Determining Fatal Heart Failure Risks in Patients Diagnosed with Chronic Kidney Disease: A Machine Learning Approach

by

Adiba Haque
21241097
Anika Nahian Binte Kabir
21241094
Maisha Islam
18101585
Mayesha Monjur
21241096

A thesis submitted to the Department of Computer Science Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2022

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Adiba Haque
21241097

Anika Nahian Binte Kabir
21241094

Maisha Islam
18101585

Mayesha Monjur
21241096

Approval

The thesis titled "Determining Fatal Heart Failure Risks in Patients Diagnosed with Chronic Kidney Disease: A Machine Learning Approach" submitted by

1. Adiba Haque (21241097)
2. Anika Nahian Binte Kabir (21241094)
3. Maisha Islam (18101585)
4. Mayesha Monjur (21241096)

of Fall2021, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Computer Science on 18-01-2022.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Khalilur Rhaman
Associate Professor
Computer Science and Engineering
Brac University

Co-supervisor:
(Member)

Moin Mostakim
Senior Lecturer
Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Computer Science and Engineering
Brac University

Head of Department:
(Chair)

_____
Dr. Sadia Hamid Kazi
Associate Professor
Computer Science and Engineering
Brac University

Abstract

The connection between heart and kidney has been medically established. The presence of a condition affecting one of the organs impairs the other. Hence, the causal associations between two of the most long-term conditions: Chronic Kidney Disease and fatality of Heart Failure are supported by the outcomes of Machine Learning techniques. The novelty of this research lies in its techniques that successfully find patterns in on of the stages of Cardiorenal Syndrome. We employed two disease datasets to perform predictive analysis with five classifiers: Random Forest, XG-Boost, CatBoost, Logistic Regression and Support Vector Machine, and analyzed the feature importance scores of the models to gauge the relationship between the conditions. The top predictors in our research were Random Forest, XGBoost and CatBoost classifiers with accuracy of the models for heart failure ranging from 70% to 76% and the accuracy for CKD prediction varied from 97% to 99%. Numerous features of the top predictors of HF and CKD were shared such as serum creatinine and diabetes. Individually, the CKD dataset ranked highest importance for haemoglobin levels, the imbalance of which causes anemia, and anemia is a key component in the HF dataset. The results of the visualization techniques of ML also yielded outcomes that were medically sound. This analysis of the physiological attributes and their importance with the help of machine learning, aided in successfully reaffirming the medical findings of a crucial stage of Cardiorenal syndrome.

Keywords: Cardiorenal; Machine Learning;Feature Importance; Cardiovascular Disease; Chronic Kidney Disease

Acknowledgement

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Chronic diseases have a tendency to progress into other life-threatening conditions in the future [1]. Chronic Kidney Disease (CKD) and Cardiovascular Disease (CVD) are two such long-term and persistent diseases. Chronic kidney disease is a condition that comprises deterioration of kidney function over a period of time. Heart Failure (HF) is a condition associated with cardiovascular disease that increases the risk of mortality in patients [2], [3]. Patients with heart failure can experience the adverse symptoms of chronic kidney disease and result in death or shorter life span [4].

Heart failure impairs the heart's capability to pump blood and often, pressure accumulates in the main pathway connecting the kidneys, causing a blockage. The kidney function exacerbates over time, due to insufficient oxygen or blood supply. This renders the kidney unable to remove wasteful substances from the blood. On the other hand, when the kidney function is compromised, it is unable to regulate blood pressure that exerts pressure on the heart to supply blood to the kidneys [5]. This can lead to structural changes in the heart that contribute to the condition of heart failure. In the medical sector, the primary focus of the study of **Cardiorenal syndrome** is on the pathophysiology of chronic renal end-stages, structural changes in the heart unique to kidney disease progression and it concludes that CKD is prevalent in 60% cases of heart failure (HF) [6]. In medical research the strong association of the two conditions have been proven and extensively studied. Identifying people with CKD, HF or those who are at risk of developing one of the conditions in the presence of the other, is a cornerstone of treatment or preventive assistance.

Our primary objective is to use machine learning techniques in order to develop personalized models that will separately predict chronic kidney disease (CKD) and fatal heart failure (HF) in patients. Then, using feature importance to gain insight into the predictive models and data, we identify an association between the two conditions [7-9], corroborating the medical theory linking terminal heart failure and chronic kidney disease.

Annually, more than 250,000 people die from heart failure in the United States. Among them 50% of the deaths are sudden [10]. In spite of the improvements in healthcare, heart failure may be difficult to diagnose initially [11]. Utilization of existing data on persistently occurring conditions is essential in finding subtle patterns

or behaviour of one condition in relation with another.

In previous literature, several studies have been conducted to apply machine learning techniques on patients' data. Abdar M. experimented with 10 machine learning algorithms to classify Coronary Artery Disease. The best three algorithms were optimized along with genetic algorithm and particle swarm optimization [12].

Muhammad L. J. uses instances from two Nigerian hospitals and applies traditional machine learning algorithms for prediction of CAD. Lastly, they visualized the influencing features by generating a decision tree [13].

Several techniques have been employed to detect chronic kidney diseases. In a research project, Pujari et al. proposes image processing techniques to classify stages of chronic kidney diseases [14]. Moreover, Almansour et al. implemented ANN and support vector machines for the prediction of kidney diseases with high accuracies [15]. Using the UCI Chronic Kidney Disease Dataset, Kunwar et al. implemented naïve bayes and ANN to draw a comparison [16] between the two techniques. Overall, it was observed that research endeavors have been made to improve disease prediction and as well as progression. However, limited research has been conducted to see how one disease could potentially progress into another.

More studies have successfully illustrated associations between conditions such as cardiovascular diseases and chronic kidney disease by employing open datasets [17-19]. However, in the status quo, machine learning methods have been used in limited ways to reaffirm associations between diseases. Chaudhuri A.K. et al [20] used data mining techniques to detect early cardiovascular disease in kidney disease patients. The authors have used Chi-square automatic interaction detection method to analyse and rank the features from the two datasets.

1.1 Scope and Novelty of the Research:

The novelty of our approach is twofold. We incorporate two existing datasets by UCI machine learning repository: one including clinical records for heart failure [21] and another containing records for chronic kidney disease [22], and employ classifier models to predict the outcomes for each of the instances. Then, we explore the relationship between the two closely related conditions that are: fatal heart failure and chronic kidney disease. In other words, we use machine learning to explore a stage of Cardiorenal syndrome, which has not been investigated using machine learning in previous literature. Medical journals have repeatedly revealed the strong relationships between the two conditions.

The potential for integrating the two disciplines is endless. A well developed machine learning approach to analyse cardiorenal syndrome can aid in early diagnosis of heart failure in kidney disease patients and the reverse. Machine learning algorithms consider variables like ejection fraction, anemia or creatinine levels to predict the diseases concerned. More often their decisions are dependent on a set of features

with varying importances. When assessing the feature relevance scores, it is essential to select the best performing models. In our case, the best performing models were: Random Forest, XGBoost and CatBoost. Moreover, out of all 13 features of heart failure, about 6 of them were common with chronic kidney disease feature set which confirms that the diseases are related to a large extent. Our interest, therefore lies within these common features and their correlations with the other unique features for each condition.

Machine learning is becoming increasingly important in the realm of healthcare. In situations involving two or more diseases where practitioners often find it difficult to gauge disease progression and their interrelations [23], machine learning can aid finding patterns. Thus, we devised this approach to study a set of chronic diseases and understand their patterns in crucial stages in order to confirm biological insights. In the first stage of our experiment we used classification algorithms to predict disease outcomes from the open datasets. Then, we determined the best predictors or importances of the top features. Additionally, we evaluate the correlation matrix with a similar aim. Lastly, we employ data visualisation techniques to prove the hypothesis stated in medical literature.

