**Technical Report**

**Applications of Machine Learning in Medicine**

Faculty of Engineering, Ariel University, Department of Industrial Engineering and Management

**Abstract**: This report investigates the effectiveness of supervised machine learning techniques for predicting Chronic Kidney Disease (CKD).

**Written by**:

* + - * Amit Fallah: 318510070
      * Karin Lev-Sofi: 206682577
      * Ido Zimels: 206070682
      * Matan Shimon: 314716648

**Lecturer:** Dr.Orit Refaeli

**Date**: 28.07.2024

**Table of Contents**

1. **Abstract...........................................................................................3**
2. **Introduction and Purpose of the Work............................................ 3**
3. **Methods4...........................................................................................**
   * Data Collection4....................................................................
   * Preprocessing......................................................................4
   * Handling Outliers…………………………………………………………4
   * Exploratory Data Analysis (EDA)...........................................5
   * Feature Selection.................................................................5
   * Handling Class Imbalance...................................................7
   * Model Training and Evaluation.............................................8
4. **Explanatory AI: SHAP Model Analysis…………………………………….9**
5. **Subgroup Analysis Using K-means Clustering.............................10**
6. **Finding ..........................................................................................11**
7. **Conclusion………………………………………………………………….……12**
8. **Bibliography13..................................................................................**

**Abstract**

This report explores the development of a predictive model for chronic kidney disease (CKD) using advanced data analysis and machine learning techniques.

Initial exploratory data analysis (EDA) identified key relationships between critical variables such as creatinine levels and glomerular filtration rate (eGFR), diastolic and systolic blood pressure, and eGFR and age.

The study found significant correlations consistent with clinical literature, emphasizing the importance of these variables in CKD progression.

Following these insights, we implemented dimensionality reduction using Principal Component Analysis (PCA) and applied unsupervised learning algorithms like TSNE, KMeans, and hierarchical clustering to divide the data into homogeneous groups.

Initially, a two-cluster approach revealed that the majority of positive CKD cases were concentrated in one cluster, prompting a deeper investigation into more granular clustering. By increasing the clusters to five, we achieved improved model performance, indicating that detailed clustering enhances predictive accuracy.

The final logistic regression model was selected based on its optimal recall and balanced F1 score, crucial for minimizing the risks associated with false negatives in medical predictions. This comprehensive analysis underscores the importance of multi-faceted data exploration and tailored machine learning approaches in improving CKD prediction and understanding.

**Introduction and Purpose of the Work**

Chronic Kidney Disease (CKD) is a significant public health issue affecting millions of people worldwide, leading to severe complications such as end-stage renal disease, cardiovascular diseases, and premature mortality.

Early diagnosis of CKD is crucial as it allows for early intervention, which can slow the progression of the disease and improve patients' quality of life.

In the era of electronic health records (EHR), there is a vast amount of routinely collected medical data available that can be harnessed to develop advanced predictive models for diseases.

Machine learning offers powerful tools for analyzing these data and identifying patterns that may be challenging to detect using traditional methods.

This project aims to develop a machine learning model to predict the risk of developing CKD based on electronic medical record data.

We will focus on the process of data preparation, selection of relevant features, addressing class imbalance, model training and evaluation, and interpretation of prediction results.

By leveraging this research, we aim to demonstrate the potential of machine learning as a tool for early prediction of CKD and provide insights that could improve clinical practice and reduce the global health burden of CKD.

This study not only explores the predictive power of various supervised machine learning models but also seeks to identify the clinical features most strongly associated with CKD development.

**Methods**

**Data Collection**

The dataset used in this project includes laboratory test results, clinical symptoms and demographic data.

Data was collected from Electronic records collected in hospitals that include data about CKD.  
The dataset consists of 491 instances and 24 features.

The data collection phase involved gathering relevant data.

The collected data included features such as age, BMI, medical history, and various lab results, which were crucial for building predictive models.

**Preprocessing**

Data preprocessing involved several steps:

* **Handling Missing Values**: We handled missing values ​​for the TriglyceridesBaseline, HgbA1C columns by looking at the distribution of the data.

We chose to remove the outlier data from the 'TriglyceridesBaseline' column, thus obtaining a normal distribution that is characterized symmetrically around the mean, so that we fill the missing values ​​in the column with the mean, because the mean represents the center of the data distribution and fills missing values ​​in a value that is not extreme.

This is a good way to keep the data stable and prevent missing values ​​from affecting the results.

For the 'HgbA1C' column, after we divided the values ​​according to the patient's diabetes history, we saw that the standard deviation decreased significantly for each cluster separately, so we fill in the average according to the division (first cluster - positive past for diabetes, second cluster - negative past).

This step was crucial for maintaining the dataset’s integrity and ensuring that no valuable information was lost.

* **standardization of numerical variables:** We first explored the data by density matrix. We can see a Gaussian distribution in the columns, HgbA1C, CreatnineBaseline, eGFRBaseline, sBPBaseline, dBPBaseline, BMIBaseline. (Figure 1).

**Handling Outliers**

Outliers checking revealed that the outlier values in the 'TriglyceridesBaseline' column belong to the 'HistoryDLD' class equal to 1 (Figure 2,3).

Similarly, outlier values in the 'HgbA1C' column are associated with the 'HistoryDLD' = 1 and 'HistoryDiabetes' = 1 classes.

There is a statistically significant dependence between the categorical variables 'HistoryDLD' and 'HistoryDiabetes' according to the chi-square test. (Figure 4,5).

Additionally, outlier values in the 'BMIBaseline' column belong to the 'HistoryCHD' = 0 class (Figure 6).

