ARIEL UNIVERSITY

Chronic Kidney Disease (CKD) Prediction Using Machine Learning

Industrial Engineering and Management

Machine Learning in Healthcare

Dr. Orit Raphaeli

Amit Gido (315754606)

Bar Mizrahi (315042994)

Sahar Sokolik (205937287)

Sarah Hazziza (337891675)

Table of contents

Abstract **3**
Keywords **3**
Introduction **3**
Methodology **4**
Data Collection and Exploration: 4
Data Sources 4
Exploratory Data Analysis (EDA) 4
Data Preprocessing 5
Handling Missing Values 5
Normalization and Standardization 5
Model Building 6
Feature Selection 6
Train-Test Split 7
Data Balancing 7
Algorithm Selection 8
Performance Evaluation 8
Evaluation Metrics 8
Subgroup Analysis 9
Results **10**
Best Overall Model 10
Performance model by metrics 10
Explanatory AI 11
Subgroup Analysis 12
Conclusions **12**
Appendices **13**
Bibliography **14**

Abstract

The project aims to predict Chronic Kidney Disease (CKD) using supervised machine-learning techniques.

By leveraging patient health data, various machine learning models were developed and evaluated to determine their effectiveness in predicting CKD. The study involved data exploration, preprocessing, feature selection, model training, and evaluation.

The results indicate that supervised machine learning can predict CKD with good accuracy, demonstrating significant potential for these methods in clinical settings to aid in early diagnosis and treatment planning.

Keywords

Chronic Kidney Disease (CKD), Supervised machine learning, Patient health data, Predictive models, Early diagnosis.

Introduction

Chronic Kidney Disease (CKD) is a progressive condition characterized by the gradual loss of kidney function, impacting millions of individuals globally. Stages 3-5 CKD represent more advanced and severe reductions in kidney function, often necessitating intensive medical intervention and significantly affecting patient health. Early detection and intervention are critical for managing CKD and improving patient outcomes.

The main question is how to leverage advanced machine learning techniques to predict the risk of developing stage 3-5 CKD and improve early intervention strategies.

The primary purpose of this project is to develop predictive models using supervised machine learning to assess the risk of stage 3-5 CKD based on patient health data.

By analyzing various health metrics and employing different machine learning techniques, this work seeks to provide a robust tool for early detection, which can be integrated into clinical practice to enhance patient care and management.

Methodology

Data Collection and Exploration:

Data Sources

The dataset was obtained from a medical database. Initially, it contained 24 columns and 491 rows, 56 patients were sick with stage 3 to 5 CKD (the result variable). The columns included a comprehensive range of health indicators, in addition to the result variable column EventCKD35, including:

- Demographic details - Gender, AgeBaseline, Age.3.categories (0 – under 50, 1 – from 50 to 65, 2 – above 65).
- Patient's history - HistoryDiabetes, HistoryCHD (Coronary Heart Disease), HistoryVascular, HistorySmoking, HistoryHTN (Hypertension), HistoryDLD (Dyslipidemia), HistoryObesity.
- Medication - DLDmeds, Dmmeds (Diabetes), HTNmeds, ACEIARB (inhibitors).
- Clinical measurements - CholesterolBaseline, TriglyceridesBaseline, HgbA1C, CreatnineBaseline, eGFRBaseline, sBPBaseline, dBPBaseline, BMIBaseline.
- Others: TimeToEventMonths (time from follow-up starts to severe CKD event or last visit).

Exploratory Data Analysis (EDA)

Initial data exploration was performed to understand the distribution of the data, identify outliers, and detect missing values. Visualization tools such as histograms and box plots were used to gain insights into the data structure and relationships between variables.

Data Preprocessing

Handling Missing Values

The first step in handling missing values was to identify the columns with such gaps. The column TriglyceridesBaseline had 5 missing values. Since none of these rows had the value of 1 in the result variable column, we decided to drop these rows with missing values to maintain data integrity.

