar University, Tirunelveli - **627007**, nancy_lordwin@rediffmail.com, ORCID: 0000-0002-8626-2604

ABSTRACT :

Millions of people worldwide suffer from chronic kidney disease (CKD), a serious public health concern. Effective treatment and better patient outcomes depend on early diagnosis. Using clinical and biochemical data, this study predicts CKD using machine learning approaches. We seek to develop a predictive model using two datasets, one raw (ckd.csv) and one preprocessed (Preprocessed.csv), each including 25 features that range from demographic data (e.g., Age, Blood Pressure) to medical test findings (e.g., Red Blood Cells, Packed Cell Volume). The preprocessed version of the raw data addresses missing values, discrepancies, and categorical data. Machine learning algorithms can use this cleaned dataset because variables have been standardized, encoded, and missing values have been imputed. Through the use of different models, including decision trees, we investigate how accurate each method is at predicting the diagnosis of chronic kidney disease .In order to help medical practitioners make prompt and well-informed decisions regarding kidney health management, this study offers a thorough investigation of how data preparation and various machine learning models might improve the identification of CKD.

Keywords: Chronic Kidney Disease (CKD), Data Preprocessing, Missing Value Handling, Dataset Cleaning, Binary Encoding

1. Introduction :

The degenerative illness known as chronic kidney disease (CKD) is typified by the progressive loss of kidney function, which frequently results in serious side effects like cardiovascular disease or renal failure[1] . Chronic kidney disease (CKD) affects millions of people globally and places a heavy cost on public health systems. The diagnosis of chronic kidney disease (CKD) can be difficult because of the wide range of symptoms and risk factors, but early detection and treatments are essential to stopping its progression[2] . Conventional diagnostic techniques depend on a mix of laboratory testing and clinical evaluations, which may not always yield precise or timely predictions, especially in the early stages.

With the promise to improve diagnostic precision and offer early warning systems for chronic illnesses like chronic kidney disease, machine learning has become a valuable instrument in the healthcare industry in recent years[3] . Large patient data sets can be analyzed by machine learning algorithms, which can then uncover hidden patterns and provide prediction insights that surpass traditional methods. The purpose of this research is to use machine learning approaches to forecast the emergence of chronic kidney disease (CKD) based on a variety of clinical and biochemical indications, such as age, blood pressure, and laboratory test results.

This study will investigate the efficacy of various machine learning models in CKD prediction using a real-world dataset[4] . To guarantee accuracy and consistency, the raw dataset—which includes inconsistent and missing data—is preprocessed. This allows the models to generate accurate predictions. In order to identify the best accurate and effective method for CKD prediction, this study compares a number of methods, including support vector machines, random forests, and decision trees[5] . Contributing to the creation of machine learning-powered diagnostic tools that can help medical professionals make better judgments regarding patient care is the ultimate objective.

2.Literature Review:

M.A. Islam et al. (2023) conducted a study on the early detection of CKD using machine learning[6] . They worked with a dataset of 400 cases, featuring 24 attributes—13 categorical and 11 numerical. After preprocessing, Principal Component Analysis (PCA) was applied to determine key features for CKD prediction. The XgBoost classifier outperformed other algorithms, reaching an accuracy of 98.33% with the original data and improving to 99.16% after PCA was applied. Other classifiers also achieved an accuracy of 98.33% before PCA.

R. Sawhney et al. (2023) developed AI models to predict and assess CKD, using a dataset with 400 cases and 24 features, both categorical and numerical[7] . They utilized a Multilayer Perceptron (MLP) with backpropagation, integrating two feature extraction and three feature selection techniques to improve

efficiency. The Artificial Neural Network (ANN) model they created outperformed other classifiers, achieving a perfect testing accuracy of 100%, significantly higher than the Logistic Regression (LR)and Support Vector Machine (SVM) scores of 96% and 82%, respectively.

Alsekait et al. (2023) developed an ensemble deep learning model to predict CKD using a dataset of 400 cases with 24 features[8] . The process involved data preprocessing, including label encoding and outlier detection, followed by feature selection through methods like mutual information and Recursive Feature Elimination (RFE). The model used a stacked approach combining RNN, LSTM, and GRU models, with a Support Vector Machine (SVM) for meta-learning. This model achieved high performance metrics, with an accuracy, precision, recall, and F1 score all around 99.69%.

Arif M.S. et al. (2023) designed a machine learning model to predict CKD, incorporating advanced preprocessing, feature selection using the Boruta algorithm, and hyperparameter optimization. Their method involved iterative imputation for missing values and a novel sequential data scaling technique that included robust scaling, z-standardization, and min-max scaling[9] . The model, tested on the UCI CKD dataset with 400 cases and 24 features, achieved an accuracy rate of 100% using k-Nearest Neighbors (KNN) algorithm and grid-search CV for optimization.

K.M. Almustafa (2021) proposed a classification system for kidney diseases using a dataset of 400 cases with 24 features. Various machine learning classifiers were tested, with J48 and Decision Tree (DT) performing best, achieving 99% accuracy[10] . Post-feature selection, these classifiers, along with Naive Bayes and KNN, showed improved accuracy, indicating the effectiveness of feature selection in enhancing model performance.