We used pie charts to see the distribution of binary columns and boxplots, histograms, and violin graphs to examine numerical columns. Additionally, a chi-squared test was employed to test statistical dependence between categorical variables.

These visualizations and statistical tests provided deeper insights into the data, highlighting significant patterns and relationships crucial for understanding and predicting CKD.

**Exploratory Data Analysis (EDA)**

In the EDA part, we performed a preliminary study using pie charts to see the distribution of the binary variables and histograms to see the distribution of the numerical variables.

We examined relationships between categorical variables using statistical tests and relationships between numerical variables using heat maps and regplot graphs.

The conclusions we reached are

**Correlation Between eGFR and Age:**

A moderate negative correlation was found between eGFR and age.

As age increases, eGFR decreases.

This finding is consistent with research reporting a decline in kidney function with increasing age. *(article 3).* (Figure 7).

**Correlation Between Creatinine Levels and Glomerular Filtration Rate (eGFR):**

A moderate negative correlation was found between creatinine levels and eGFR.

As creatinine levels increase, eGFR decreases.

This finding supports clinical literature reporting that a decrease in eGFR is a significant marker for assessing CKD. *(article 1).*

Specifically, the negative correlation is notably stronger in patients with

'HistoryCHD' = 1 (correlation of -0.81). (Figure 8).

**Correlation Between Diastolic Blood Pressure (dBP) and Systolic Blood Pressure (sBP):**

A moderate positive correlation was observed between dBP and sBP.

As diastolic blood pressure increases, so does systolic blood pressure. *(article 2).*

This finding aligns with clinical literature emphasizing the importance of both measures in predicting kidney disease progression. (Figure 9).

**Feature Selection**

Before we started the feature selection, we divided the data into a training set and a test set.

The split is designed to prevent overmatching and information leakage between sets that should be separate in advance.

The motivation is that during the process we will not learn anything from the test set itself.

A two-step approach combining statistical tests and wrapper methods was employed for feature selection:

1. **Statistical Filtering:**

* **Chi-Square Test**: Applied to categorical features to measure association strength with the target variable. (figure 10).
* Strengths:

Simplicity: Easy to implement and interpret.

Efficiency: Quickly identifies the strength of association between categorical features and the target variable.

* Limitations:

Independence Assumption: Assumes that features are independent, which may not always hold true in practice

* **ANOVA F-test**: Used for numerical features to assess linear relationships with the target variable. (Figure 11).
* Strengths:

Interpretability: Measures how well numerical features correlate with the target variable in a straightforward manner.

Linear Relationships: Effective for identifying linear relationships between features and the target.

* Limitations:

Linearity Assumption: Assumes a linear relationship, which might not capture complex or non-linear associations.

1. **Wrapper Method:**

* **Recursive Feature Elimination (RFE)**: Utilized to select features based on model performance while considering feature interactions. (Figure 12).
* Strengths:

Interaction Consideration: Takes into account interactions between features, which can be beneficial for model performance.

Versatility: Can be used with various algorithms to select features based on model performance.

* Limitations:

Computational Intensity: Can be computationally expensive, especially with a large number of features.

Risk of Overfitting: May overfit if the dataset is small, as it relies heavily on model performance metrics.

The final selected features for the model included:

* DMmeds
* eGFRBaseline
* HistorySmoking
* CholesterolBaseline
* CreatinineBaseline
* BMIBaseline
* ACEIARB
* HTNmeds
* TimeToEventMonths
* HistoryCHD
* Age.3. categories
* HistoryDiabetes
* HgbA1C', 'DLDmeds

**Handling Class Imbalance**

Class imbalance is a common problem in machine learning where some classes are underrepresented compared to others, leading to biased models that perform poorly on the minority class.

To address this issue in our dataset, we explored various sampling techniques and evaluated their effectiveness.

Applying class balancing techniques to the X\_train\_selected dataset is crucial for addressing class imbalance and improving model training.

By balancing the training data (X\_train\_selected and y\_train), we ensure that the model learns from a dataset with equal representation of each class, which helps reduce bias and enhances the model's ability to generalize to the minority class.

It is important to keep the test data (X\_test and y\_test) unchanged to maintain an unbiased evaluation of the model's performance on real-world data. (Figure 13).

**SMOTE (Synthetic Minority Over-sampling Technique)**

**Explanation:** SMOTE generates synthetic samples for the minority class by interpolating between existing minority class samples. (Figure 14).

* Advantages: Increases diversity by creating new, synthetic examples.

Can improve classifier performance by balancing the dataset.

* Disadvantages: Risk of overfitting due to similarity of synthetic samples.

Computationally intensive.

May introduce data complexity that isn’t representative.

Suitable For: Imbalanced datasets with continuous features where interpolation is meaningful.

**Random Over-Sampling**

**Explanation:** Random Over-Sampling duplicates existing examples from the minority class to balance class distribution. (Figure 15).

* Advantages:

Simple and easy to implement.

Effective in improving performance for moderate class imbalance.

* Disadvantages:

Risk of overfitting due to duplication of examples.

Does not increase data diversity.

Suitable For: Moderate imbalance in small to medium datasets.

**Why Use SMOTE?**

SMOTE (Synthetic Minority Over-sampling Technique) is recommended for balancing datasets when you aim to enhance model performance by generating synthetic samples.

Unlike simple over-sampling, which duplicates existing samples, SMOTE creates new, synthetic instances by interpolating between existing minority class samples.

This approach helps in capturing more nuanced patterns and relationships within the data, especially useful for complex datasets where the creation of additional synthetic data points can significantly improve the learning process of the model.

Therefore, using SMOTE (X\_train\_smote and y\_train\_smote) can lead to better model performance and a more robust classifier.