1.2   Thesis Organization:

The work is organized into following sections: In section 2, we discussed the relevance of this study and the purpose for it, the breakdown of the models and the reasons for their success (or lack), and finally, the alternative strategies we used. Section 3 overviews the experimental setup that demonstrates the work progression, description of the datasets, preprocessing techniques, hyperparameter tuning, discussion of the best predictors; followed by section 4, where the results and feature contributions are extensively discussed. Section 5 concludes the paper.
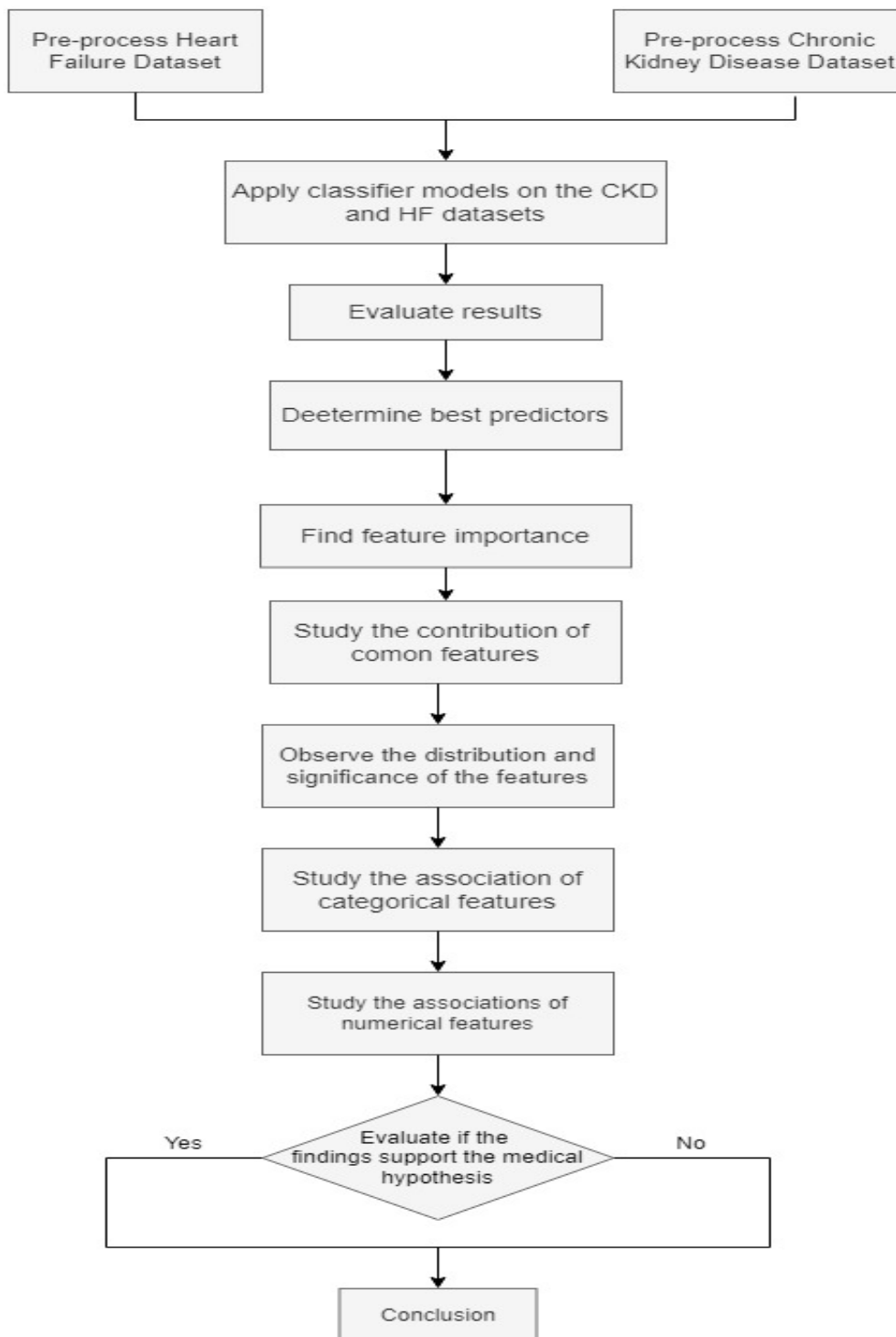
Figure 1.1: Workflow Diagram of Our Research

Chapter 2

Related Work and Theoretical Background

Prior medical research has frequently established a relationship between the two terminal conditions by structured review assessment of the literature to assess the evidence of heart failure in people with chronic kidney disease or vice-versa [24]. While each of these health states is linked with a high rate of morbidity or mortality, the presence of the dyad indicates an even worse prognosis. Thus, using machine learning to determine the association between these conditions can aid in forecasting the likelihood of a patient having one ailment getting the other.

In our experiments, we have used Logistic Regression, Support Vector Machines, XGBoost classifier, Random Forest classifier and CatBoost classifier for classification. Then we fine-tuned the models to choose a set of optimal parameters for the algorithms. We further applied K-fold cross validation to assess the model performance on data and observe models' improvements. The next stage was to determine the feature importance of the best predictor models and to compare the common contributing features of the two medical conditions. Finally, we sought to establish a mathematical relationship between chronic kidney disease and heart failure.

2.1   Significance of the Research:

We developed a concept for this research that would combine artificial intelligence with the domain of healthcare. The contributions include the following:

- Machine learning techniques can be used to find hidden patterns accurately and aid in the prognosis of diseases and conditions. This further reduces the diagnostic tests for chronic diseases, thus reducing the overall medical cost.

- Classification, association and clustering in machine learning contributes to finding traits efficiently. Each of these has a unique set of guidelines when applied on clinical data. Effective application of these strategies will benefit in the understanding of the progression of chronic diseases.

- A methodology based on artificial intelligence can provide significant insights and aid physicians in making more informed decisions about patient care. This work is a significant step forward towards that direction.

- The UCI repository harbours several datasets pertaining to different diseases, which are essentially lab reports of instances. Our work demonstrates the use of two openly available and easily accessible datasets that can be used in sophisticated research investigations.It proves to be beneficial to rely on documented records or big data and ML techniques, which make the process to procure results and links of diseases faster and more effective.

- This research can serve as a trajectory for the future research conducted in finding association of diseases.

2.2 Logistic Regression

Logistic regression is a supervised machine learning approach that provides a straightforward yet efficient solution to binary and linear classification problems. It is an extensively employed linear algorithm with a non-linear transform on output. Logistic regression removes noise as it considers removing outliers and possibly mis-classified instances from the training data. Our experiment highly relies on determining feature importance and logistic regression can interpret model coefficients as indicators of feature importance. In addition, during hyperparameter tuning we included L1 regularization, which helps avoid over-fitting in these cases, even though logistic regression is very little inclined towards the possibilities of over-fitting. In addition, we used L1 regularization during hyperparameter tuning to assist avoid over-fitting in these situations, even though logistic regression has a low tendency to over-fit.
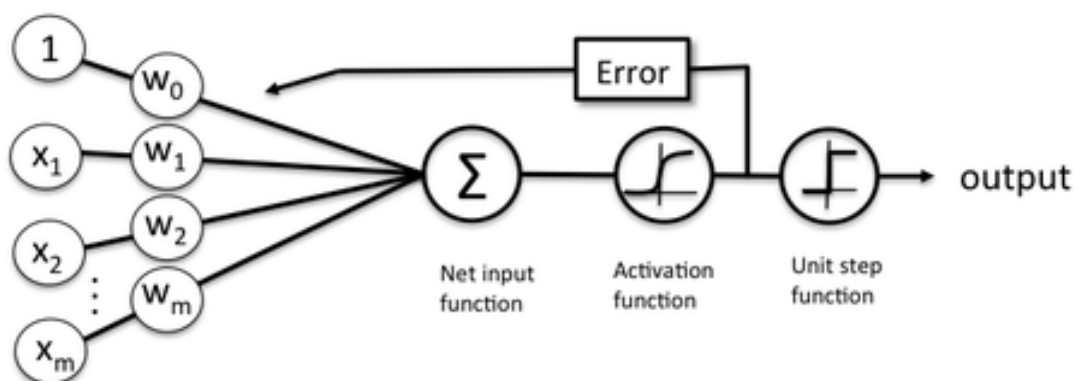
Schematic of a logistic regression classifier.

