ARTIFICIAL INTELLIGENCE AND EDGE COMPUTING

Project Report On
Chronic Kidney Disease Classification

Submitted By

GROUP-2

S.N	Name of the students	Registration No.
1.	Mousumi prava Pradhan	2141016419
2.	Spruha Das	2251004004
3.	Pratyusha Mohanty	2141013231
4.	Sudipta Ranjan Tripathy	2141020044

B. Tech. (ECE) 7th Semester (Section–B)

DEPT. OF ELECTRONICS & COMMUNICATION ENGINEERING
Institute of Technical Education and Research

SIKSHA 'O' ANUSANDHAN
DEEMED TO BE UNIVERSITY

Bhubaneswar, Odisha, India.

(February 2024)

Contents

DECLARATION

We certify that

a. The work contained in this report is original and has been doneby us.

b. The work has not been submitted to any other Institute for any degree or diploma.

c. We have followed the guidelines provided by the Department in preparing the report.

d. Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due creditto them by citing them in the text of the report and giving their details in the reference.

Name of the students	**Registration no.**
Mousumi prava Pradhan	2141016419
Spruha Das	2251004004
Pratyusha Mohanty	2141013231
Sudipta Ranjan Tripathy	2141020044

DATE: 30/09/2024
SIKHSHA 'O' ANUSANDHAN, ITER

AIM OF THE PROJECT: -

To develop a Machine Learning model to predict, diagnose, and monitor Chronic Kidney Disease (CKD) using patient data.

MOTIVATION OF THE PROJECT: -

The kidney is one of the most important body organs that filtrates all the wastes and water from human body to make urine.

Chronic Kidney Disease (CKD), also commonly known as chronic renal disease or chronic kidney failure is a life-threatening disease.

It leads to the continuous decrease of Glomerular Filtration Rate (GFR) for a period of 3 months or more and is a universal health problem.

CKD is caused by a variety of underlying factors, including diabetes, high blood pressure and other diseases that damage the kidneys.

Early symptoms of CKD can be subtle and may include fatigue, swelling and decreased urine output which is why it often goes undiagnosed until the later stages.

Early detection and treatment can help for slow the progression of the disease and prevent complications Machine Learning (ML) techniques can be used to predict, diagnose and monitor.

DATASET: - https://www.kaggle.com/datasets/akshayksingh/kidney-disease-dataset

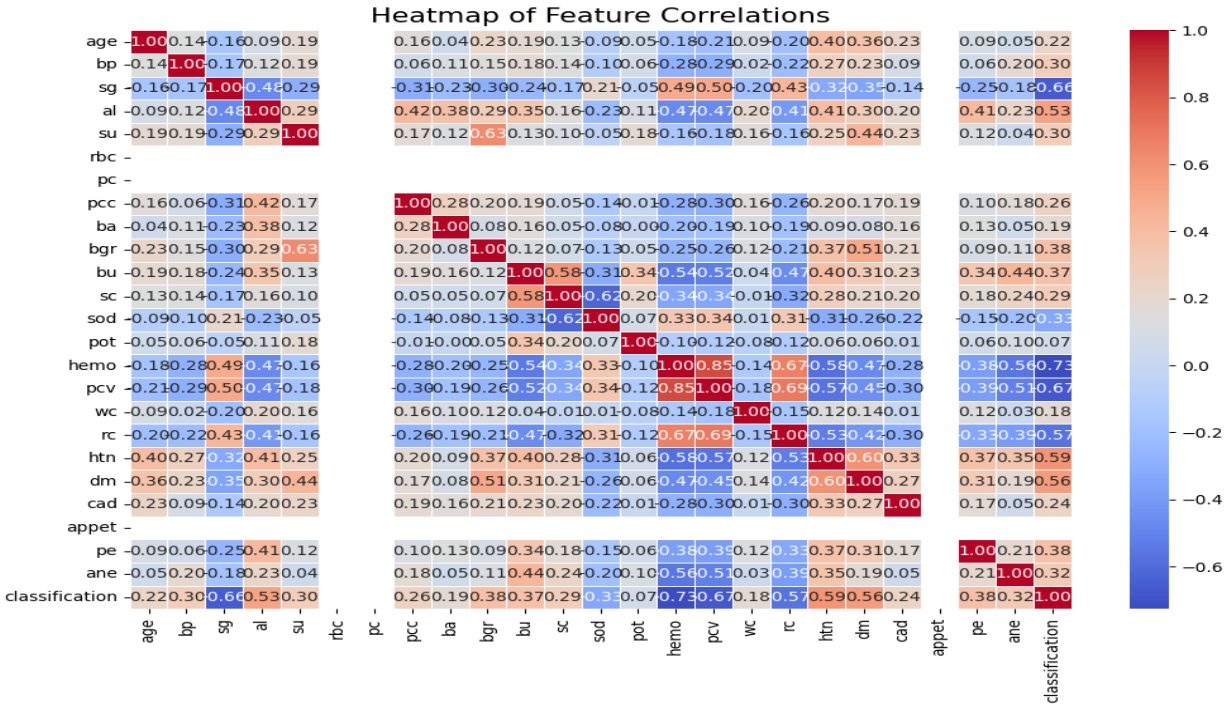
EDA OF THE DATA: -

Null Value Identification

For numerical columns: Impute missing values with the median.

