10/07/2024

**Lecturer's name: Dr. Orit Rafali**

PRESENTERS: MeiTAL PADwA, YAEL dickstein, SHIR SAADI, ORIYA SELA

**Final report**

Predicting Chronic Kidney disease (CKD) using machine learning

# Table of content

[Table of content 1](#_Toc171504944)

[1. Summary 2](#_Toc171504945)

[2. A brief introduction and the purpose of the work 2](#_Toc171504946)

[introduction 2](#_Toc171504947)

[literature overview 2](#_Toc171504948)

[3. Methodology 3](#_Toc171504949)

[Data Collection 3](#_Toc171504950)

[The database 3](#_Toc171504951)

[Data Description**:** 3](#_Toc171504952)

[Data Preprocessing and Data Visualization-EDA 3](#_Toc171504953)

[Visualization techniques applied in the project are as followed: 4](#_Toc171504954)

[Insights and observations: 4](#_Toc171504955)

[Feature Selection and class imbalance 4](#_Toc171504956)

[Techniques Used**:** 4](#_Toc171504957)

[Results for feature selection: 5](#_Toc171504958)

[Class imbalance 5](#_Toc171504959)

[Model Development 6](#_Toc171504960)

[Model Evaluation and tuning 6](#_Toc171504961)

[evaluation**:** 6](#_Toc171504962)

[Hyperparameter Tuning**:** 7](#_Toc171504963)

[conclusions: 7](#_Toc171504964)

[EXplanatory AI Techniques 7](#_Toc171504965)

[Subgroup Analysis 8](#_Toc171504966)

[Homogeneous Groups Identification**:** 8](#_Toc171504967)

[Model Development for Subgroups**:** 8](#_Toc171504968)

[4. Results and conclusions 10](#_Toc171504969)

[The effectiveness of ML for CKD Prediction 10](#_Toc171504970)

[research ideas including usage of different ML methods 10](#_Toc171504971)

[5. appendix 12](#_Toc171504972)

[6. Bibliography 15](#_Toc171504973)

# Summary

In this project, we developed a machine learning-based Chronic Kidney Disease Prediction (ML-CKDP) model aimed at enhancing dataset preprocessing and optimizing the prediction of CKD. Our proposed model incorporates a comprehensive data preprocessing protocol that includes converting categorical variables to binary values, imputing missing data, and normalizing through robust scaling.

To extract meaningful insights, we performed data visualization and employed a variety of feature selection techniques, such as Correlation Analysis, Chi-Square tests, ANOVA tests, and Recursive Feature Elimination, to refine the datasets. The model leverages four classifiers—Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT)—to predict CKD.

We applied cross-validation and hyperparameter tuning to optimize the model's performance. The effectiveness of the models was assessed by measuring their confusion matrix statistics, generating (ROC) curves, and calculating the Area Under the Curve (AUC). Notably, the Random Forest and Support Vector Machine classifiers achieved the highest rates.

To enhance model interpretability, we employed explanatory AI techniques such as LIME and SHAP. These methods clarify the importance and effects of features, providing transparency and helping understand the model's predictions.

Finally, we conducted a Subgroup analysis to identify homogeneous groups within the population, refining the model development for one specific homogenic group. This subgroup analysis revealed patterns that further informed our predictive model, enhancing its applicability and precision.

# A brief introduction and the purpose of the work

## introduction

Chronic Kidney Disease (CKD) is a significant public health issue affecting millions globally, characterized by the gradual loss of kidney function. Early detection is essential for slowing disease progression and improving patient outcomes. Traditional diagnostic methods can be invasive, time-consuming, and costly.

Advancements in machine learning offer the potential for more efficient, non-invasive, and accurate predictions. This project aims to use machine learning techniques to predict early-stage CKD (CKD35) using patient health records.

By applying and evaluating various algorithms, **we seek to develop a predictive model for early CKD identification**, improving treatment.

This report outlines the methodology, data preprocessing steps, model selection, evaluation metrics, and results, providing a comprehensive overview of using machine learning for CKD prediction.

## literature overview

To understand the significance of developing a machine learning model for CKD detection and to gain insight into the medical condition, we reviewed relevant literature.