Figure 2.1: Architecture of Logistic Regression

2.3 Support Vector Machine

Support Vector Machine(SVM), one of the most popular supervised learning algorithms, is used for both classification and regression problems. The model(s) creates a decision boundary which can divide n-dimensional space into classes. This makes the process of putting a new datapoint into the correct category easier for future steps. SVM decides the hyperplane, which is the best decision boundary, by choosing support vectors and data points that are closest to the hyperplane. In our experiment, we have used SVC out of the many classes of SVM. The training complexity of the model depends highly on the size of the dataset, hence it produces higher accuracy for our datasets. However, in order to avoid the possibilities of overfitting, L1 regularization was included during hyperparameter tuning.
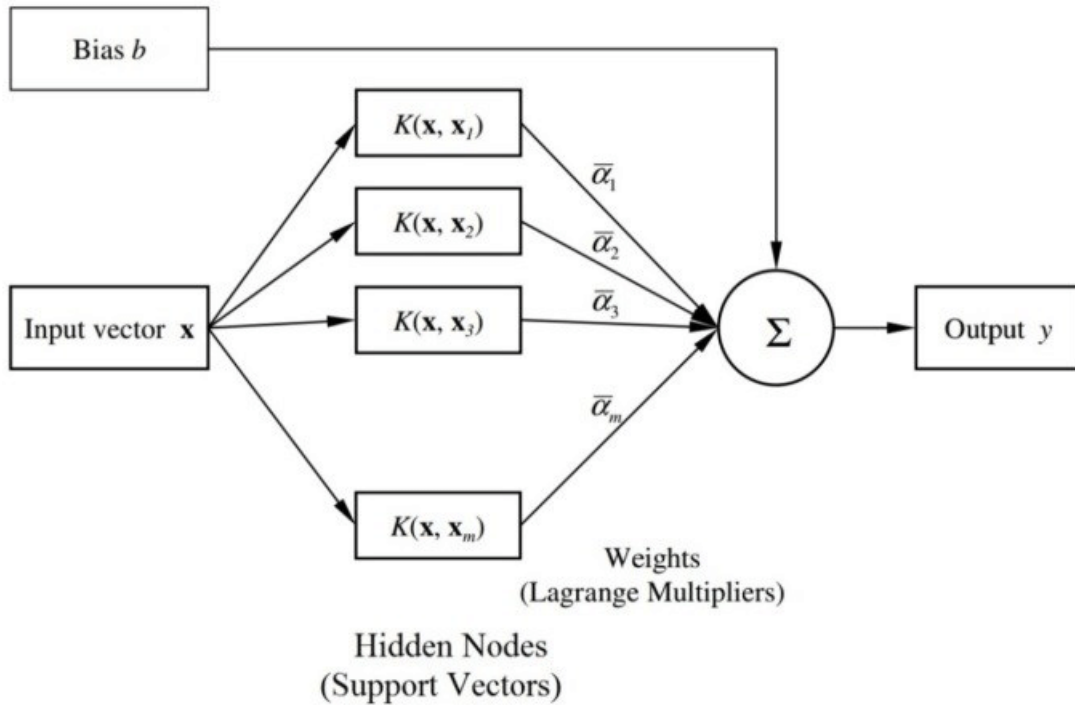
Figure 2.2: Architecture of SVM

2.4 Random Forest

Random forests are ensemble learning models consisting of a large number of individual decision trees that can be used for classification, regression and other tasks. The large number of relatively uncorrelated models act as a committee. The whole dataset is used to build individual trees unless sub-sample size is controlled with the max_samples parameter - bootstrap=True(default). The use of bagging(bootstrap aggregation) and featuring randomness while building each tree in order to create uncorrelated trees leads to improved predictive accuracy as well as overfitting-control, ultimately making random forest perform well in most datasets.
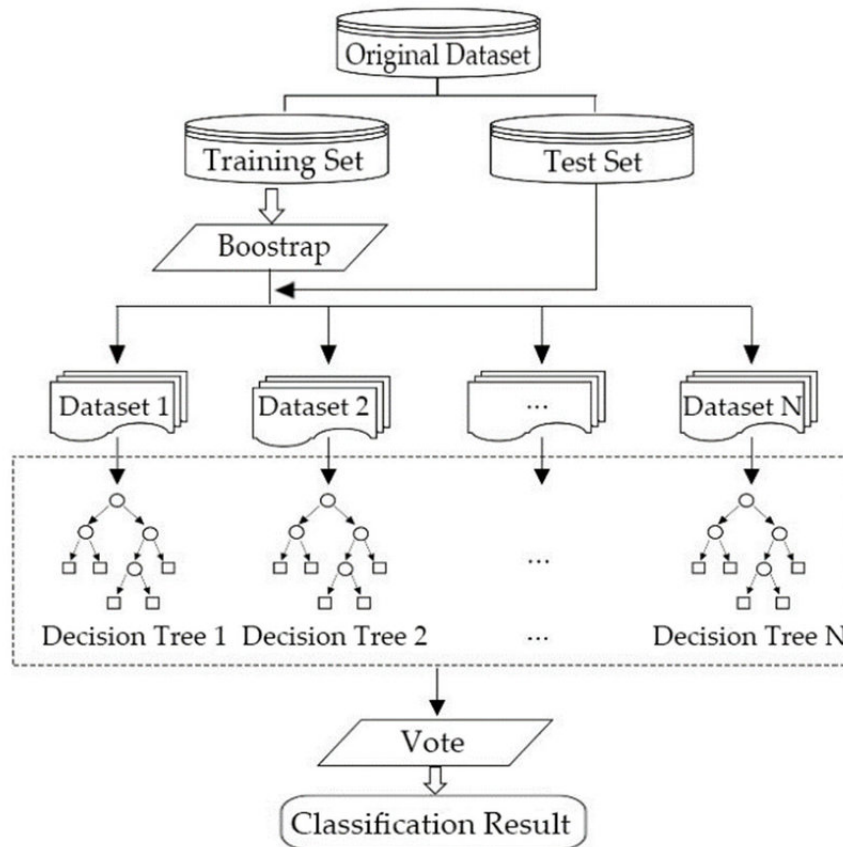
Figure 2.3: Random Forest Classifier

2.5 Gradient Boosting

Gradient Boosting can be realized as a differentiable loss function that is to be minimized. The third step, the weak learners are aggregated. The gradient descent algorithm is used to minimize the loss when adding learners.

2.6 XGBoost Classifier

XGBoost is an ensemble classifier that is built upon decision trees and Gradient Boosting. The individual decision trees, known as weak learners, decide which features are important and need focus. The misclassified examples are assigned higher weights.
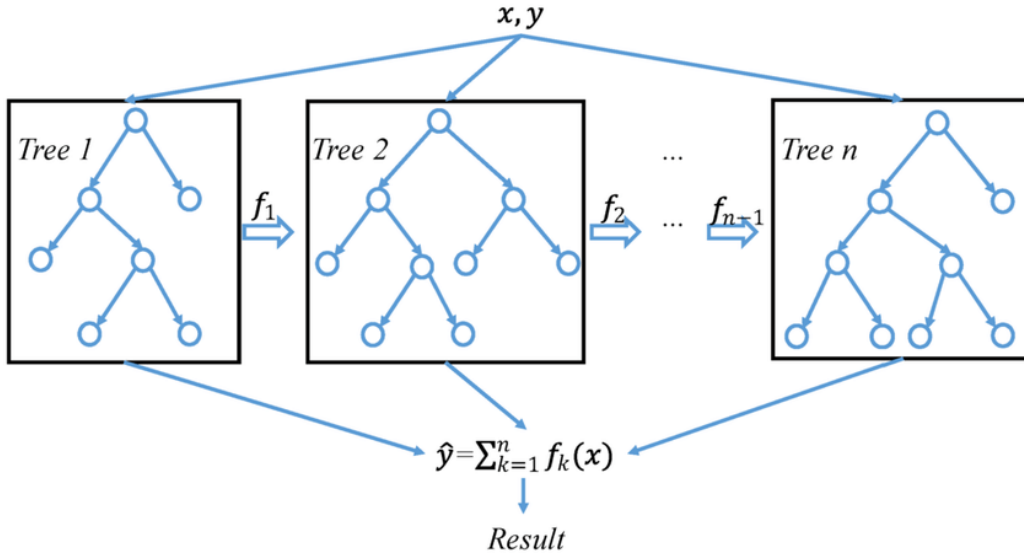
Figure 2.4: XGBoost Classifier

2.7 Explored Techniques

Initially, we intended to use nine models on both the datasets. However, we were unable to apply Naive Bayes classifier on the CKD dataset due to the non-normal distribution of the numerical data. For classification we also employed KNN classifier, a non-parametric algorithm which uses data with several classes to predict the classification of the new sample point. The classifier was then discarded because our study is based on feature importance, which is not defined in the KNN classification technique - the strategy for identifying significant features for this model during training involves feature selection method, which is not included in the scope of our work.