Next, the column HgbA1C was considered. Two rows with missing values in this column had the value of 1 in the EventCKD35 column. Removing these rows would result in losing critical data about sick patients, which is undesirable due to the limited information on this subgroup. Therefore, imputation was necessary, to decide on the appropriate technique, we first checked if the HgbA1C data was normally distributed, however as shown in Figure 1, it wasn't. Consequently, we opted for a linear regression imputation approach, using BMIBaseline, AgeBaseline, and HistoryDiabetes as predictor variables. [1]

After addressing the missing values, the dataset contained 485 rows.

Normalization and Standardization

To prepare the dataset for modeling, standardization was applied to the numerical columns to ensure that these features had a mean of zero and a standard deviation of one. This step is crucial for machine learning algorithms, as it helps to stabilize the numerical range and improve model performance by ensuring that all features contribute equally to the distance measurements. Additionally, ordinal encoding was applied to a categorical column (Age.3.categories) with values 0, 1, and 2, preserving the right order of the categories. These preprocessing steps were essential in preparing the data for effective modeling and ensuring that the algorithms could learn from the features without being affected by differing scales or categorical hierarchies.

Model Building

Feature Selection

The feature selection process involved several steps to ensure that only the most relevant features were used for modeling. The first step was to drop columns based on the correlation heatmap and Variance Inflation Factor (VIF). The correlation heatmap helps identify highly correlated features, which can lead to multicollinearity, while VIF quantifies the severity of multicollinearity by measuring how much the variance of a regression coefficient is inflated due to linear dependence with other predictors. Based on these methods, the following columns were removed: HistoryHTN, HTNmeds, CreatnineBaseline, Age.3.categories, Gender, DLDmeds and DMmeds.

The second step was to evaluate feature importance using mutual information scores. Mutual information measures the dependency between two variables and helps identify features with the highest predictive power. Columns with a mutual information score of zero, indicating no predictive power, were subsequently removed: BMIBaseline, TriglyceridesBaseline, CholesterolBaseline and sBPBaseline.

The third step involved Recursive Feature Elimination (RFE), a wrapper technique that recursively removes the least important features and builds models on the remaining features. This process continues until the optimal number of features is selected, enhancing model performance by eliminating redundant features. The optimal columns were: AgeBaseline, HistoryDiabetes, ACEIARB, HgbA1C, eGFRBaseline, dBPBaseline, TimeToEventMonths.

Finally, Elastic Net regularization was applied. Elastic Net combines the penalties of Lasso (L1) and Ridge (L2) regression to select features while handling multicollinearity and maintaining model accuracy. An L1 ratio of 0.1 was found optimal, indicating that the model primarily applies Ridge (L2) regularization but incorporates a small component of Lasso (L1) regularization. This comprehensive feature selection process ensured that only the most relevant and impactful features were used in the final models,

leading to more robust and reliable predictions. The selected columns are: AgeBaseline, HistoryDiabetes, ACEIARB, HgbA1C, eGFRBaseline, dBPBaseline, TimeToEventMonths.
After performing RFE and ElasticNet the following columns were removed: HistoryCHD, HistorySmoking, HistoryDLD, HistoryObesity, HistoryVascular. Therefore, the final selected features for the model were: AgeBaseline, HistoryDiabetes, ACEIARB, HgbA1C, eGFRBaseline, dBPBaseline, TimeToEventMonths.

Train-Test Split

The dataset was divided into training and testing sets using an 80-20 split to evaluate model performance on unseen data. Stratification was applied during the split to ensure that the proportion of the target class, indicated by the EventCKD35 column, was maintained in both the training and testing sets. This step was particularly important due to the class imbalance, it helps to preserve the original class distribution, improving the reliability and validity of the model evaluation.

Data Balancing

To address the class imbalance in the training data and improve model performance, two techniques were applied only to the training data to avoid overfitting and ensure that the model generalizes well to unseen data.