A notable paper, "ML-CKDP: Machine Learning-Based Chronic Kidney Disease Prediction with Smart Web Application," provides an extensive study on developing a machine learning model for CKD prediction. This study emphasizes improved dataset preprocessing and the creation of a web-based application for efficient CKD prediction.

CKD is characterized by a gradual decline in kidney function, often leading to severe complications such as hypertension, anemia, and renal failure. The disease progresses slowly and is typically asymptomatic, resulting in late-stage diagnoses. CKD has seen a significant rise in global prevalence, affecting over 500 million individuals, particularly in developing regions. Early detection and management are crucial to mitigate the economic and health impacts, especially where advanced treatments like Renal Replacement Therapy (RRT) are often inaccessible.

The paper's model involves comprehensive data preprocessing, feature selection, and employs various classifiers for CKD prediction:

1. **Data Preprocessing**: Converting categorical variables to numerical values, imputing missing data, and normalizing via Min-Max scaling.
2. **Feature Selection**: Techniques include Correlation, Chi-Square, Variance Threshold, Recursive Feature Elimination, Sequential Forward Selection, Lasso Regression, and Ridge Regression.
3. **Classifiers Used**: Random Forest (RF), AdaBoost (AdaB), Gradient Boosting (GB), XgBoost (XgB), Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT).

The study showed that RF and AdaB achieved 100% accuracy across various validation methods. Additionally, a real-time web-based application was developed to operationalize the model, enhancing accessibility for healthcare practitioners.

In conclusion, this study significantly advances predictive diagnostics for CKD by developing a sophisticated machine learning model and a user-friendly web application. Its methodologies and insights will serve as a valuable foundation for our model development, **guiding our selection of optimal models and feature selection methods for predicting CKD.**

# Methodology

## **Data Collection**

### **The database**

The data is based on electronic medical record data of patients admitted to the hospital. It includes demographic characteristics of the patient, blood and urine test data and history of disease data. The target variable is the existence/absence of EventCKD35.

### **Data Description:**

The data set contains 491 records and 25 columns. Features are described on  ***Fig 1, data information***.

## **Data Preprocessing** and **Data Visualization**-EDA

**Data Cleaning**: We addressed null values in the **TriglyceridesBaseline** and **HgbA1C** columns by imputing them with the mode to avoid bias from extreme values. No further inconsistencies or missing values were found, allowing us to continue with preprocessing.

**Categorical Variable Conversion**: Categorical variables were encoded using **One-Hot Encoding**. This method creates binary columns for each category, preventing any false ordinal relationships and ensuring no category is unfairly prioritized.

**Data Normalization**: We applied the **Robust Scaler** to normalize numerical variables. Unlike standardization, which can be skewed by outliers, the Robust Scaler uses the median and interquartile range, making it more robust for data with potential outliers, such as clinical test results.

### **Visualization techniques applied in the project are as followed:**

1. We began by presenting **summary statistics** for each column to understand the statistical values following preprocessing and establish our baseline.
2. **Histogram for Numerical features** was generated to all numerical columns to see their distributions.
3. **Correlation matrix** used to find possible interesting relations between features.
4. Lastly, we generated **Bar plots** for categorial features for comparison purposes.

### **Insights and observations:**

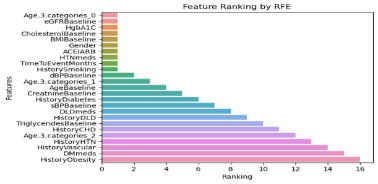
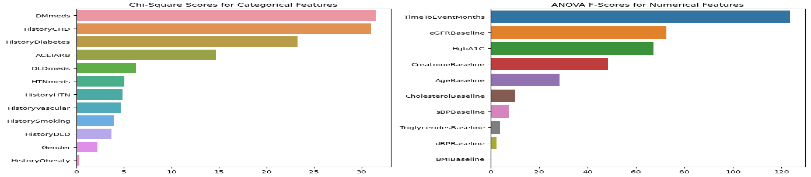
1. most of our numerical data is presenting a normal distribution. this will assist in model selection in future analysis.
2. correlations:
   1. We can see A negative correlation (r = -0.59) suggests that older individuals have reduced kidney filtration ability, which aligns with clinical expectations, and makes sense.
   2. A strong negative correlation indicates that lower eGFR values correspond to higher creatinine levels, reinforcing the relationship between poor kidney function and waste accumulation.
   3. Finally, an interesting yet not strong negative corr is found between the number of months to event and the existence of CKD. Seems like a longer time span decreases the likelihood of CKD35.
3. about the categorial features:
   1. most of the patients have a history of Dyslipidemia and are on DLD meds.
   2. most of the patients also suffer from HTN and take HTN meds
   3. Although most of the pts have a history of diabetes, almost only a half of them are on diabetic meds. this can imply on wrong community treatment that may cause future complication and is an interesting lead for us.