**Model Training and Evaluation**

Initial model training involved the following steps:

* **Model Selection**: We experimented with various machine learning algorithms including Logistic Regression, Decision Trees, XGBoost, and Support Vector Machines (SVM).

These models were chosen based on their suitability for classification tasks and their performance in previous studies.

* **Model Training**: Model training: Models were trained on the balanced data set created using SMOTE on the variables selected in the feature selection stage.

These actions ensured that the models were not biased towards the majority class and could correctly predict the minority class.

* **Model Evaluation**: The models were evaluated using metrics such as accuracy, precision, recall, and F1 score. (Figure 16).

These metrics provided a comprehensive understanding of the models' performance, highlighting their strengths and weaknesses.

* **Hyperparameter Tuning**: We performed hyperparameter tuning to optimize model performance.

This involved adjusting various parameters such as regularization strength in Logistic Regression, depth of trees in Decision Trees and Random Forests, and kernel parameters in SVM.

**Grid Search for Hyperparameter Tuning**

Grid search is a technique for hyperparameter tuning that involves an exhaustive search through a manually specified subset of the hyperparameter space.

It systematically works through multiple combinations of parameter values, cross-validating as it goes to determine which set of parameters yields the best performance.

The advantages of grid search include its simplicity and thoroughness, as it explores all possible combinations within the provided grid. However, this method can be computationally expensive and time-consuming, especially with large datasets or numerous hyperparameters.

Despite these drawbacks, grid search is a robust method to find optimal hyperparameters and enhance model performance when computational resources allow.

**Performance Measures**

**Precision:** Vital in cases where false positives come at an expensive cost.

**Recall:** Important when the expense of false negatives is significant.

**F1-Score:** The harmonic mean of precision and recall, useful when dealing with unbalanced classes.

**ROC-AUC:** Indicates the model's ability to distinguish between classes.

**Precision-Recall AUC:** More informative when dealing with unbalanced datasets.

Considerations for Choosing a Model

In medical prediction, the cost of treatment for a wrong diagnosis and the implications of missing a positive case classified as negative are significant.

Late detection or failure to detect kidney disease can lead to irreversible damage to the patient. Therefore, based on the above considerations and the plotted curves, we prioritize the recall metric and choose logistic regression.

This model not only demonstrates optimal performance in terms of recall but also achieves a favorable optimal F1 score, balancing the consideration of false positive alerts.

We looked at (*article 4)* which deals with chronic kidney diseases, and according to this article, we chose the evaluation indices.

**Explanatory AI: SHAP Model Analysis**

SHAP (SHapley Supplemental Explanations) offers a comprehensive framework for interpreting machine learning models by assigning each feature an importance value that reflects its contribution to a particular prediction (in our case using logistic regression).

This methodology provides a clear understanding of how individual attributes affect model predictions, enabling more transparent and actionable insights. (Figure 17).

In our analysis, the SHAP summary plot for the top five variables highlights several key findings. `eGFRBaseline` shows a significant negative tendency in the plot.

The predominance of red samples on the negative side indicates that lower eGFR values are associated with a higher risk prediction for chronic kidney disease. *(article 3).*

This aligns with clinical knowledge where decreased eGFR typically reflects impaired kidney function and an elevated risk of CKD.

The variability in the plot suggests that while eGFR is a strong predictor, its influence on risk predictions can vary across different cases.

`TimeToEventMonths` is predominantly positioned on the negative side of the plot, with red-tinted samples. This negative association implies that a longer time to the critical health event corresponds to a decreased prediction of the positive outcome.

Clinically, this suggests that a prolonged period until the event might be linked to a reduced likelihood of immediate disease progression, which could influence the urgency of medical interventions and monitoring strategies.

`ACEIARB` (Angiotensin-Converting Enzyme Inhibitors and Angiotensin Receptor Blockers) displays a uniform distribution across both sides of the SHAP plot, indicating no strong or consistent influence on the model’s predictions.

This suggests that ACEIARB may not be a significant predictor in this context, or its impact is inconsistent among different samples.

This calls for further investigation into the variable's role and its potential variability in influencing the outcome.

`HgbA1C` reveals a mixed distribution with both positive and negative SHAP values.

Higher levels of HgbA1C, which indicate poorer glucose control, are associated with increased CKD risk.

However, the variability in the plot points to a threshold effect, where the influence of HgbA1C on predictions becomes more pronounced beyond a certain level.

This variability underscores the importance of careful management and assessment of HgbA1C levels in evaluating CKD risk. *(article 5).*

Lastly, `CholesterolBaseline` shows a balanced distribution across both positive and negative sides, with mixed colors indicating varying effects.

This suggests that cholesterol levels impact model predictions in a complex manner, with both high and low values being associated with different risks.

Clinically, this implies that cholesterol should be managed alongside other risk factors for a comprehensive assessment of CKD risk. *(article 6).*

In summary, SHAP has provided valuable insights into how different variables influence the prediction of CKD risk.

While eGFR and HgbA1C show significant and variable impacts, TimeToEventMonths and CholesterolBaseline demonstrate more complex relationships. The uniform distribution of ACEIARB suggests limited predictive value in this model. These findings emphasize the need for a nuanced approach in incorporating these variables into CKD risk assessments.

**Subgroup Analysis Using K-means Clustering**

**Clustering Analysis and Insights**

Our clustering analysis aimed to refine the understanding of the dataset and improve predictive accuracy for chronic kidney disease (CKD). The process involved several key methodologies, including the elbow method, Principal Component Analysis (PCA), silhouette analysis, and hierarchical clustering, culminating in a detailed examination of cluster divisions.