SMOTE: Synthetic Minority Over-sampling Technique was used to generate synthetic samples for the minority class. By creating new, synthetic instances of the minority class (sick individuals in the EventCKD35 column), SMOTE helps to balance the class distribution.

Random Under-Sampling: This technique involves reducing the number of samples in the majority class (non-sick individuals) to balance the class distribution.

Algorithm Selection

Various machine learning algorithms were tested:

- Logistic Regression: A simple yet effective model for binary classification.
- Random Forest: An ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction.
- Gradient Boosting: An ensemble technique that builds models sequentially to correct errors made by preceding models.
- Support Vector Machine (SVM): Effective in high-dimensional spaces and used for classification tasks.

To optimize performance, hyperparameter tuning was performed for each algorithm using GridSearchCV and k-fold cross-validation across three types of training data: regular, SMOTE, and under-sampled. After determining the best hyperparameters, models were evaluated on each data version.

Performance Evaluation

Evaluation Metrics

The models were evaluated using a range of performance metrics, including:

- AUC-ROC (Area Under the Receiver Operating Characteristic Curve): Measures the ability of the model to distinguish between classes.
- F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both.
- Precision and Recall: Precision measures the accuracy of positive predictions, while recall measures the ability to capture all positive instances.
- ROC and Precision-Recall Curves: Visual tools to assess model performance across different threshold settings.
- Confusion Matrix: Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives.

Subgroup Analysis

In addition to the primary analysis, a subgroup analysis was performed to explore whether focusing on a homogeneous subset of the data could enhance model performance. This process involved several key steps:

1. Defining Thresholds and Testing Significance: Thresholds were established to identify a homogeneous group within the dataset, specifically for CreatnineBaseline (≥60) and HgbA1C (≥6). The significance of these thresholds was evaluated using T-tests to ensure that the identified subgroup was statistically distinct.
2. Visualizing Group Distributions: The distributions of features within the subgroup were visualized to assess homogeneity. This step was crucial for confirming that the subgroup was uniform in terms of its characteristics (Figure 2).
   In the subgroup the number of patients remaining is 176, from them 43 are sick.
3. Model Training and Evaluation: For the identified homogeneous subgroup, all previous steps—excluding feature selection—were repeated. Class imbalance with under-sampling was excluded as well, due to insignificant results.

The aim was to determine if concentrating on a more homogeneous group would improve the model's predictive performance for CKD. By comparing the results from the subgroup analysis with the original dataset, insights were gained into whether focusing on a smaller, more uniform group could enhance the accuracy and effectiveness of CKD prediction models.

Results

The performance of various machine learning models was evaluated across three different datasets—SMOTE, Undersampling, and Regular—and for a homogeneous subgroup.

Best Overall Model

The Random Forest model with SMOTE preprocessing emerged as the best overall model, achieving an AUC-ROC of 0.90, an F1 Score of 0.62, a Recall of 0.82, and a Precision of 0.50. High recall is essential in a medical context as it minimizes the number of false negatives, ensuring that patients who have chronic kidney disease are correctly identified, which is crucial for timely and effective treatment (figure 3).

Performance model by metrics

	SMOTE	UNDERSAMPLING	REGULAR
LOGISTIC REGRESSION	AUC-ROC: 0.88 F1 Score: 0.50 Recall: 0.82 Precision: 0.36	AUC-ROC: 0.87 F1 Score: 0.30 Recall: 0.91 Precision: 0.18	AUC-ROC: 0.89 F1 Score: 0.47 Recall: 0.36 Precision: 0.67
RANDOM FOREST	AUC-ROC: 0.90 F1 Score: 0.62 Recall: 0.82 Precision: 0.50	AUC-ROC: 0.84 F1 Score: 0.44 Recall: 0.73 Precision: 0.32	AUC-ROC: 0.89 F1 Score: 0.53 Recall: 0.45 Precision: 0.62
GRADIENT BOOSTING	AUC-ROC: 0.90 F1 Score: 0.42 Recall: 0.45 Precision: 0.38	AUC-ROC: 0.84 F1 Score: 0.39 Recall: 0.73 Precision: 0.27	AUC-ROC: 0.85 F1 Score: 0.50 Recall: 0.45 Precision: 0.56
SVM	AUC-ROC: 0.80 F1 Score: 0.34 Recall: 0.45 Precision: 0.28	AUC-ROC: 0.16 F1 Score: 0.44 Recall: 0.82 Precision: 0.30	AUC-ROC: 0.88 F1 Score: 0.00 Recall: 0.00 Precision: 0.00