## **Feature Selection** and class imbalance

### **Techniques Used:**

1. **Chi-Square Test for Categorical Features:**
2. **Advantages:**Simple and fast, helps in identifying the strength of association between categorical features and the target variable.
3. **Disadvantages:** Assumes independence of features, which might not always be true.
4. **ANOVA F-test for Numerical Features:**
5. **Advantages:**Measures the linear relationship between features and the target variable, easy to interpret.
6. **Disadvantages:**Assumes linear relationship, may not work well with non-linear relationships.
7. **Wrapper Method- Recursive Feature Elimination (RFE):** Elimination (RFE) was used with a logistic regression model to select important features.
8. **Advantages:**Considers feature interactions, works well with various algorithms, selects features based on model performance.
9. **Disadvantages:**Computationally expensive, may be overfit on small datasets.

### **Results for feature selection:**



|  |  |
| --- | --- |
| **The Top features selected by statistical filtering approach** | **The Top features selected by RFE** |
| 'DMmeds', 'HistoryCHD', 'HistoryDiabetes', 'ACEIARB', 'DLDmeds', 'TimeToEventMonths', 'eGFRBaseline', 'HgbA1C', 'CreatnineBaseline', and 'AgeBaseline'. | 'HistorySmoking', 'HTNmeds', 'ACEIARB', 'Gender', 'Age.3.categories\_0', 'CholesterolBaseline', 'HgbA1C', 'eGFRBaseline', 'BMIBaseline', and 'TimeToEventMonths' |

### **Class imbalance**

The target variable, "Event CKD35," consists of two classes: 1 for diagnosed CKD patients and 0 for healthy patients. The classes are not equally represented, necessitating balance in their representation. We evaluated two methods to address this imbalance:

1. **SMOTE (Synthetic Minority Over-sampling Technique):**This method generates synthetic samples for the minority class to balance the class distribution.
   1. **Advantages:**
      1. Produces synthetic samples rather than duplicating existing ones.
      2. Balances the class distribution without information loss.
   2. **Disadvantages:**  
        May introduce noise if the synthetic samples are not generated properly.
2. **Random Under-sampling:**This technique reduces the number of samples in the majority class to achieve balance.
   1. **Advantages:**
      1. Simple and quick to implement.
      2. Reduces the size of the dataset, which can speed up training.
   2. **Disadvantages:**  
        May lead to loss of important information from the majority class.

Based on the results seen on ***figure 2 , Distribution of the target variable***, we will apply SMOTE to the training dataset to develop a well-trained model using balanced data.

* **On the following Phase, we will explore two approaches for model development:**

1. **Training models using only the top features selected during the feature selection phase.**
2. **Training models using all available features.**

**The method that achieves the best evaluation scores will be selected as the optimal approach for our model.**

## Model Development

**Classifiers Used:**

1. Random Forest (RF)
2. Support Vector Machine (SVM)
3. Logistic Regression (LR)
4. Decision Tree (DT)

In the research we introduced above, there was a usage of XGBoost method as well, but after a test of it on our data we found it retrieves bad results, therefore we did not continue with it.

**Model Training:**

We split the data to train and test set using "train test split" method, assigning 20% of the data to the test set and the rest to the train set (Random state was given to preserve the random split for future command runs).