**Methodology and Process**

We began our analysis by employing the elbow method to determine the optimal number of clusters. This involved plotting the inertia (within-cluster sum of squares) against the number of clusters and identifying the "elbow" point, (Figure 18) where the rate of decrease in inertia slows significantly. (Figure 19).

This graphical approach helps in selecting an appropriate number of clusters by balancing between model complexity and performance.

Following this, we applied Principal Component Analysis (PCA) for dimensionality reduction. PCA helps in simplifying the dataset by reducing its number of features while retaining most of the variance in the data. (Figure 20).

This step is crucial for visualizing and interpreting clusters more effectively, as it transforms the data into a lower-dimensional space.

With the reduced dimensions, we then used silhouette analysis to evaluate the quality of clusters. Silhouette scores measure how similar each sample is to its own cluster compared to other clusters.

Higher silhouette scores indicate well-defined clusters.

We visualized the silhouette scores through graphs, which helped in understanding how different numbers of clusters affect the model.

Next, we employed hierarchical clustering using the WARD method. The WARD method minimizes the total within-cluster variance by combining clusters in a way that results in the smallest increase in variance. (Figure 21).

This technique ensures that the clusters formed are cohesive and distinct.

We compared the results from hierarchical clustering with silhouette scores and t-SNE (t-Distributed Stochastic Neighbor Embedding) visualizations (Figure 22) to assess the quality and distinctiveness of the clusters.

**Insights from Clustering Analysis**

The initial clustering analysis with K=2, based on the highest silhouette scores and t-SNE visualizations, provided a silhouette score of 0.1597. This division revealed a significant concentration of positive samples in Cluster 0, particularly in the 'HistoryCHD', 'HistoryVascular', and 'EventCKD35' columns. However, focusing the CKD prediction model on this cluster alone resulted in a lower F1 score of 0.5556, (Figure 23) indicating the need for a more refined approach.

To address this, we further explored clustering the data into K=5 clusters. This decision was driven by the realization that the initial K=2 clusters did not capture enough granularity. The division into 5 clusters was motivated by the need to investigate smaller, more heterogeneous groups and improve the prediction model’s performance. In this finer clustering, we observed that Cluster 0 continued to contain a significant portion of positive samples.

Implementing the predictive model on this cluster yielded an improved F1 score of 0.6667,

(Figure 24) reflecting the benefits of a more detailed clustering approach.

**Conclusions and Future Directions**

The improved results from the K=5 clustering underscore the value of detailed and specific clustering.

The transition from K=2 to K=5 clusters enhanced the predictive accuracy and provided more nuanced insights into the data.

Smaller, more focused clusters allowed for better model performance and a clearer understanding of CKD risk factors.

Future research should involve adding more samples and further refining the clustering approach to enhance prediction efficiency.

Expanding the dataset could create larger clusters, potentially improving the predictive model's performance and providing deeper insights into CKD risk factors.

This iterative approach emphasizes the importance of detailed clustering in improving predictive modeling and guiding subsequent research efforts.

**Findings**

**1. Relationship between Creatinine Levels and eGFR:**

* A moderate negative correlation of -0.81 was found between 'CreatnineBaseline' and 'eGFRBaseline'.
* This finding aligns with clinical literature indicating that increased creatinine levels correlate with decreased eGFR, an important indicator of CKD progression. *(article 1).*

**2. Relationship between Diastolic and Systolic Blood Pressure:**

* A moderate positive correlation was observed between 'dBP' and 'sBP' (0.64).
* This result supports existing research highlighting the significance of blood pressure measurements in predicting kidney disease progression**.** *(article 2).*

**3. Relationship between eGFR and Age:**

* A moderate negative correlation was found between 'eGFR' and age (-0.66).
* This is consistent with numerous studies reporting that kidney function declines with increasing age. *(article 3).*

**4. Model Evaluation:**

Logistic Regression demonstrated high performance with a precision of 0.714, recall of 0.833, F1-score of 0.769, and ROC-AUC of 0.942.

**5. Dimensionality Reduction:**

- PCA effectively reduced dimensions, facilitating clearer clustering analysis (10 columns with 0.9 Variation explained).

**6. Initial cluster (K=2):**

* Grouping after reducing dimensions into 2 groups showed an optimal silhouette score of 0.1597
* Prominent positive sample concentrations in 'HistoryCHD', 'HistoryVascular' and 'EventCKD35' were observed in cluster 0 which resulted in a homogeneous division (cluster 0 contains 0.98 of the positive samples in the target variable).

**7. Cluster-Specific Prediction:**

Focusing on Cluster 0, which contained most positive samples, resulted in an F1 score of 0.5556, suggesting the need for more detailed clustering.

**8. Further Clustering (K=5):**

* Transitioning to 5 clusters improved the F1 score to 0.6667.
* This approach maintained relationships between variables like 'HistoryCHD' and 'EventCKD35' and provided better accuracy compared to the K=2 division.

**Conclusions**

**1. Outlier Analysis:**

* Outlier values in the 'TriglyceridesBaseline' column are associated with 'HistoryDLD' = 1.
* Outlier values in the 'HgbA1C' column are associated with 'HistoryDLD' = 1 and 'HistoryDiabetes' = 1.
* A statistically significant dependence was found between the categorical variables 'HistoryDLD' and 'HistoryDiabetes' (Chi-square test).
* Outlier values in the 'BMIBaseline' column are associated with 'HistoryCHD' = 0.

**2. SHAP Summary Plot Insights:**

* eGFRBaseline: Lower values are associated with higher risk predictions.
* TimeToEventMonths: Longer times to event are linked with reduced probability of the event.
* ACEIARB: No significant influence observed.
* HgbA1C: Mixed influence, suggesting variability in risk prediction.
* CholesterolBaseline: Balanced effect, indicating complex interactions with other variables.