The variables contained in the dataset consist of both categorical and continuous types. K-means clustering requires variables to be strictly numerical. However, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) can find clusters even in the presence of outliers. It forms clusters based on regions of high density. It consists of two hyperparameters: minPoints and epsilon. To initiate the process the algorithm starts by randomly locating a point and then finds a minimum of points separated by distance of epsilon. DBSCAN uses Euclidean distance by default. However, it also avails the option of using Gower's similarity coefficient which allows one to compare two elements and assigns terms based on the nature of the feature. The outcome of this was a single cluster assigned the number 0 since the numerical features had a limited range. Moreover, to validate the nature of the merged dataset, hierarchical clustering was used. The dendrograms formed had no distinguishable cluster.

2.8 Correlation Matrix

Generally, correlation matrices are the statistical measure of strength of the relationship between two variables. In machine learning we use correlation to determine the relationship between the data features. However, during the investigation of our work we have learnt that studies in the medical sector provide evidence of causal assertion overreach drawn from correlational findings. In finding patterns between interconnected diseases, correlation matrices were deemed as unreliable because the critical evaluation in a medical work requires statistical knowledge in addition to clinical knowledge. Furthermore, coefficients are often influenced by outliers which may skew the actual relationship [25]. We evidently reached the same conclusion using machine learning approaches as well. For instance, for the heart failure correlation matrix, ejection fraction only showed a negative correlation of only 0.32 and serum creatinine showed a correlation coefficient of positive 0.36. For CKD, serum creatinine had a correlation of 0.3, despite the medical hypotheses. Hence some of the correlations are statistically underestimated and it is not completely reliable when predicting the importance of features [26].
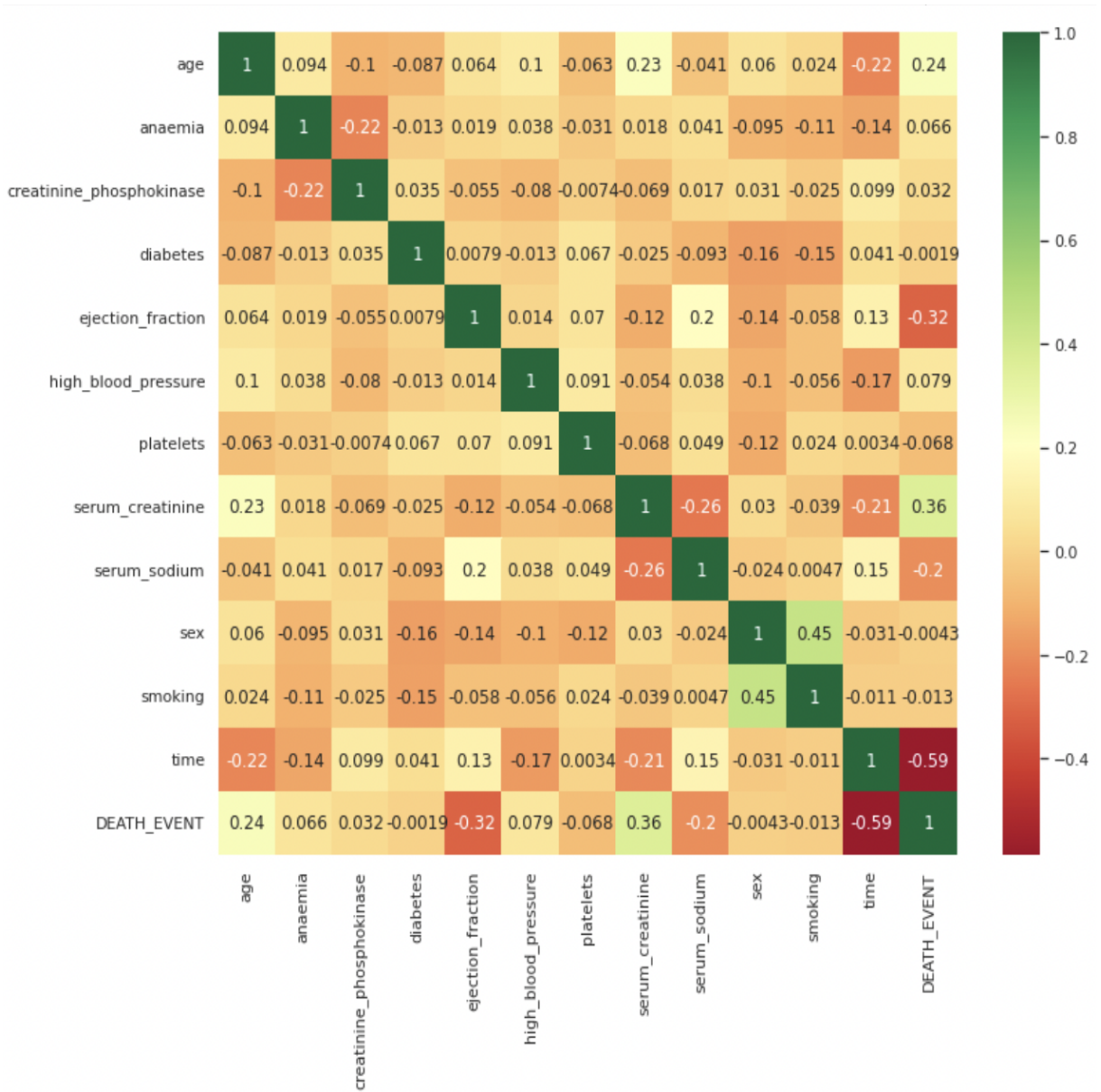
Figure 2.5: Correlation Matrix for Heart Failure

Figure 2.6: Correlation Matrix for Chronic Kidney Disease

Chapter 3

Experimental Setup

3.1 Dataset

We collected the datasets from UCI Machine Learning Repository. The Heart Failure Clinical Records Dataset with 13 features and 299 instances was used for the supervised training of the models to predict death due to HF. In this dataset, the instances fall into two categories - death due to heart failure and non-fatal heart failure, which has the potential to become fatal. The 13 characteristics include:

- age: age of the patient (years)

- anaemia: reduction of red blood cells or hemoglobin (boolean)

- high blood pressure: status of hypertension (boolean)

- creatinine phosphokinase: amount of the CPK enzyme in the blood (mcg/L)

- diabetes: presence of diabetes (boolean)

- ejection fraction: amount of blood leaving the heart with each contraction (%)

- platelets: platelets in the blood (kilo platelets/mL)

- sex: gender (binary)

- serum creatinine: amount of serum creatinine in the blood (mg/dL)

- serum sodium: level of serum sodium in the blood (mEq/L)

- smoking: the patient smokes or not (boolean)

- time: follow-up period (days)

- [*outcome*] death event: whether the patient deceased from HF (boolean)

We used Chronic Kidney Disease Dataset with 25 features and 400 instances, for the training of CKD prediction. The individuals in the chronic kidney disease dataset are divided into two categories - patients with chronic kidney disease and those without kidney disease. In this dataset the 25 are:

- age: age in years (numerical)

- blood pressure: bp (mm/Hg)

- specific gravity (nominal): sg (1.005,1.010,1.015,1.020,1.025)

- albumin: represented as al (nominal)

- sugar: represented as su (nominal)

- red blood cells: rbc count being normal or abnormal

- pus cell: pc count is normal or abnormal (nominal)

- pus cell clumps: pcc- presence or absence (nominal)

- bacteria: ba is present or not present (nominal)

- blood glucose random: bgr (mgs/dl)

- blood urea: bu (mgs/dl)

- serum creatinine: sc (mgs/dl)

- sodium: sod (mEq/L)

- potassium: pot (mEq/L)

- hemoglobin: hemo (gms)

- packed cell volume: pcv (numerical)

- white blood bell: wc count (cells/cmm)

- red blood cell: rc count (millions/cmm)

- hypertension: presence or absence of hypertension, htn (binary)

- diabetes mellitus: presence or absence of dm (binary)

- coronary artery disease: cad (binary)

- appetite: appet being good or bad (binary)

- pedal edema: pe (nominal)

- anemia: ane (nominal)

- [*outcome*] CKD: weather the patient has CKD or not (binary)

Generally, reducing the features to only the most relevant ones when training machine learning models plays an important part in the training process. In the case of chronic diseases, models' in-built feature importance scores can be used to infer the most contributing attributes to the disease. Hence, we used a subset of a larger set of medical indicators of the two conditions. The datasets were split for the training and test sets at 80:20 ratio (respectively).