## **Model Evaluation** and tuning

### **evaluation:**

1. **Metrics used:**

**We used the following metrics on every trained model and on both approaches:**

* 1. **Precision**: Precision measures the proportion of true positive predictions among all positive predictions, indicating the accuracy of the positive predictions made by the model.
  2. **Recall**: Recall, or sensitivity, measures the proportion of true positive predictions among all actual positive cases, indicating the model's ability to identify all relevant instances.
  3. **F1-score**: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns, particularly useful when the class distribution is imbalanced.
  4. **ROC-AUC**: The ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) measures the model's ability to discriminate between positive and negative classes, with a higher value indicating better overall performance.
  5. **Cross-Validation**: We used K-Fold Cross-Validation to improve the model stability.It was used during the Hyperparameter tuning phase as a default choice of our chosen method.
  6. **ROC and Precision-Recall curves:**

1. **Considerations:**
   1. If the cost of missing CKD is higher (false negatives), **recall** is crucial.
   2. If the cost of incorrectly predicting CKD is higher (false positives), **precision** is important.
   3. ROC-AUC and Precision-Recall AUC provide overall model performance.
2. **Results received:**

**Please see full results on Fig *3. models evaluation results:***

The best model's outcomes is for **LogisticRegression**:

Precision: 0.47Recall: 0.67F1-score: 0.55ROC-AUC: 0.78

|  |  |
| --- | --- |
| **Evaluation metrics for top features only** | **Evaluation metrics for all available features** |
| תמונה שמכילה טקסט, תרשים, קו, עלילה  התיאור נוצר באופן אוטומטי |  |

### **Hyperparameter Tuning:**

The method used is Grid search. For both approaches the results are the same:

1. log\_reg: {'C': 10, 'solver': 'liblinear'}
2. dec\_tree: {'max\_depth': 10, 'min\_samples\_split': 10}
3. rand\_forest: {'max\_depth': 15, 'n\_estimators': 100}
4. svm: {'C': 10, 'kernel': 'rbf'}

### **conclusions:**

Based on the evaluation metrics and plotted curves, we have selected the **Logistic Regression Classifier** as it achieved the best overall results across all metrics.

Since we chose Logistic Regression, we concluded that using all available features, rather than undergoing feature selection, is preferable. This conclusion is supported by the analysis of the ROC and Precision-Recall curves-

The ROC curve showed *only* a slight improvement (from 0.89 without feature selection to 0.9 with feature selection), whereas the Precision-Recall curve indicated a **more significant decrease in performance** (from 0.69 without feature selection to 0.66 with feature selection). Therefore, **the approach where all features are included enhances the model's precision and recall, leading to more reliable predictions.**

## EX**planatory AI Techniques**

**SHAP** is an approach to explain the output of any machine learning model. It connects game theory with local explanations. We showed few explanations plot from the SHAP models, to clarify our model's results.

1. **Explaining a single prediction from the test set:**

In this explanation, a "healthy" predicted patient was analyzed based on individual feature contributions. The key findings are that "HTNmeds"=0 and  DMmeds=0 are most likely to lead to a positive CKD prediction. The features eGFRBaseline= 0.9324 and  age.3.categories\_0=1 (youngest group)  drew the same prediction to be at 0.

This is understandable as **untreated hypertension or diabetes increases the risk of kidney damage, younger individuals are less likely to have CKD, and high eGFRBaseline values indicate healthy kidney function.**

1. **Summery plot:**

The plot described on ***Fig 4 SHAP Summery plot*** highlights the most important features and their effects on the model's predictions. Conclusions:

**TimeToEventMonths**: higher values reduce CKD likelihood, implying the longer time passes from last event, the less likely it is a pt. will have CKD.

**CreatinineBaseline**: High values increase CKD likelihood, which is reasonable as high creatinine values appears at kidney disfunction.

**Gender: women are more likely to impact on positive CKD prediction**

Medications: untreated HTN pts (HTNMEDS=0) are more likely to be diagnosed with CKD

**Other Features**: like baseline health metrics also influence predictions but not significantly.

**LIME** explains the predictions of any classifier by approximating it locally with an interpretable model**.**

Here in ***Fig 5 LIME plot***, It was found that **Gender** and **eGFRBaseline** have the most significant negative contributions, indicating they strongly influence the model towards predicting CKD, while **HTNmeds** and **dBPBaseline** have the most substantial positive contributions, suggesting they decrease the likelihood of CKD in this instance.