**3. Implications of Detailed Clustering:**

Smaller, more specific clusters enhanced predictive accuracy and provided deeper insights into CKD risk factors.

**4. Future Directions:**

Expanding the dataset and refining clustering methods could further improve prediction efficiency and insights into CKD risk factors.

**Summary**

The analysis began with a comprehensive preliminary evaluation of the dataset, achieving an F1-score of 0.769 without clustering.

To explore data patterns and improve model performance, dimensionality reduction techniques were applied in the unsupervised learning phase.  
Initial clustering into 2 groups revealed that one cluster contained the majority of positive outcomes, indicating the formation of relatively homogeneous groups.

However, the lack of diversity within this cluster led to less effective predictions.

Consequently, the dataset was re-clustered into 5 groups, which enhanced the prediction performance with fewer samples.

Throughout the process, SMOTE was used to balance the dataset, and variable selection techniques were applied to refine feature relevance.

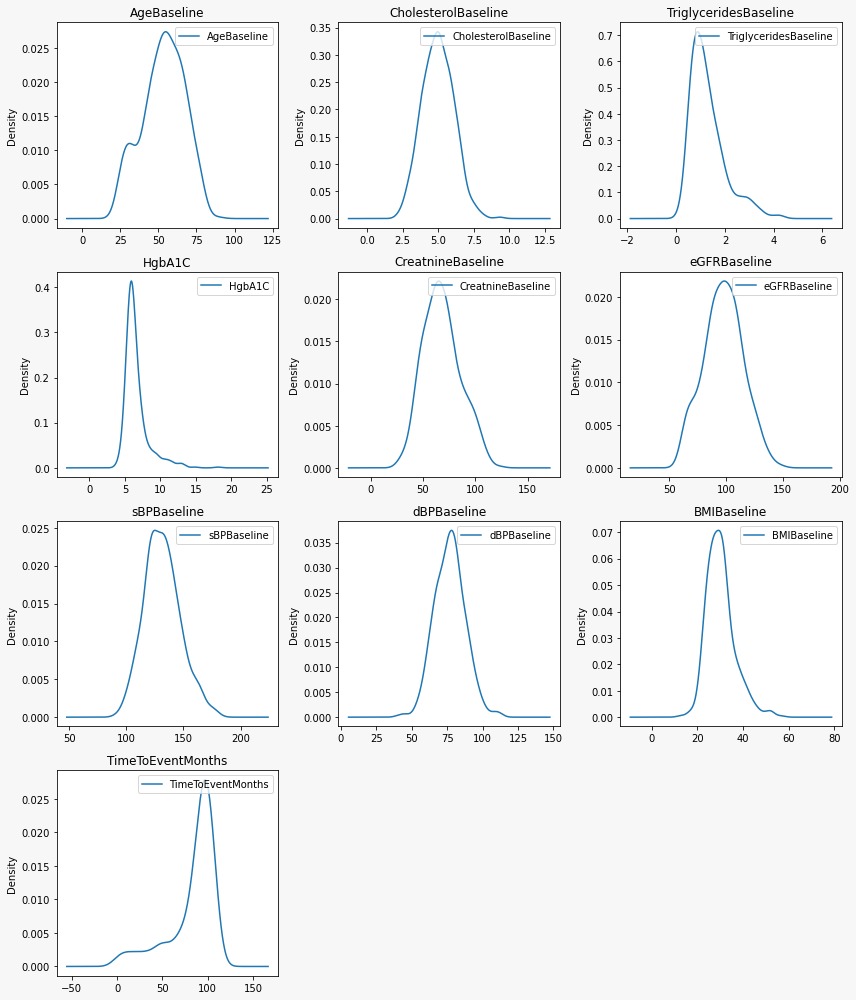
Future work should focus on expanding the dataset and further refining clustering methods to optimize prediction accuracy and gain deeper insights into the data.

**Bibliography**

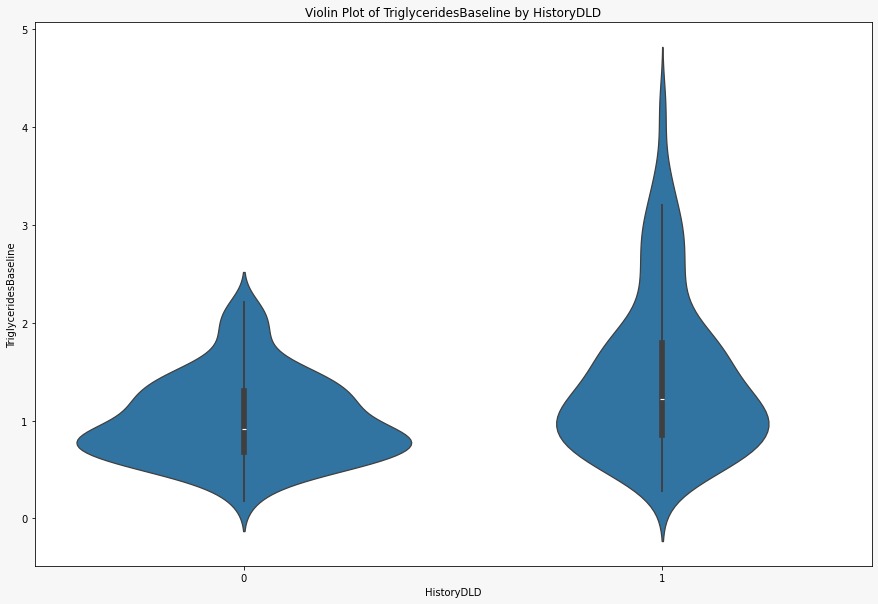
1. Utility and Validity of Estimated GFR–Based Surrogate Time-to-Event End Points in CKD: A Simulation Study: <https://www.kidney.org/sites/default/files/CKDEndpoints_Reference09.pdf>
2. Pulse pressure modifies the association between diastolic blood pressure and decrease in kidney function: the Japan Specific Health Checkups Study:

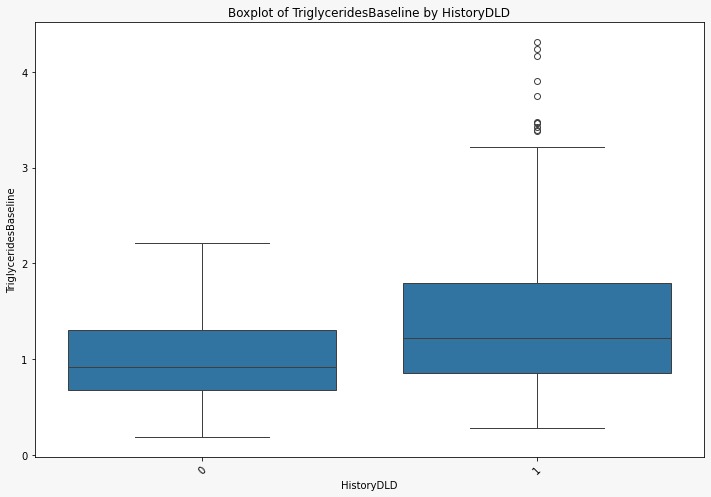
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11153873/>

1. Longitudinal Blood Pressure in Late-Stage Chronic Kidney Disease and the Risk of End-Stage Kidney Disease or Mortality (Best Blood Pressure in Chronic Kidney Disease Study): <https://www.ahajournals.org/doi/full/10.1161/HYPERTENSIONAHA.117.09855>
2. Machine Learning Techniques for Chronic Kidney Disease Risk Prediction: <https://www.mdpi.com/2504-2289/6/3/98>
3. How does CKD affect HbA1c? : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1753-0407.12624>
4. Cholesterol Disturbances and the Role of Proper Nutrition in CKD Patients: <https://www.mdpi.com/2072-6643/11/11/2820>

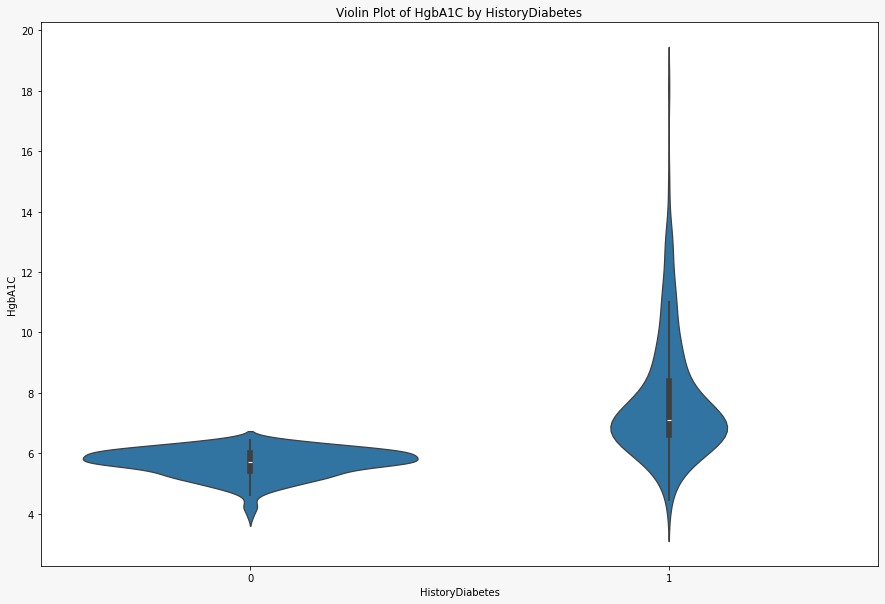
 **Figure 1 - Numerical columns density plot**