3.Methodology:

3.1 Data Preprocessing:

- Load the dataset (ckd.csv) and replace missing values (?) with NaN.
- Convert categorical values (e.g., "normal", "abnormal") into binary numerical values using a mapping dictionary[11].
- Fill missing numeric values with the mean of their respective columns to ensure no gaps remain in the dataset[12].

3.2 Dataset Splitting:

- Separate features (X) and target labels (Y).
- Split the dataset into training (70%) and testing (30%) subsets using train_test_split.

Training Decision Tree Classifiers:

Train two decision tree models with different criteria:

- **Gini Index:** A decision tree is trained using the Gini impurity measure.
- **Entropy:** Another decision tree is trained using entropy for information gain.

Model Evaluation:

- Predict CKD outcomes for the test set using the trained models.
- Evaluate models using metrics like:

Confusion Matrix.
Accuracy Score.
Classification Report (precision, recall, F1-score).

Visualization:

- Compare the accuracies of both models (Gini and Entropy) using a bar plot.

3.6 Data Export:

- The cleaned and preprocessed dataset is saved as final.csv for further use.
- The file is made downloadable for external use.

This methodology ensures a complete pipeline from data cleaning to model training, evaluation, and export for Chronic Kidney Disease prediction.

4.Result and Analysis :

The results and analysis for this methodology can be divided into several key areas: **model performance**, **accuracy comparison**, and **evaluating key metrics**. The following steps summarize the results and provide insights:

1.Model Performance

After preprocessing the dataset (handling missing values, converting categorical features to binary, and filling in missing numerical values), two decision tree classifiers are trained: one using the *Gini Index* and the other using *Entropy* as splitting criteria. These models are trained on 70% of the dataset and evaluated on the remaining 30%.

	Precision	Recall	F1-score	support
ckd	0.99	0.96	0.97	80
not ckd	0.93	0.97	0.95	40
Accuracy			0.97	120
Macro avg	0.96	0.97	0.96	120
Weighted avg	0.97	0.97	0.97	120

Table 1. Model Performance metrics Using Gini Index

	Precision	Recall	F1-score	support
ckd	0.99	0.96	0.97	80
not ckd	0.93	0.97	0.95	40
Accuracy			0.97	120
Macro avg	0.96	0.97	0.96	120
Weighted avg	0.97	0.97	0.97	120

Table 2. Model Performance metrics Using Entropy

Confusion Matrix

For each model (Gini and Entropy), the confusion matrix helps in assessing the true positives, true negatives, false positives, and false negatives. This matrix provides insights into how well the model performs on the test set and where the model is making errors.

- *True Positives (TP)*: Correctly predicted CKD positive cases.
- *True Negatives (TN)*: Correctly predicted non-CKD negative cases.
- *False Positives (FP)*: Incorrectly predicted CKD positive cases (Type I Error).
- *False Negatives (FN)*: Incorrectly predicted CKD negative cases (Type II Error).

The confusion matrix will look like this for each model:

[[TN, FP],
[FN, TP]]

Where:

- *TN* is the number of healthy individuals correctly identified.
- *TP* is the number of CKD-positive individuals correctly identified.
- *FP* is the number of non-CKD individuals incorrectly predicted as CKD-positive.

FN is the number of CKD individuals incorrectly predicted as non-CKD.

Accuracy

Accuracy is calculated as:

Accuracy=TP + TN /TP+TN+FP+FN*100

Accuracy gives us the overall performance of the model in terms of correctly classified instances. A higher accuracy indicates better performance, but accuracy alone may not be enough if the dataset is imbalanced.

Classification Report

The classification report provides several important metrics for evaluation:

- *Precision*: The proportion of true positives among all predicted positives.
  Precision=TP/TP+FP
- **Recall (Sensitivity)**: The proportion of true positives among all actual positives.
  Recall=TP/TP+FN
- **F1-Score**: The harmonic mean of precision and recall, which balances both metrics.
  F1=2×Precision+Recall/Precision×Recall

Accuracy Comparison

Following training and prediction, the accuracy of both models is computed and compared using Gini and Entropy. As seen in Figure 1, a bar chart that graphically depicts each model's accuracy is used to illustrate the comparison.

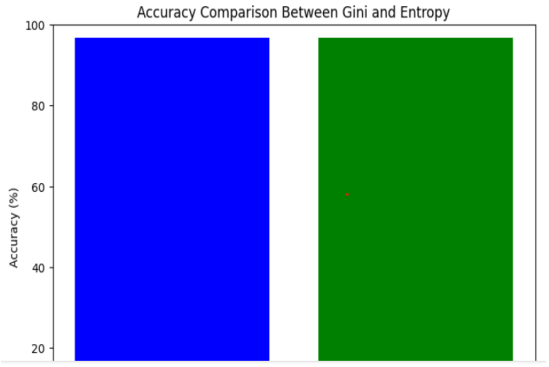

Figure 1 Accuracy comparison graph

- *Gini Index Model*: Gini impurity is focused on minimizing the likelihood of misclassification by splitting the data into homogeneous sets. This model's performance will depend on how well the features can separate CKD from non-CKD cases.
- *Entropy Model*: Entropy, which measures the uncertainty in the data, strives to reduce disorder by creating splits that maximize information gain. It can sometimes perform better when dealing with more complex relationships between features.

