# CHRONIC KIDNEY DISEASE DETECTION USING ENSEMBLE LEARNING

Dhanush Uppala pati

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

# Introduction:

The project, titled "Chronic Kidney Disease Detection Using Ensemble Learning," focuses on predicting Chronic Kidney Disease with high accuracy using machine learning. By employing algorithms such as K-Nearest Neighbor, Naïve Bayes, and others, alongside advanced ensemble techniques, the study achieves around 99% prediction accuracy. This work underscores the potential of applying machine learning to enhance healthcare diagnostics with a streamlined set of features.

# Reference Paper:

Title: Chronic Kidney Disease Detection using AdaBoosting Ensemble Method and K-Fold Cross Validation

Author: N.Mohana Suganthi, Jemin V.M, P.Rama

Year: 07 February 2023

Journal Name: **IEEE *Xplore*** *(*Click here for Reference paper*)*

Why I choose this paper:I selected this paper for its detailed exploration of AdaBoost ensemble learning and K-Fold Cross Validation in CKD detection, aligning with my project's aim for high predictive accuracy. It offers practical insights into enhancing machine learning models for healthcare, demonstrating significant success with a 99% accuracy rate using AdaBoost-Random Forest, which directly informs and supports the methodologies applied in my work.

Models used in this project: The paper utilizes AdaBoost with ensemble methods like Random Forest for Chronic Kidney Disease prediction, employing K-Fold Cross Validation for accuracy assessment.

Other References:

1.Detection of chronic kidney disease by using ensemble classifiers|IEEE
2.Chronic kidney disease prediction using boosting techniques based on clinical parameters|PLOS ONE

# Problem Statement

This initiative seeks to refine the prediction methods for Chronic Kidney Disease (CKD) by employing advanced ensemble learning methods. CKD is a critical health issue marked by a slow but steady decline in kidney functionality, which often escapes early detection due to the lack of pronounced symptoms. The project's ambition is to harness a combination of sophisticated machine learning techniques to construct a highly accurate predictive model. This model is expected to identify CKD with exceptional precision, utilizing a concise array of indicators, thereby contributing to improved early detection and management strategies in medical practice.

# Dataset Overview:

- The dataset on kidney disease comprises 400 entries with 26 attributes each, blending clinical, demographic, and laboratory data pertinent to kidney function and disorders. Attributes include patient age, blood pressure, specific gravity of urine, levels of albumin, sugar, red and pus cells, hemoglobin, and more, alongside the presence of conditions like hypertension, diabetes, and anemia. It features both numerical and categorical data, with some missing values across various columns. Primarily, it categorizes patients into those with chronic kidney disease (CKD) and those without, serving as a valuable resource for analyzing kidney health indicators and potentially developing predictive models for CKD.
- Features include:
  Age
  Blood pressure
  Specific gravity of urine
  Albumin levels
  Sugar levels
  Red and pus cells in the urine
  Hemoglobin count
- Additionally, it notes the presence of conditions such as: Hypertension Diabetes Anemia

Click here for Dataset

# Motivation

The project is driven by the objective to enhance the precision in diagnosing Chronic Kidney Disease (CKD) through sophisticated ensemble machine learning models. The focus is on early and reliable detection due to CKD's subtle onset, aiming to prevent its progression to critical stages. By targeting a near-perfect accuracy with fewer diagnostic indicators, the initiative aspires to equip medical professionals with a powerful tool for better patient care, promoting timely interventions and improving health outcomes for those at risk of CKD.

# Summary of the project

Objective: Enhance CKD detection accuracy using machine learning.

Method: In this study, a variety of machine learning techniques were utilized to develop models for the accurate prediction of Chronic Kidney Disease. These techniques include the K Nearest Neighbour (KNN), Naïve Bayes (NB), Linear Discriminate Analysis (LDA), Light Gradient-Boosting Machine (LGBM), AdaBoost Classifier (ABC), Decision Tree (DT), Logistic Regression (LR), and Multi-Layer Perceptron (MLP). Each of these algorithms was employed to train distinct models, aiming to ensure the reliability of CKD predictions.

Rationale: Combines multiple algorithms for improved accuracy, compensates for missing data, and leverages ensemble learning to strengthen prediction capability.

Outcome: Achieved approximately 99% accuracy, demonstrating the effectiveness of ensemble methods in medical diagnosis applications. To further enhance prediction accuracy, we adopted a cross-validation strategy using the K-Fold Stratified method. Following this, we employed an ensemble modeling approach, integrating the strengths of various algorithms to bolster the overall performance of our CKD prediction model.

# Timeline

1. March 1, 2024: Literature review - Complete analysis of related studies.
2. March 8, 2024: Data acquisition - Finish collecting necessary datasets.
3. March 15, 2024: Data cleaning - Begin preprocessing and cleaning of data.
4. March 22, 2024: Initial model development - Start training with selected ML techniques.
5. March 29, 2024: Model refinement - Optimize parameters and algorithms.
6. April 5, 2024: Cross-validation implementation - Apply K-Fold Stratified validation.
7. April 12, 2024: Ensemble model integration - Combine models for improved accuracy.
8. April 19, 2024: Performance evaluation - Assess model effectiveness and tweak as necessary.
9. April 26, 2024: Final adjustments - Make last-minute adjustments based on feedback and testing results.

This timeline provides a clear structure for progressing through the project phases on a weekly basis, ensuring a comprehensive approach to achieving the project's objectives.