**Figure 2 - violin of TriglyceridesBaseline by HistoryDLD**

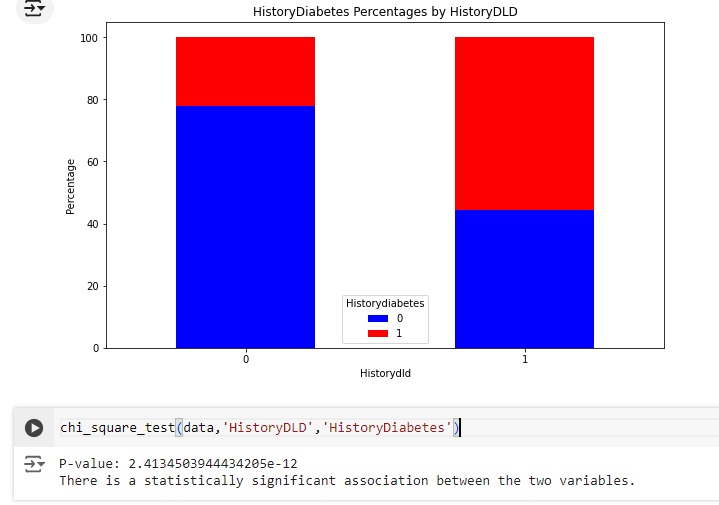


 **Figure 3 - boxplot of TriglyceridesBaseline by HistoryDLD**

**Figure 4 - violin of HgbA1C by HistoryDiabetes**



**Figure 5 - HistoryDiabetes by HistoryDLD**



**Figure 6 - violin of BMIBaseline by HistoryCHD**

A blue and black shapes

Description automatically generated with medium confidence

**Figure 7 – eGFR vs age**

A blue dotted line with black border

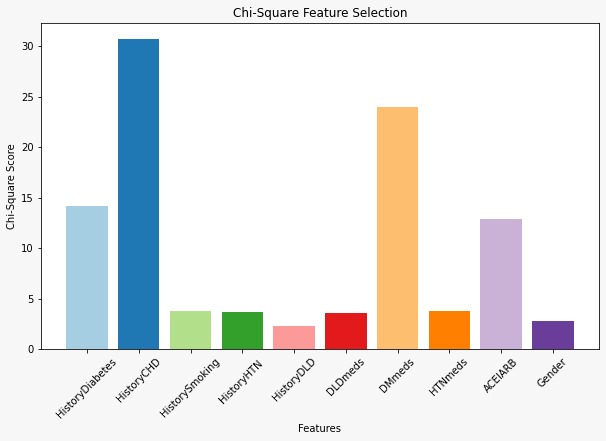
Description automatically generated

תמונה שמכילה צילום מסך

התיאור נוצר באופן אוטומטי**Figure 8 – eGFR vs Creatine**

תמונה שמכילה צילום מסך, פתית שלג

התיאור נוצר באופן אוטומטי**Figure 9 – sBP vs dBP**

 **Figure 10 - Chi-Square Feature Selection**

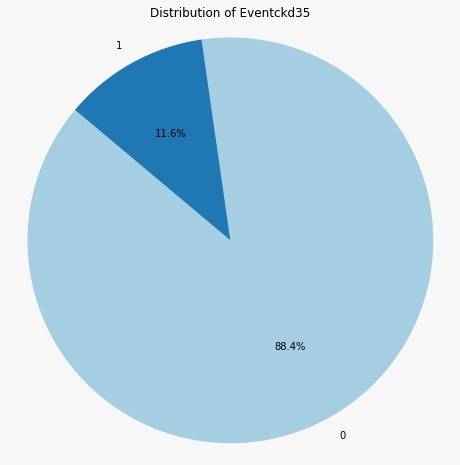
A graph of different colored bars

Description automatically generated **Figure 11 - ANOVA F-test Feature Selection**

A graph showing the ranking of a ranking

Description automatically generated with medium confidence **Figure 12 - Feature Ranking by RFE**

**Figure 13 - Class Imbalance**

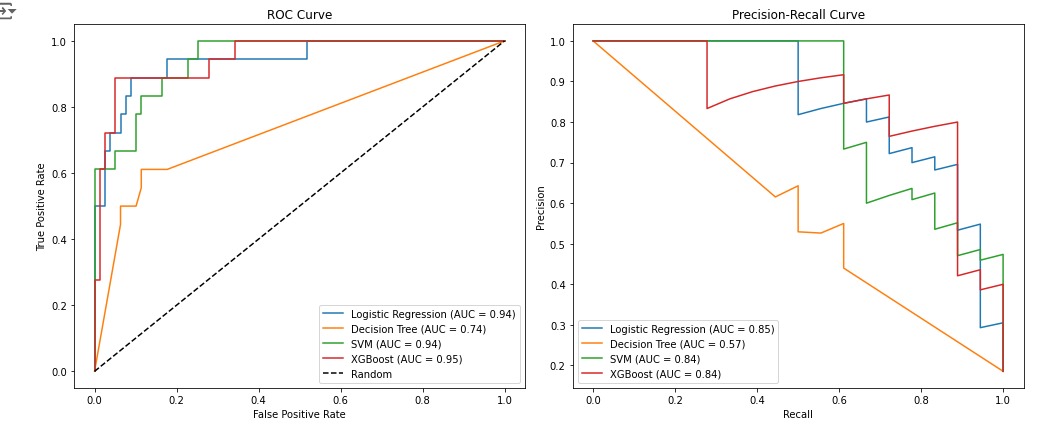


A graph of a pie chart

Description automatically generated with medium confidence **Figure 14 - Class Imbalance SMOTE**

A graph of a pie chart

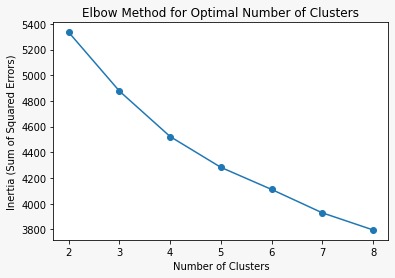
Description automatically generated with medium confidence **Figure 15 - Class Imbalance Random Over-Sampling**

 **Figure 16 – ROC and Precision-Recall AUC**

A graph with blue and white text

Description automatically generated**Figure 17 - Bar Plot: Ranks features based on their importance.**