3.2 Preprocessing of the Input Data

As the first step we have applied log transform to the continuous features of the heart failure dataset to correct the skewness of the variables. The original goal was to make the continuous variables in the heart failure dataset have log-normal distribution. The continuous variables such as creatine phosphokinase, ejection fraction, platelet counts and serum sodium were scaled using MinMaxScaler in order to transform the features to a particular range. This is supposed to refine the performance of the models used in our experiment. In the CKD dataset StandardScaler was used to transform red blood cell counts, white blood cell counts, packed cell volume, haemoglobin, serum sodium, blood urea, blood glucose random, blood pressure and age to ensure the skewed distributions have zero mean and unit variance. Potassium, serum creatinine, pus cell, red blood cell, sugar, albumin and specific gravity were scaled using MinMaxScaler. StandardScaler shifted the mean to zero; however, the skewness of the distributions did not change due to the presence of outliers. Hence the Yeo-Johnson transformation technique was used to improve the distribution curve (Figure 8). Unlike Box-Cox transformation, Yeo-Johnson transformation does not require the variables to be positive [27].
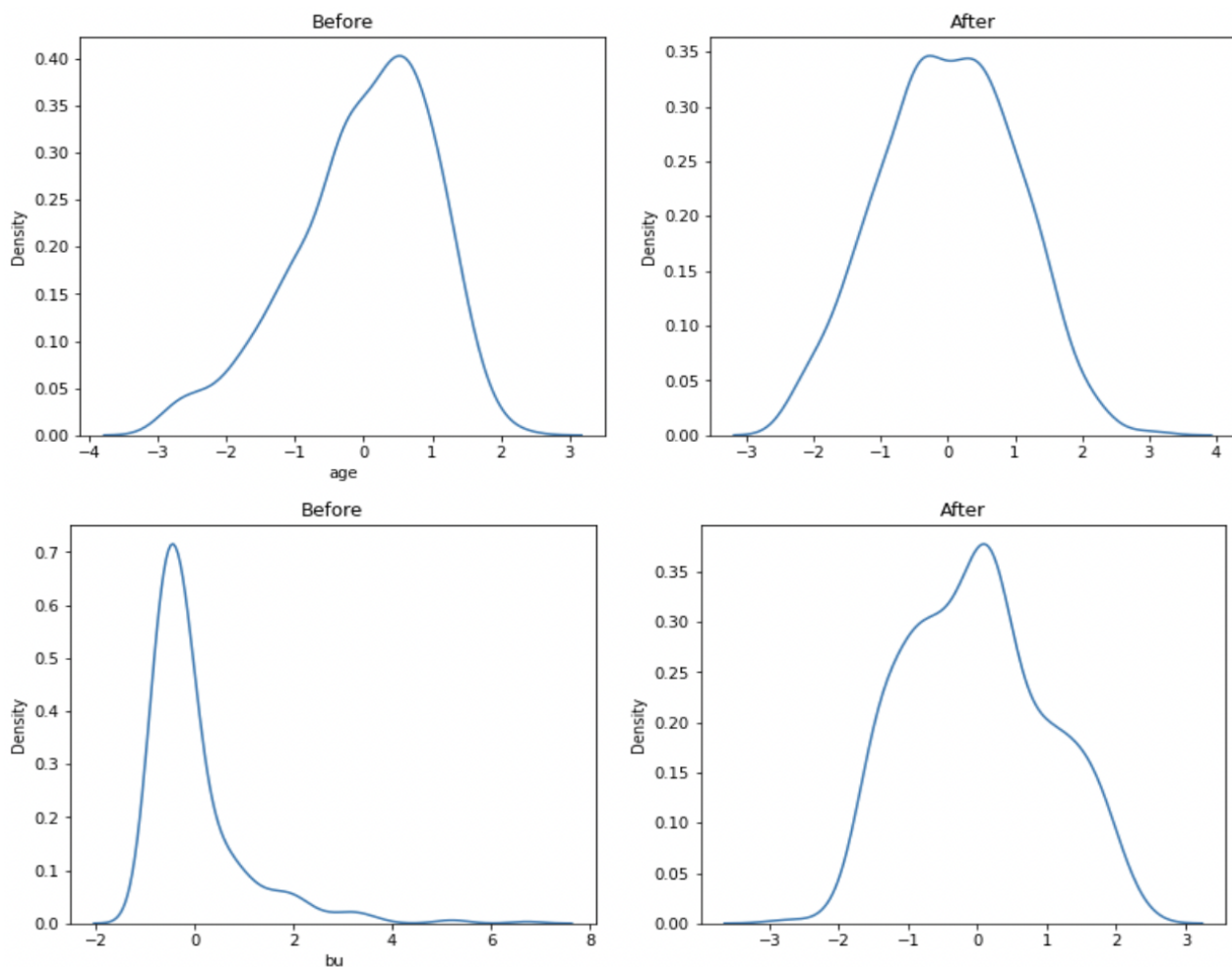

Figure 3.1: Comparison of Before and After Yeo-Johnson Transformation

The heart failure records dataset did not have any missing values in general. However, the Chronic Kidney Disease Dataset had substantial missing data, which were handled via missing data imputation. Traditional missing data imputation techniques such as replacement with mean, mode or most occurring values are not suitable for prediction of diseases. Hence, we employed the K-Nearest Neighbour imputation technique for handling missing data. KNN imputation method involves usage of Euclidean distance measure and finding out the contributing measure for each of the predictions. It finds the samples which are the closest to it and fills in with the mean of the nearest samples.

3.3 Hyperparameter Tuning

The best outcomes for heart failure prediction were created with the addition of L1 penalty to the coefficients. To minimize the cost function, the SAGA solver is utilized. Moreover, SVM was optimized with the kernel parameter set to 'poly' and 15 maximum iterations. The XGBoost classifier was implemented with 1000 estimators and maximum depth set to 3. The learning rate was set to 0.001. For the Random Forest classifier, the optimal depth was 10, number estimators 500 and with random state set to 1. Several machine learning models were fine-tuned for Chronic Kidney Disease prediction as well. The SVM could achieve high generalization ability with a linear kernel and 15 iterations. The XGBoost classifier was initialized with 1000 estimators and maximum depth of exactly 3.

3.4 Determining the Best Predictors

At the outset, seven machine learning models were applied on both the datasets. From a theoretical perspective Naive Bayes algorithm has the underlying assumption that all the features contribute equally to the outcome of classification. Due to the nature of the K-Nearest Neighbour Classifier, its feature importance scores cannot be calculated using traditional techniques. Hence, we have eliminated the aforementioned models.

We then narrowed down to the best five classifiers, all of which produced high accuracy on the chronic kidney disease Dataset. The Random Forest classifier fared exactly as well as XGBoost and CatBoost, which performed the best on the test data. The Logistic Regression classifier also generated good results.

Lastly, the linear and kernel SVCs were used for prediction and they yielded very favourable outcomes. These models exhibited further improvements with K-fold cross validation.

We applied five of the above mentioned algorithms on the heart failure dataset. The models did not achieve the same level of accuracy for this dataset as they did

for the CKD dataset, but the findings were still reliable. The best predictors however, were the same as the ones in CKD dataset - XGBoost, CatBoost and Random Forest classifiers. Logistic Regression classifier generated close enough results and a good score. The SVC performed poorly in this dataset.

3.5 Finding Feature Importance of Top Predictors

Analysis of feature importances helps us gauge the degree of relevance. Apart from gaining a better understanding and performance of the model, it can lead researchers towards a better understanding of biological phenotypes or features of the dataset and the correlation between them . Furthermore, it is essential to analyse how inbuilt feature importance measures work. Random Forest algorithm computes its feature importances using Gini Importance and Mean Decrease Accuracy.