## **Subgroup Analysis**

### **Homogeneous Groups Identification:**

In our subgroup analysis, we used K-means clustering to segment the data into different numbers of groups (2-9) and assessed cluster quality using the Silhouette Index and Sum of Squared Errors (SSE).

1. **SSE (Sum of Squared Errors)**: Measures the compactness of the clusters; lower SSE indicates tighter clusters.
2. **Silhouette Score**: Evaluates how similar an object is to its own cluster compared to other clusters; higher scores indicate better-defined clusters.

The best silhouette and SSE scores were observed with 8 clusters. Furthermore, the most distinct partition in the silhouette plot was achieved with the 8-cluster solution.

1. **Elbow Method**: We plotted SSE against the number of clusters to identify the optimal number of clusters, where the SSE begins to plateau, forming an 'elbow'.

Surprisingly, the significant change in the SSE was found around the 5-6 cluster range, but a notable smaller decrease was also seen in the 7-8 cluster range.

**After evaluation, we determined that clustering into 8 groups provided the best balance of SSE (3278.80) and Silhouette Score (0.22), indicating well-defined clusters.**

### **Model Development for Subgroups:**

We focused on "group\_1," the subgroup with the most diagnosed CKD patients, replicating the EDA and preprocessing stages from the entire dataset, to see if we get better insights on the specific group.

* + 1. **EDA:**
  1. **The numeric data** is sparser and mostly non-normal. Baseline values are higher, and the age baseline has shifted upwards, suggesting that the group is containing more elderly patients.
  2. **correlations**:

The correlation between EGFR and Age has weakened from 0.59 to -0.49 but still indicates that older individuals have reduced kidney function.

The negative correlation between EGFR and CreatinineBaseline has strengthened, suggesting that lower EGFR corresponds to higher creatinine levels.

The correlation between the number of months to the event and CKD presence is now insignificant (0.06).

* 1. **Categorial features:**

The most significant change found in the current EDA is in the amount of people that consume ACE\ARB meds. It aligns with the high number of hypertensive patients in the group. ACE\ARB meds are consumed by patients with high BP and Kidney disease, as their purpose is to decrease the blood volume.

Next, there is a higher proportion of male patients. It gives us insight on the characteristics of our group.

Following, the number of diabetic patients has increased, with more patients taking diabetes medications than before.

To conclude the EDA part, this subgroup consists mainly of elderly CKD-diagnosed men with hypertension and diabetes, treated primarily with HTN and diabetes medications. Key correlations between age, EGFR, and creatinine remain consistent.

* + 1. **Class imbalance and feature selection:**

1. There wasn't a significant change in the feature selection, but some features aren't in the top features now.
2. The classes are almost balanced in the sub-group analysis; hence no class imbalance needed this time.
   * 1. **Models training and evaluation:**

After testing again all the models, Logistic Regression (LR) and Random Forest (RF) performed best, with identical scores (whereas before the clear leading model was LR):

Precision: 0.75

Recall: 0.75

F1-score: 0.75

ROC-AUC: 0.80

* + 1. **Hyperparameter search and models comparison:**
       1. The hyperparameters have changed slightly. Showing a different "C" for LR (1 VS 10), max depth for DT (5 vs 10) and for RF (5 VS 15). Also, different kernel suggested for SVM (Linear VS rbf).
       2. The comparison between ROC and Precision-Recall curve of all models showed no difference and LR is still the preferred model. In addition, **now it is preferred to use only the top selected features instead of all features**, as it gives us the best over-all evaluation results.
    2. **SHAP Explanatory AI:**

On ***Fig 6. Subroup SHAP summery plot,*** the clear key features influencing CKD predictions are: **HistoryCHD, eGFRBaseline, DMmeds, Age.3.categories\_2,** and **TimeToEventMonths**.

The clinical conclusions driven are that a history of CHD is the most significant predictor, though based on limited data. Low eGFR baseline indicates poor kidney function, as expected. Interestingly, DM medications increase CKD likelihood. Additionally, **TimeToEventMonths, age categories, and diastolic blood pressure** significantly impact model predictions, with varying influence compared to previous analyses. These findings underscore the value of machine learning in identifying subgroup-specific factors, aiding clinical decision-making.