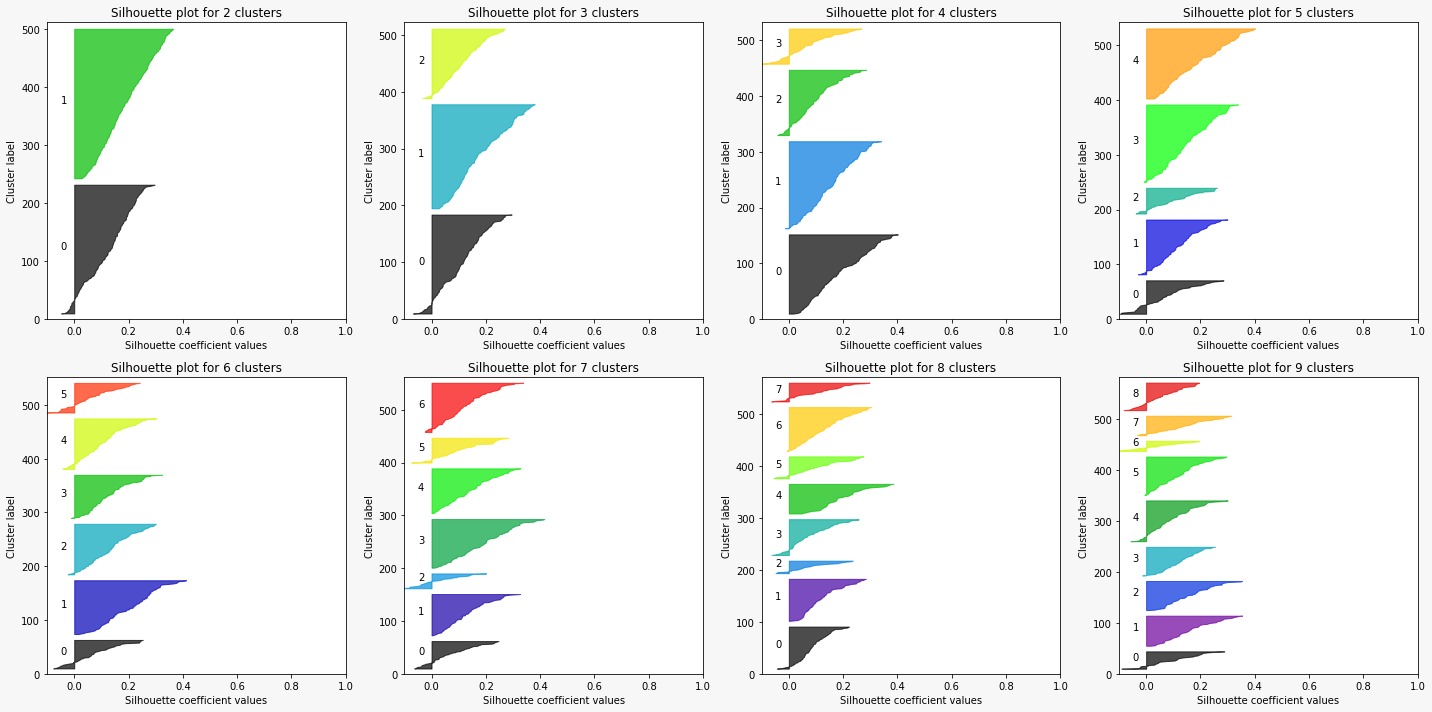
**Figure 18 - Elbow Method**

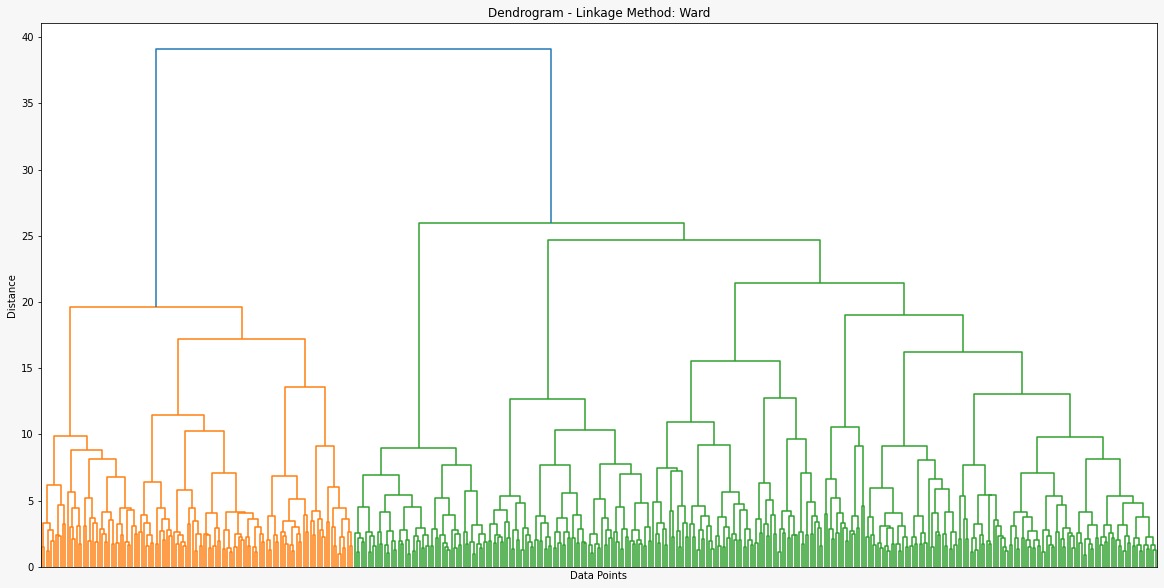


**Figure 19 - Inertia Method**

A graph of a number of inertia

Description automatically generated

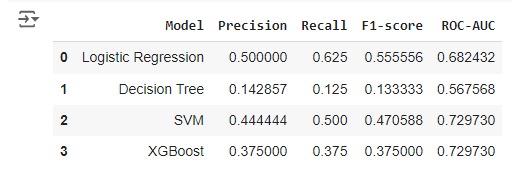
 **Figure 20 - silhouettes after applying PCA**

**Figure 21 – WARD Method**

**Figure 22 - tsne visualization**



**Figure 23 - evaluate models with cv K=2**



A screenshot of a graph

Description automatically generated **Figure 24 - evaluate models with cv K=5**

**Figure 24 – Model Evaluation**

