Supervised Learning Classifiers and Data Preprocessing on Chronic Kidney Disease Dataset

Table of Contents

1. Introduction

Chronic Kidney Disease (CKD) is a major global health concern, with early diagnosis playing a pivotal role in improving patient outcomes. This study investigates the Chronic Kidney Disease dataset, focusing on applying machine learning techniques to predict the presence of CKD. The dataset includes both numerical and categorical features, which will be preprocessed and used for clustering and classification tasks. Various supervised learning algorithms, including ensemble methods, will be evaluated to determine the best approach for CKD prediction.

2. Dataset Overview

The Chronic Kidney Disease dataset comprises 25 features that capture various health metrics. These features include both continuous variables, such as age and blood pressure, and categorical variables, such as hypertension status and diabetes. A summary of the features is provided below:

Column	Description
id	Unique identifier for each patient.
age	Age of the patient.
bp	Blood pressure readings.
sg	Specific gravity of urine, indicating kidney function.
al	Albumin levels in urine, another kidney function indicator.
su	Sugar levels in urine, related to diabetes.
rbc	Red blood cell count.
pc	Pus cell count, indicating infection.
pcc	Pus cell clumps, indicating infection.
ba	Bacteria presence.
pcv	Packed cell volume, related to anemia.
wc	White blood cell count, indicating immune response.
rc	Red blood cell count.
htn	Hypertension status.

Column	Description
dm	Diabetes mellitus status.
cad	Coronary artery disease status.
appet	Appetite status.
pe	Presence of pedal edema.
ane	Anemia status.
classification	Final diagnosis (target variable): 0 = no CKD, 1 = CKD.

3. Data Preprocessing

3.1. Handling Missing Values

Missing values were imputed using the KNNImputer for numerical features, and the mode imputation was applied to categorical features. This ensures the integrity of the data, minimizing any loss of useful information.

3.2. Data Splitting

The data was split into numerical and categorical columns. Numerical columns underwent MinMax scaling, ensuring that all values fall within a normalized range between 0 and 1. Categorical features were encoded using LabelEncoder to convert them into numerical values.

3.3. Handling Class Imbalance

To address class imbalance in the target variable, the RandomOverSampler method was used, resulting in an equal number of instances for both the classes (CKD and non-CKD).

3.4. Dimensionality Reduction

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset from 24 to 18 features, retaining a majority of the dataset's variance.

4. Clustering and Unsupervised Learning

4.1. KMeans Clustering

KMeans clustering was performed, and the optimal number of clusters was determined using the elbow method. The results indicated that the best number of clusters is 2, which corresponds to the two distinct classes in the target variable.

- Silhouette Score: 0.3844
- Davies-Bouldin Index: 1.3125

4.2. 4.2 DBSCAN Clustering

DBSCAN was applied as a density-based clustering algorithm. The results were less effective than KMeans in separating the two classes.

- Silhouette Score: -0.1184
- Davies-Bouldin Index: 1.3914

4.3. 4.3 Clustering Evaluation

KMeans showed better performance in creating well-separated clusters compared to DBSCAN, which failed to define distinct clusters in the dataset.

5. Feature Selection

Feature selection was conducted using Random Forest as an estimator. The most influential features identified were age, specific gravity (sg), albumin (al), red blood cell count (rbc), and packed cell volume (pcv). These features were retained for training the supervised learning models.

6. Supervised Learning Classifiers

The following classifiers were applied to the preprocessed dataset:

1. Logistic Regression (Linear Classifier)
2. K-Nearest Neighbors (KNN) (Non-linear Classifier)
3. Support Vector Classifier (SVC) (Non-linear Classifier)
4. Random Forest (Ensemble Method)
5. AdaBoost (Boosting Algorithm)

7. Results and Evaluation

The performance of each classifier was evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Below are the results:

Classifier	**Accuracy**	**Precision (0)**	**Recall (0)**	**F1-Score (0)**	**Precision (1)**	**Recall (1)**	**F1-Score (1)**	**Confusion Matrix**
Logistic Regression	0.98	1.00	0.96	0.98	0.96	1.00	0.98	[[44 2] [0 54]]
K-Nearest Neighbors (KNN)	0.98	1.00	0.96	0.98	0.96	1.00	0.98	[[44 2] [0 54]]
Support Vector Classifier (SVC)	0.98	1.00	0.96	0.98	0.96	1.00	0.98	[[44 2] [0 54]]
Random Forest	1.00	1.00	1.00	1.00	1.00	1.00	1.00	[[46 0] [0 54]]
AdaBoost	1.00	1.00	1.00	1.00	1.00	1.00	1.00	[[46 0] [0 54]]

8. Conclusion

This study demonstrates that machine learning techniques, particularly ensemble methods like Random Forest and AdaBoost, are highly effective in predicting Chronic Kidney Disease based on a diverse set of health indicators. The Random Forest and AdaBoost classifiers achieved perfect accuracy, showing the importance of feature selection and preprocessing in improving model performance. Overall, the integration of various data preprocessing and feature extraction methods contributed significantly to the success of the analysis.

9. References

1. Kaggle: Chronic Kidney Disease Dataset
2. Scikit-learn Documentation: https://scikit-learn.org/stable/
3. PCA Tutorial