The *bar chart* provides a clear comparison between the two algorithms, indicating which criterion (Gini or Entropy) gives better results.

3.Visual Representation

The bar chart comparing the accuracies of Gini and Entropy is a key part of the analysis. It shows:

- *X-axis*: The decision tree criterion used (Gini vs. Entropy).
- *Y-axis*: The accuracy percentage.

This visualization helps quickly assess which model performs better on the dataset.

4.Evaluation of key Metrics

Based on the confusion matrix and classification report, here is an analysis of the key metrics:

- *Precision*: Measures how well the model avoids false positives (wrongly classifying a non-CKD patient as CKD-positive).
- *Recall*: Measures how well the model detects actual CKD cases. A higher recall means fewer false negatives (wrongly classifying a CKD patient as non-CKD).
- *F1-Score*: Balances the precision and recall, providing a more complete measure of a model's performance, especially when the dataset is imbalanced.

By analyzing these metrics, we can get a deeper understanding of the model's ability to identify CKD cases and avoid misclassifications.

5.Future Enhancement:

Future work may involve exploring more complex algorithms, such as random forests or gradient boosting, and conducting hyperparameter tuning to enhance prediction accuracy further. Additionally, integrating clinical data with socio-demographic factors could provide a more comprehensive understanding of CKD risk factors.

To enhance the CKD prediction model, scaling data, handling missing values with advanced imputation, and performing feature selection using techniques like RFE can improve accuracy. Model performance can be boosted through hyperparameter tuning and exploring algorithms like Random Forest or XGBoost. Additionally, handling imbalanced data with methods like SMOTE or class weighting is crucial. Automating workflows using pipelines and deploying the model via web interfaces or integrating it with Electronic Health Records (EHR) systems can increase usability and impact in clinical settings. These improvements ensure better accuracy, reliability, and real-world application of the model.

6. Conclusion:

In order to differentiate between fake and real news, this study employed a machine learning technique implementing logistic regression. Preprocessing methods like TF-IDF vectorization, stop word removal, and punctuation eradication were applied to the dataset, which consisted of labeled false and authentic news articles. The study's main goal was to assess the model's performance with different sample sizes and examine how well it predicted the veracity of news articles. The Logistic Regression model accurately classified news items as either authentic or fake. Accuracy showed consistent performance throughout the data subsets and improved with increasing sample size. Notably, the model demonstrated its effectiveness in resource-constrained contexts by maintaining a respectable predictive performance even with less datasets. The model's accuracy was further confirmed by the classification report, which displayed excellent performance metrics for both the fake and real news categories. Additionally, an external validation dataset was used to evaluate the model, and it accurately determined new statements' authenticity. This illustrates how the model may be applied to actual situations, making it a useful instrument for halting the spread of false information.

REFERENCES :

1. Aljaaf A.J., Al-Jumeily D., Haglan H.M., et al. 2018 IEEE Congress on Evolutionary Computation (CEC) IEEE; 2018. Early prediction of chronic kidney disease using machine learning supported by predictive analytics; pp. 1–9.
2. Almasoud M., Ward T.E. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Soft Comput Appl. 2019;10.
3. Arora M., Sharma E.A. Chronic kidney disease detection by analyzing medical datasets in weka. Int J Comput Mach Learn Algor New Adv Mach Learn. 2016;3:19–48.
4. Banik S., Ghosh A. Prevalence of chronic kidney disease in Bangladesh: a systematic review and meta-analysis. Int Urol Nephrol. 2021;53:713–718. doi: 10.1007/s11255-020-02597-6.
5.Charleonnan A., Fufaung T., Niyomwong T., Chokchueypattanakit W., Suwannawach S., Ninchawee N. 2016 Management and Innovation Technology International Conference (MITicon) IEEE; 2016. Predictive analytics for chronic kidney disease using machine learning techniques. pp. MIT–80.
6.Chen Z., Zhang X., Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int Urol Nephrol. 2016;48:2069–2075. doi: 10.1007/s11255-016-1346-4.
7.Chittora P., Chaurasia S., Chakrabarti P., et al. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access. 2021;9:17312–17334.
8.Gudeti B., Mishra S., Malik S., Fernandez T.F., Tyagi A.K., Kumari S. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) IEEE; 2020. A novel approach to predict chronic kidney disease using machine learning algorithms; pp. 1630–1635.
9.Saringat Z., Mustapha A., Saedudin R.R., Samsudin N.A. Comparative analysis of classification algorithms for chronic kidney disease diagnosis. Bull Elect Eng Inform. 2019;8:1496–1501.
10.Subasi A., Alickovic E., Kevric J. CMBEBIH 2017. Springer; 2017. Diagnosis of chronic kidney disease by using random forest; pp. 589–594.