For categorical columns: Impute missing values with the mode.

HEAT MAP ANALYSIS



DATA TYPE MODIFICATION

Convert Categorical Columns:
Convert relevant categorical features to numerical using label encoding or one-hot encoding.
Data Cleaning:
Columns like pcv, wc, and rc should be converted from object to numeric after addressing non-numeric values.

ML MODEL JUSTIFICATION: -

Machine Learning Model Justification

Context and Problem Definition

Chronic Kidney Disease (CKD) is a significant public health concern, impacting millions of individuals globally. Early diagnosis and management are crucial for preventing disease progression and complications. Traditional diagnostic methods often rely on subjective assessments and may not capture the complexities of patient data effectively. To improve diagnostic accuracy, we utilize machine learning techniques to analyze medical attributes and classify patients as having CKD or not.

Why Machine Learning?
Handling Complex Data: The dataset comprises a variety of medical features (e.g., age, blood pressure, blood test results) that can be interdependent and nonlinear. Machine learning algorithms, particularly ensemble methods like Random Forest, can effectively model such complexities.

Scalability: As healthcare systems increasingly digitize patient records, machine learning models can be trained on large datasets, allowing them to generalize well and remain effective as more data becomes available.

Automated Decision-Making: Machine learning models can automate the classification process, reducing the burden on healthcare professionals and increasing the speed of diagnosis. This can be particularly beneficial in resource-limited settings.

Performance Evaluation: Machine learning frameworks provide robust metrics for evaluating model performance, such as accuracy, precision, recall, F1-score, and specificity. These metrics are critical for assessing the effectiveness of the model in real-world scenarios.

Choice of Model: Random Forest Classifier
The Random Forest classifier was selected for the following reasons:

Robustness: Random Forest is less sensitive to overfitting compared to individual decision trees, especially with high-dimensional data. It averages the predictions from multiple trees to improve generalization.

Feature Importance: This algorithm can evaluate the importance of different features in making predictions, offering insights into which medical attributes are most influential in diagnosing CKD.

Flexibility: Random Forest can handle both numerical and categorical data, making it suitable for our diverse dataset.

Good Baseline Performance: Random Forest typically performs well across various datasets and is a strong baseline model for classification tasks.

Evaluation Metrics
To ensure the model is reliable, we used several metrics for evaluation:

Accuracy: Overall correctness of the model's predictions.
Precision: Proportion of true positive predictions among all positive predictions, indicating the model's ability to minimize false positives.
Recall: Proportion of true positive predictions among all actual positive cases, highlighting the model's ability to detect actual CKD cases.
F1-Score: Harmonic mean of precision and recall, providing a balance between the two metrics.
Specificity: Ability of the model to correctly identify negative cases (non-CKD patients), minimizing false negatives.
Conclusion
In summary, using a machine learning approach, particularly the Random Forest classifier, allows for an effective and efficient means of classifying Chronic Kidney Disease. This model leverages the rich dataset of medical features to provide insights that can significantly aid in early diagnosis and treatment planning, ultimately improving patient outcomes. The model's robustness, ability to handle complex relationships, and interpretability make it a suitable choice for this critical health issue.

ML MODEL CODE:

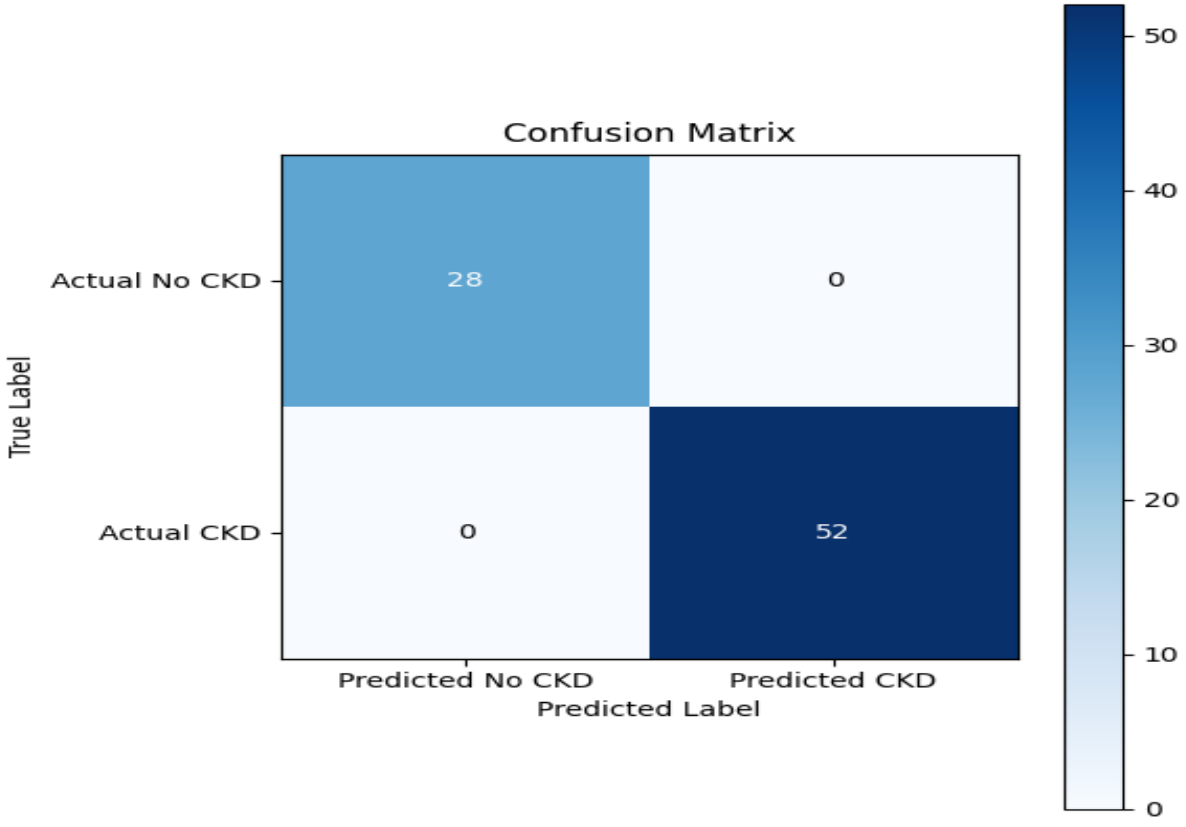
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import precision_score, classification_report, confusion_matrix, accuracy_score, recall_score, f1_score

# Train a RandomForestClassifier
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = rf_model.predict(X_test)

# Calculate basic metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)
tn, fp, fn, tp = conf_matrix.ravel()

# Calculate Specificity
specificity = tn / (tn + fp)

# Display results in percentage format
print(f"Accuracy: {accuracy * 100:.2f}%")
print(f"Precision: {precision * 100:.2f}%")
print(f"Recall: {recall * 100:.2f}%")
print(f"F1-Score: {f1 * 100:.2f}%")
print(f"Specificity: {specificity * 100:.2f}%")

# Confusion matrix interpretation
print("\nConfusion Matrix:")
print(conf_matrix)
```

```
print(f"True Negatives: {tn}")
print(f"False Positives: {fp}")
print(f"False Negatives: {fn}")
print(f"True Positives: {tp}")
```

METRICES:-

Accuracy: 100.00%
Precision: 100.00%
Recall: 100.00%
F1-Score: 100.00%
Specificity: 100.00%

Confusion Matrix:
[[28 0]
 [0 52]]
True Negatives: 28
False Positives: 0
False Negatives: 0
True Positives: 52



INFERENCES:-

Model Overview

The Random Forest classifier was utilized to predict Chronic Kidney Disease (CKD) based on various health indicators.

Performance Summary

The model demonstrated exceptional performance, achieving perfect scores in accuracy, precision, recall, F1-score, and specificity. The confusion matrix confirmed that all CKD and non-CKD cases were correctly classified, with no false positives or negatives.

Conclusion

This model shows high reliability in predicting CKD, making it an effective tool for clinical decision support. Future efforts should focus on validating the model with independent datasets to ensure its robustness in real-world applications.

Scope Of Enhancement:-

The scope of enhancement for this Chronic Kidney Disease (CKD) prediction project could include:

Feature Engineering:

Create new features derived from existing medical data (e.g., interaction terms between blood pressure and age, or cumulative risk scores).
Include time-series data to track progression of kidney function over time for better predictions.

Model Improvement:

Experiment with different ML algorithms such as Random Forest, Gradient Boosting, or Neural Networks to improve prediction accuracy.
Use ensemble methods to combine predictions from multiple models.
Tune hyperparameters to optimize model performance.

Data Expansion:

Incorporate more diverse and larger datasets from different populations, including longitudinal data, to improve generalizability.
Include additional biomarkers and genetic data that could further enhance CKD prediction accuracy.

Handling Imbalanced Data:

Use techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting to handle imbalanced datasets, improving performance on minority classes (e.g., early CKD stages).

Real-time Monitoring:

Develop models that can be integrated into healthcare systems for real-time monitoring of CKD patients, providing timely alerts to clinicians.

Explainability:

Incorporate explainable AI techniques such as SHAP values or LIME to interpret the model's predictions, making it easier for healthcare professionals to understand and trust the model.
User-Friendly Application:

Build a user-friendly interface or dashboard for healthcare providers to easily input patient data, view predictions, and monitor disease progression.

Integration with Wearables/IoT:

Integrate data from wearable devices (e.g., blood pressure monitors, glucose trackers) for real-time health monitoring and improved predictive accuracy.

Cross-validation and External Validation:

Test the model across various geographic and demographic groups to ensure the model's robustness and generalizability to different populations.
These enhancements could significantly improve the model's accuracy, usability, and clinical impact.