Random Forest's built-in feature importance method is reliable given that its performance is highly accurate on the testing data. XGBoosts' in-built feature importance measure is highly dependable due its high performance accuracy. Additionally, its feature importance results are more reliable than that of the linear models. While calculating the importance scores, it also takes into account the relationship between variables [28].

The top predictors in our experiment were Random Forest, XGBoost and CatBoost classifiers respectively.

3.6 Study of Contributions of the Common Features

The datasets included 13 and 25 features each. In order to determine whether there is a link between the two chronic diseases, we studied the patterns of their common features. The HF and CKD datasets shared six characteristic features, which included category and numerical variables. In the HF dataset, the presence or absence of hypertension (htn) corresponded to the binary valued high blood pressure (bp) column in the CKD dataset. Similarly anemia (ane) and diabetes mellitus (dm) in both the datasets were binary valued and the columns comprising the patients' serum sodium (sd), serum creatinine (sc), and age were numerical data.

The contributions of these common features to the prognoses of the diseases were carefully evaluated. Medical findings suggest that in individuals with chronic kidney disease, heart failure is highly prevalent. This is a stage of Cardiorenal syndrome which encompasses a spectrum of disorders in the heart-kidney, in which chronic dysfunction in one organ may induce chronic dysfunction in the other [29]. This study:

Cardiorenal Syndrome: Classification, Pathophysiology, Diagnosis, and Treatment Strategies: A Scientific Statement From the American Heart Association shows that serum creatinine is found in the skeletal muscle during the diagnosis CKD and the prognosis is usually leaning towards heart failure with a likelihood of over 60%.

Diabetes mellitus of type-2 is common and is associated with endothelial dysfunction and vascular permeability [30]. Anemia in patients with CKD significantly contributes to the increase of HF and ischemic heart disease (IHD) in individuals.

3.7 Observation of the Distribution and Significance of Features

We have employed a handful of visualization techniques to verify the hypotheses found in literature. Heatmaps were used to illustrate the relationship between the categorical and target variables. Furthermore, in order to investigate the continuous features that were in common such as serum sodium, serum creatinine and age, violin plots were used.

Chapter 4

Result Analysis and Discussion

4.1 Prediction Results

In our experiments we have applied five machine learning models which have yielded decent results. The table below shows the results corresponding to each of the models. We have recorded the results in the table below (Table 1).

Model	CKD Dataset	HF Dataset
Random Forest	99.3%	74.8%
XGBoost	98.21%	72.4%
CatBoost	97.9%	71.7%
Logistic Regression	97.5%	70%
Support Vector Machine	97.14%	51.4%

Table 4.1: Accuracy of Each Model for the Two Datasets

In order to analyse the feature importance, we will be taking into account the top three models: Random Forest, XGBoost and CatBoost.

4.2 Feature Importance and General Findings

In order to find the link between the two diseases we studied the features from the top predictors. The best performing algorithms were Random Forest, XGBoost and CatBoost so we studied the feature importance for each. Over 40% of the features in the heart failure dataset were in common with the kidney disease dataset. This section summarizes our findings on the behavior of common characteristic features such as diabetes, serum creatinine, age etc. and the relevance of other unique factors such as ejection fraction and platelet count. Then the patterns of the categorical and continuous features are shown for improved visualization.

For the above-mentioned predictors, features that received most importance were serum creatinine, serum sodium, age and diabetes mellitus which are all common in

the two datasets. **Serum sodium** is highly associated with risk of mortality from failing heart, in patients with chronic kidney disease [31]. A study of the American Heart Association shows that when CKD is diagnosed, **serum creatinine** is detected in the skeletal muscle, and the prognosis is typically heart failure with a probability of over 60% [26]. In the same study we have seen that, in the population with heart failure with reduced ejection fraction (HFrEF) and **diabetes mellitus (DM)** present, CKD is found in 71% cases.

Biomarkers	Characteristics/Site of Origin	Diagnostic Value	Prognostic Value
Serum creatinine	Skeletal muscle	AKI, CRS	HF, CRS

Table 4.2: Role of Serum Creatinine in Dignosis of HF

Study	n	Study Design	Population	CKD	Concomitant Therapy	Baseline Renal Function	Outcome in CKD Group
ARTS - HF^{186}	1066	RCT, finerenone with dosage uptitrated vs eplerenone	HFrEF with EF <40%, DM with CKD (GFR >30 cc/min per 1.73 m^2), CKD without DM (GFR 30-60 cc/min per 1.73 m^2) & Exclude: GFR <30 mL/min per 1.73 m^2 CKD: CKD >3a: 71% & NA & GFR 53 mL/min per 1.73 m^2	Exclude: GFR <30 mL/min per 1.73 m^2 CKD: CKD >3-a: 71%	NA	GFR 53 mL/min per 1.73 m^2	Decrease in BNP >30%: same in both groups Any adverse event: finerenone less than eplerenone (76.9%) except finerenone 15–20 mg (78.5%) Death, cardiovascular hospitalization, worsening CHF: finerenone better than eplerenone except finerenone 2.5–5mg Hyperkalemia: finerenone better than eplerenone except finerenone 15–20 mg

Figure 4.1: Role of Ejection Fraction and Diabetes Mellitus

The **ejection fraction (EF)** in the heart failure dataset was also of significant importance. Though ejection fraction was not a feature in the CKD dataset, several studies have found that chronic kidney disease is common in people with heart failure and preserved or reduced ejection fraction (HFpER/HFrEF) [32].

In CKD dataset, **haemoglobin** level is of highest importance. Haemoglobin is the primary protein in our red blood cells and in the condition of anemia the

haemoglobin level drops. **Anemia** is one the common contributing factors in the CKD and HF-datasets. Anemia risks are higher in chronic kidney diseases and anemia in heart failure is considered to develop due to a complex interaction of iron deficiency and kidney disease [33]. Anemia is also a significant element in the HF-dataset. This demonstrates that chronic kidney disease and heart failure are also interlinked in terms of haemoglobin levels and anemia.

We will show the feature importance yielded in each of the models Random Forest, XGBoost and CatBoost from the subsections of code. Throughout the sections, it can be seen that our feature importance results have been consistent with the aforementioned findings from the medical literature. In other words the features shared by the two diseases had high to moderate importance scores, yielded by Random Forest, XGBoost and Catboost classifiers.

Random Forest

For the heart failure records dataset, some of the shared features that ranked high were serum creatinine, serum sodium and age. Similarly, for chronic kidney disease dataset the importance of serum creatinine was moderately high, it occupied the third position. Furthermore, diabetes mellitus, hypertension and serum sodium also showed some importance in this case. Ejection fraction, which was not in common among the two datasets, showed highest importance in the heart failure records dataset. Hemoglobin showed the highest importance in the chronic kidney disease dataset. Anemia was a common feature among the two dataset and it was observed from the figures below that anemia was a contributing factor to heart failure. Haemoglobin, which is associated with anemia, on the other hand gained the highest importance in the kidney disease dataset.
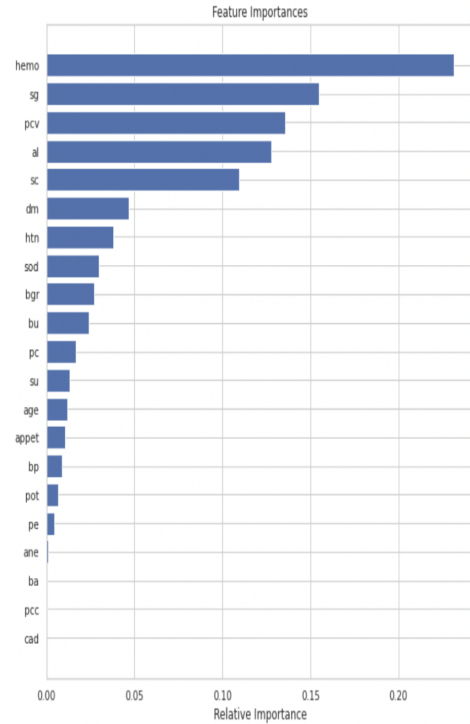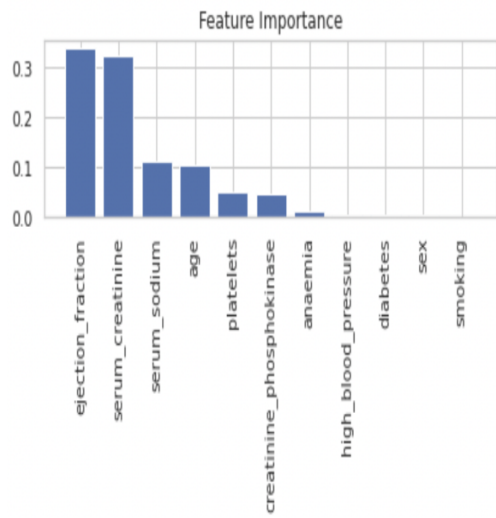
Figure 4.2: Feature Importance of Random Forest Classifier