Explanatory AI

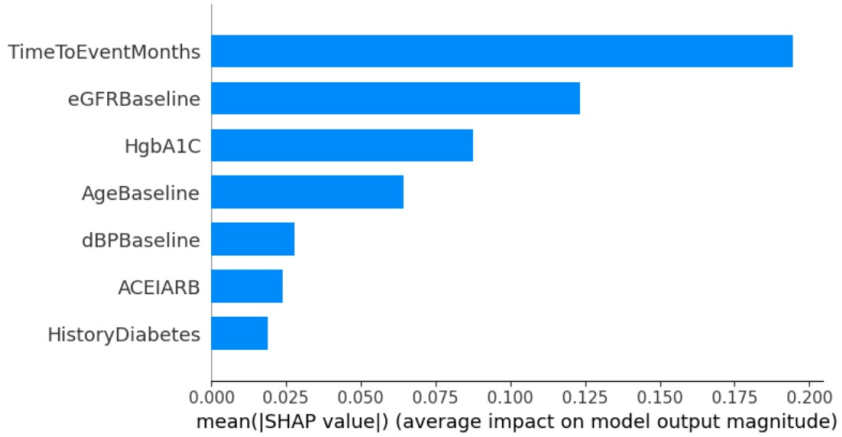

Figure 4

The SHAP summary plot highlights the importance of various features in the CKD predictive model:

- TimeToEventMonths: Most influential predictor, indicating the significance of the time until a relevant CKD event.
- eGFRBaseline: Second most important, reflecting kidney function, with lower values indicating higher CKD risk.
- HgbA1C: Third in importance, showing the impact of blood glucose levels on CKD progression.
- AgeBaseline: Older age contributes significantly to CKD risk.
- dBPBaseline: Elevated diastolic blood pressure is a notable risk factor for CKD.
- ACEIARB: Use of these medications moderately affects CKD outcomes.
- HistoryDiabetes: Presence of diabetes in medical history, while the least impactful among the features listed, still plays a role in CKD risk.

This plot underscores the importance of clinical and demographic factors in predicting CKD outcomes, aiding in identifying high-risk patients for targeted interventions.

Subgroup Analysis

The Random Forest model with SMOTE preprocessing also performed well on the homogeneous subgroup, achieving an AUC-ROC of 0.86, an F1 Score of 0.70, a Recall of 0.78, and a Precision of 0.64 (Figure 3).

Overall, the results demonstrate that the Random Forest model with SMOTE preprocessing is consistently the best performer across different datasets and metrics. However, according to the performance metrics the model performed better on the overall data rather than on this subgroup.

Conclusions

The study developed and evaluated several machine learning models for predicting chronic kidney disease (CKD). While the models demonstrated potential, the performance varied, indicating that further refinements are necessary. The results highlight the importance of data quality and the need for larger datasets to improve predictive accuracy.

Supervised machine learning techniques have shown some effectiveness in predicting CKD, with models achieving reasonable performance. However, the results suggest that current models could be enhanced for better accuracy and reliability. These models offer promise for aiding early detection and management of CKD, though further improvements are needed to fully leverage their potential in clinical practice.