# Results and conclusions

The dataset was clean and organized with minimal missing values. To address the imbalance in the target variable, SMOTE was applied to the training set, which helped avoid overfitting. Models trained on the entire dataset performed better without feature selection, while subgroup analysis benefited from selected features. **Logistic Regression was the best model for both the entire dataset and the subgroup analysis.**

The EDA for both the entire dataset and the subgroup analysis showed correlations of a negative relationship between age and eGFR , indicating reduced kidney filtration ability in older individuals, and a strong negative correlation between eGFR and creatinine levels, linking poor kidney function to high body waste. There was also a weak negative correlation suggesting a longer time from last event decreases CKD likelihood. The first EDA done showed **normal** distribution for numeric categories, while the subgroup EDA showed **sparser** data with higher values. Categorical features revealed for *entire dataset* that most patients had Dyslipidemia, Diabetes, and Hypertension, with nearly half not on diabetic medications, indicating potential community treatment issues. *On the contrary*, more patients in the *subgroup* **did consume** ACE/ARB medications, aligning with high hypertension prevalence, and there was a higher proportion of male and diabetic patients, with more on diabetes medications than before.

SHAP and LIME analyses identified TimeToEventMonths, CreatinineBaseline, Gender, and Medications as key influential features: Higher TimeToEventMonths reduced CKD likelihood, high CreatinineBaseline values increased CKD likelihood, women had a higher impact on positive CKD predictions, and untreated hypertension patients (HTNMEDS=0) were more likely to be diagnosed with CKD. Subgroup SHAP contributes less to the conclusions as it relies on limited data and can be biased.

The subgroup analysis is found to be characterized with **elderly CKD-diagnosed men with hypertension and diabetes, treated primarily with hypertension and diabetes medications**, providing insights for targeted treatment and intervention.

Finally, from all the above we can conclude that **addressing untreated diabetes or HTN in a significant portion of elderly patients could help prevent CKD complications.**

### The effectiveness of ML for CKD Prediction

Machine learning models significantly enhance CKD prediction, particularly within specific subgroups, by revealing patterns and insights that may be challenging for individual doctors to identify. **While they cannot entirely replace medical professionals,** these models process extensive patient data **quickly and accurately**, improving overall predictive capabilities.

Machine learning identifies the most important features in disease prediction, helping doctors prioritize treatments and determine which tests to perform. It also uncovers trends that may not be visible in smaller patient groups, providing a broader perspective on disease patterns.

In summary, machine learning is a valuable decision-making tool that complements traditional diagnostics, enabling doctors to make more informed decisions and improving CKD prediction and patient outcomes.

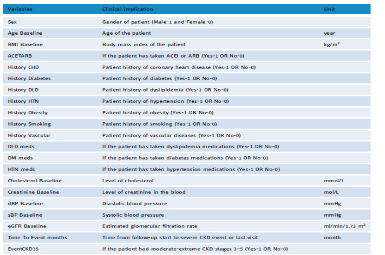
### research ideas including usage of different ML methods

We propose conducting a broader study using a larger dataset. This expanded study should include a detailed analysis of subgroups, comparing groups with CKD patients to those without to identify unique characteristics of each group.

We also recommend employing Association Rules algorithms to uncover non-obvious patient patterns and exploring neural network models to assess their accuracy compared to simpler, more interpretable models like tree-based models. While our literature review indicates that tree models can yield excellent results, our current dataset showed that the linear regression model outperformed them.

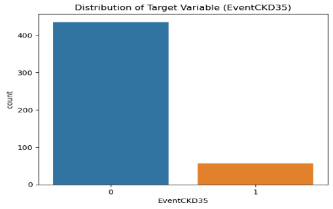
# appendix

1. Fig 1.1, data information



2.1 Distribution of the target variable

**Distribution of the target variable before applying techniques:**



|  |  |
| --- | --- |
| **Distribution of the target classes after SMOTE:** | **Distribution of the target classes after Random Under Sampling:** |
| תמונה שמכילה טקסט, צילום מסך, תרשים, מלבן  התיאור נוצר באופן אוטומטי | תמונה שמכילה טקסט, צילום מסך, מלבן, תרשים  התיאור נוצר באופן אוטומטי |
| the distribution of the target variable is:  0: 348, 1: 348 | the distribution of the target variable is:  0: 44, 1: 44 |