XGBoost

For this model, serum creatinine, serum sodium, age and high blood pressure had high importance in case of heart failure contribution. Similarly to the previous results, serum creatinine had a high importance score, followed by hypertension, diabetes mellitus and serum sodium. Among the unique features, ejection fraction as a feature has consistently been identified as the most important contributing factor to heart failure. Likewise, haemoglobin was the most important feature in the kidney disease dataset. Specific gravity and albumin was shown moderate importance by XGBoost classifier.
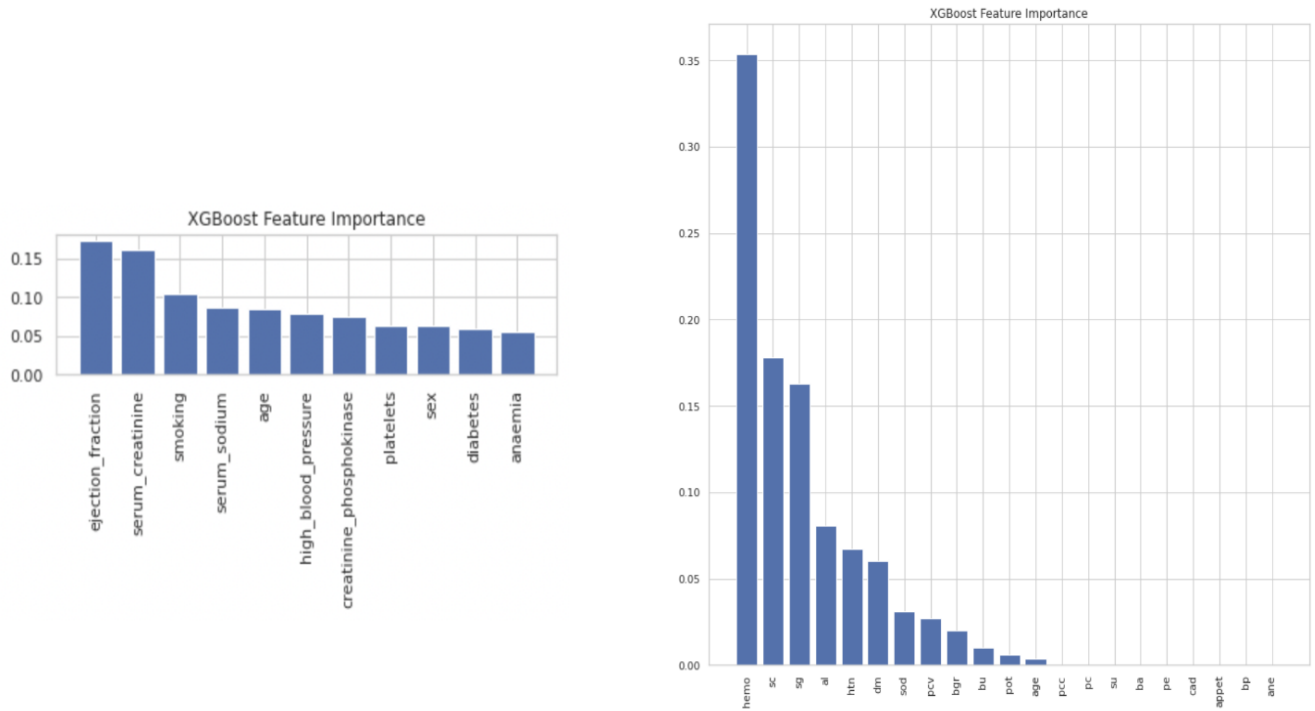
Figure 4.3: Feature Importance of XGBoost

CatBoost

Among the shared features, serum creatinine, age, serum sodium and diabetes mellitus had high to moderate importance in the prediction of heart failure. Haemoglobin levels, serum creatinine and serum sodium showed moderately high importance in detection of kidney disease. Specific gravity and albumin showed the highest importance score in case of Catboost, in the kidney disease dataset, unlike in the previous cases.
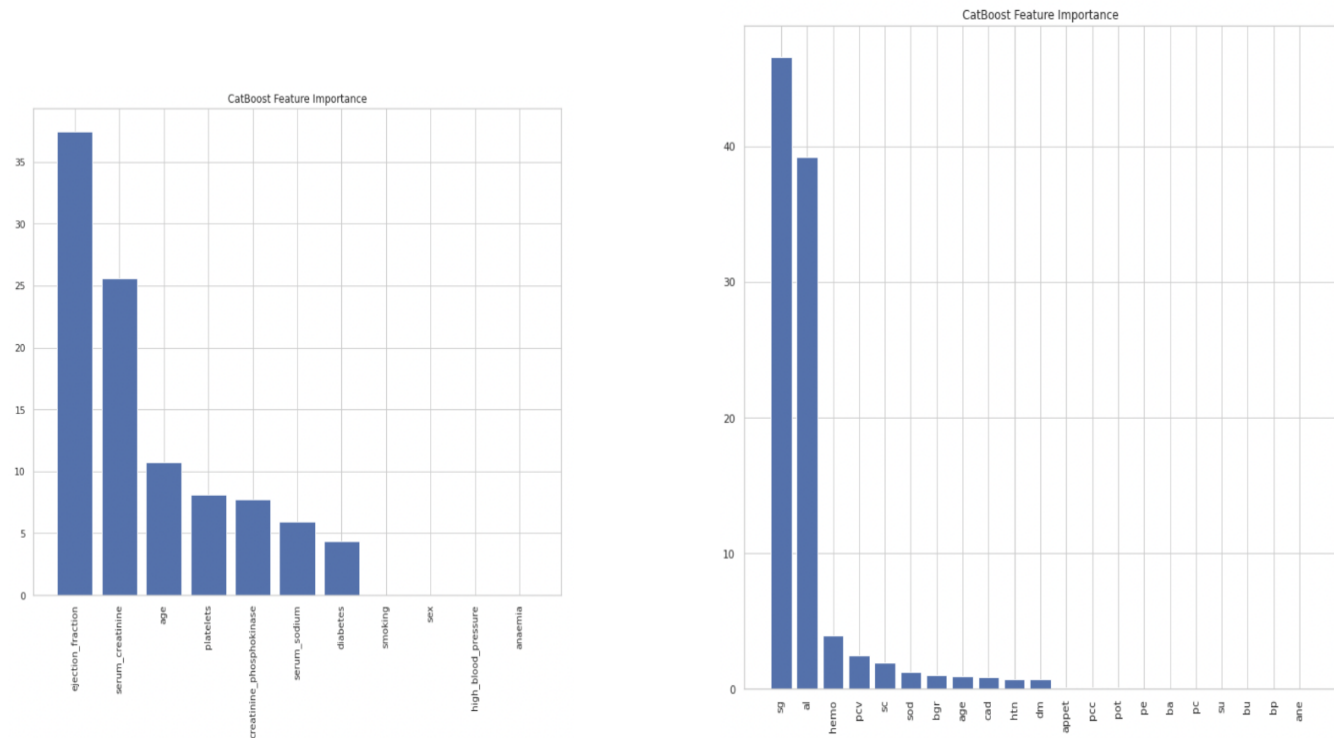
Figure 4.4: Feature Importance of CatBoost

4.3 Association between Categorical Features

Ensemble models such as XGBoost employ three different metrics for calculating feature importances: Gain, coverage and weights. Anemia, hypertension and diabetes have smaller numbers of possible values or branchings compared to other continuous variables. Each binary feature can be used once in the ensemble of trees. Hence feature importance measure could be less indicative of feature importances of binary variables in particular. In general, from the feature importance results, it was seen that categorical features had moderate to low importance scores.Therefore, the goal in this section is to analyse numerically how the categorical features have influenced the existence of both the conditions.