Future research could focus on exploring alternative machine learning approaches, such as deep learning and reinforcement learning, to uncover more complex patterns and enhance prediction accuracy. Expanding the dataset with additional and diverse sources of health information might also improve model performance. Furthermore, developing user-friendly tools for clinical integration, while addressing ethical considerations, will be crucial for the broader adoption and effective use of machine learning models in healthcare settings.

Appendices

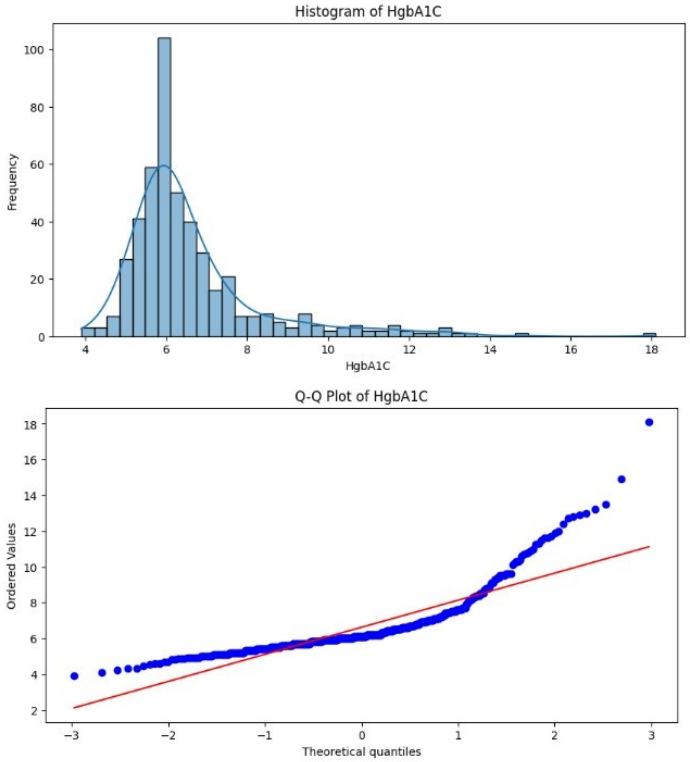

Figure1

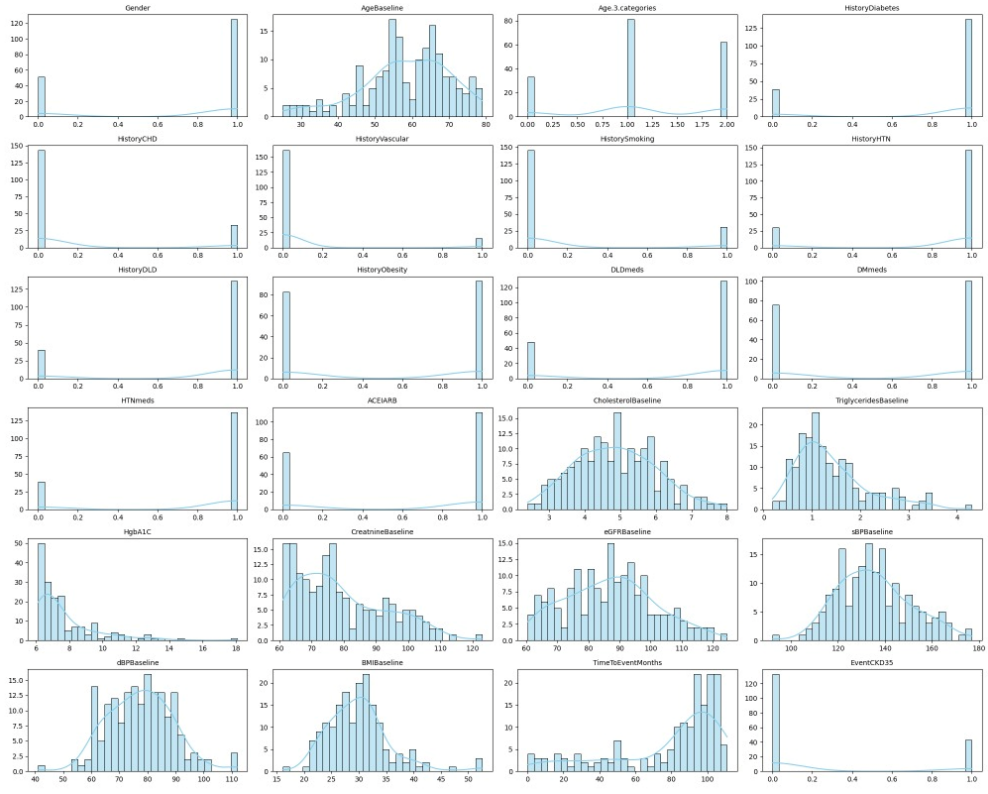

Figure 2

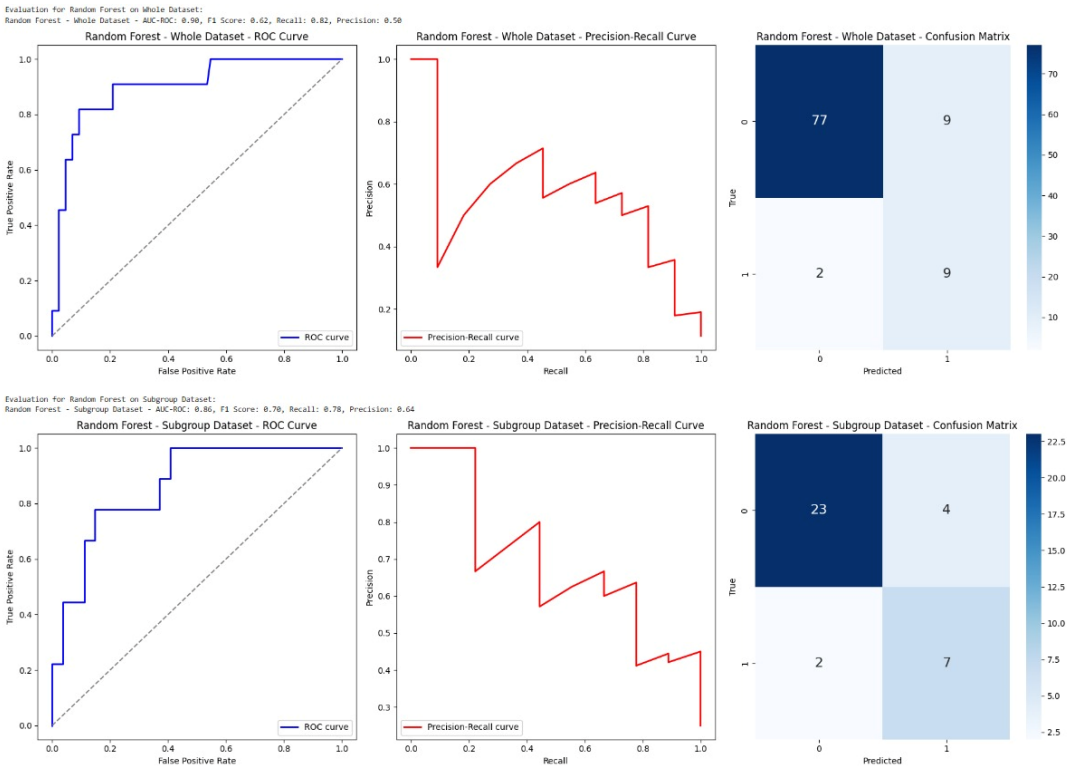

Figure 3

Bibliography

[1] Lin, W. The Association between Body Mass Index and Glycohemoglobin (HbA1c) in the US Population's Diabetes Status. Int. J. Environ. Res. Public Health 2024, 21, 517. https://doi.org/10.3390/ijerph21050517