3. models evaluation results:

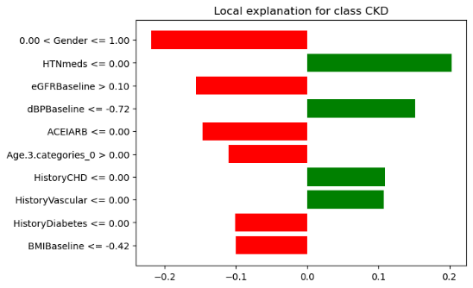
|  |  |
| --- | --- |
| **Evaluation metrics for top features only** | **Evaluation metrics for all available features** |
| * **LogisticRegression**:   Precision: 0.38  Recall: 0.75  F1-score: 0.50  ROC-AUC: 0.79   * **DecisionTreeClassifier**:   Precision: 0.44  Recall: 0.67  F1-score: 0.53  ROC-AUC: 0.78   * **RandomForestClassifier**:   Precision: 0.50  Recall: 0.33  F1-score: 0.40  ROC-AUC: 0.64   * **SVC**:   Precision: 0.37  Recall: 0.58  F1-score: 0.45  ROC-AUC: 0.72 | * **LogisticRegression**:   Precision: 0.47  Recall: 0.67  F1-score: 0.55  ROC-AUC: 0.78   * **DecisionTreeClassifier**:   Precision: 0.43  Recall: 0.50  F1-score: 0.46  ROC-AUC: 0.70   * **RandomForestClassifier**:   Precision: 0.56  Recall: 0.42  F1-score: 0.48  ROC-AUC: 0.69   * **SVC**:   Precision: 0.41  Recall: 0.58  F1-score: 0.48  ROC-AUC: 0.73   * **XGBClassifier**- done only here:   Precision: 0.45  Recall: 0.42  F1-score: 0.43  ROC-AUC: 0.67  *Due to bad results this model was neglected.* |

4 SHAP Summery plot

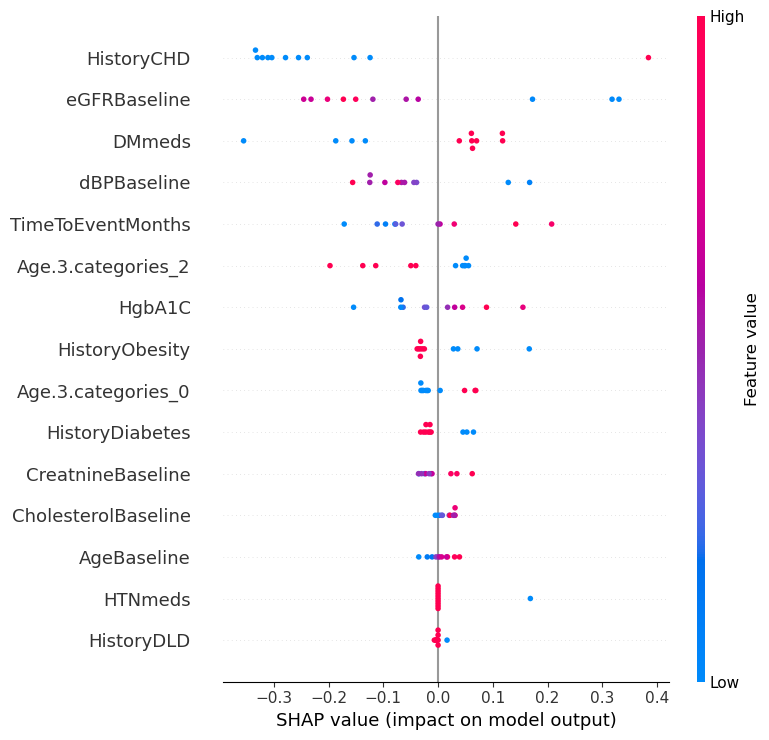
תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי

5 LIME plot



6. Subroup SHAP summery plot



# Bibliography

R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, S. Saha, R. Hossen, S. Ahmed, M. A. T. Rony, and M. F. Akter, **"ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application,"** *Journal of Pathology Informatics*, vol. 15, no. 100371, 2024.