Diabetes

Among the 96 people who died from heart failure, about 40 of them had diabetes mellitus, which goes on to prove that it has a moderate impact on heart failure fatality, as shown by the heart failure feature importance figures. In the case of chronic kidney disease, it shows a high influence. About 250 patients in total were diagnosed with kidney disease and among them 137 of them had diabetes mellitus, making it an important complication in determining the presence of kidney complications.
From the findings above, we can conclude that diabetes mellitus as a common feature has a moderate to high impact on the presence or absence of the conditions and their state.
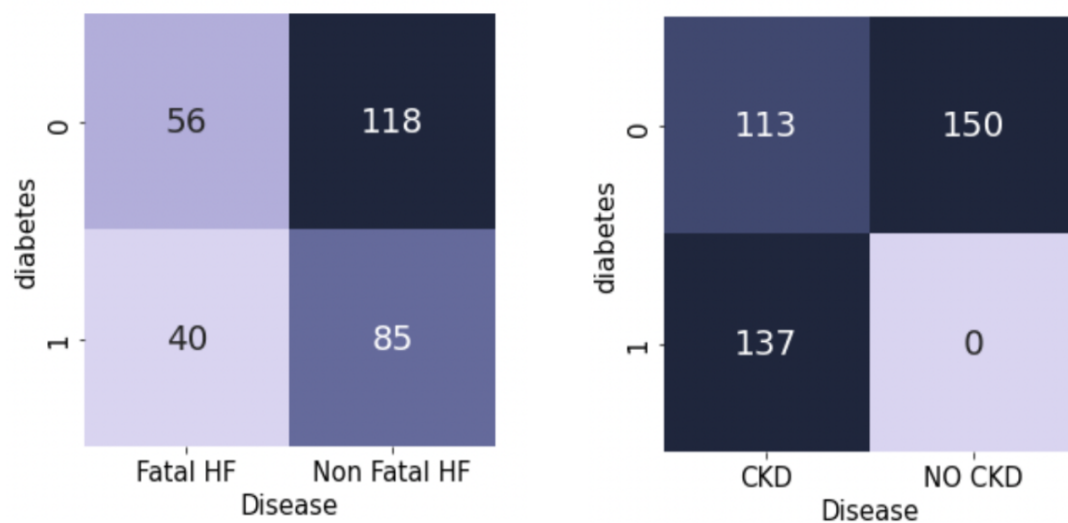
Figure 4.5: Influence of Diabetes

Hypertension

Hypertension had a low influence on the event of fatal heart failure, however it strongly impacted the presence of chronic kidney disease. From the sample of people who had heart failure, about 40% died from heart failure. On the other hand, 58.8% of the chronic kidney disease patients had hypertension. We may conclude from the findings, that the presence of hypertension does in fact increase the chances of chronic kidney disease. Likewise it increases the risk of heart failure in the presence of chronic kidney disease.
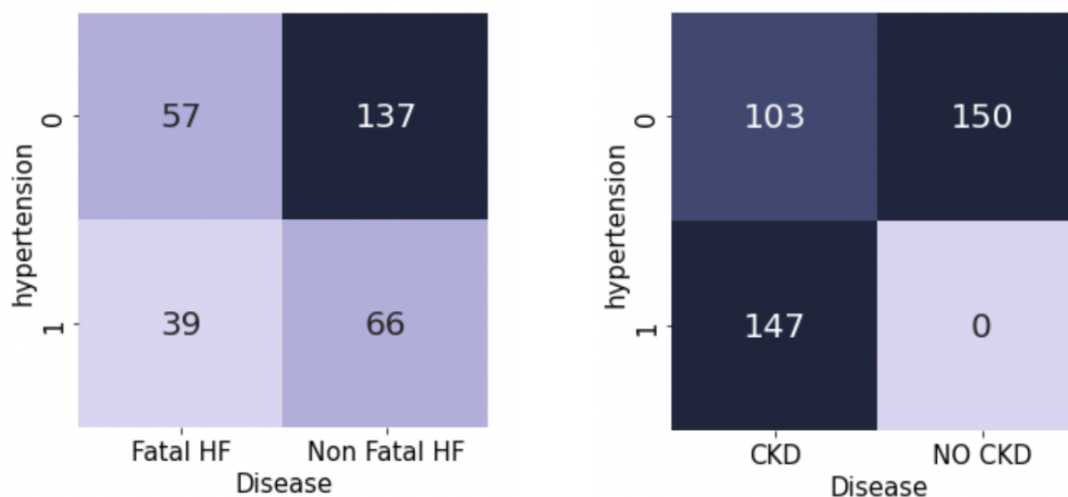

Figure 4.6: Influence of Hypertension

Anemia

From the figure below, it can be seen that among the sample of the population who were not diagnosed with chronic kidney disease, none of them were pre-diagnosed with anemia. Moreover, a moderately large subset of those diagnosed with the chronic condition were reported to have anemia. 48% of those who died from heart failure had anemia. Although anemia as a shared feature did not score significantly high in case of feature importances, numerical analysis showed that it is a highly influential factor in determining the presence of kidney disease and the possibility of the related condition, in this case, heart failure.
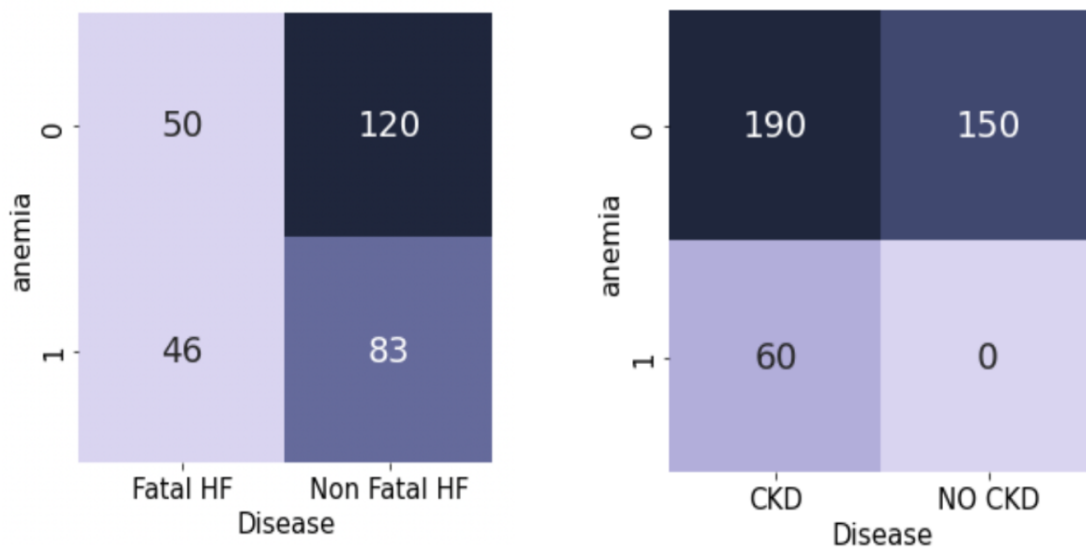

Figure 4.7: Influence of Anemia

4.4 Association between Numerical Features

In this section we analyze the numerical and continuous features to see their effect on the patients in presence of chronic kidney disease and heart failure. We have used the Violin Plot of Seaborn to visualize the quantitative categories and understand the association between the diseases. For the serum creatinine and serum sodium levels, the plots show similarity in patterns. The age distribution for the two diseases however are different, although this finding is more remarkable owing to the medical grounds supporting it. The violins are split in half to better understand the difference between two groups. For the heart failure analysis, the two parts reflect feature-behavior in cases of death from HF and non-fatal HF. Similarly, the left and right sides of the plot of chronic kidney disease data points show those of patients with and without chronic kidney disease respectively. Our work mainly focuses on the presence or fatality of the diseases i.e. the left part of the violin represented as HRT and CKD in the graph.

Serum Sodium

Serum sodium levels in cases of fatal heart failure vary from just over 120 mg/dl to 150 mg/dl, which, as seen in the plot, is closely comparable to the sd-levels in CKD. In terms of patterns, the outcome of this violin is particularly striking because the two violin plots show identical results for both presence and absence of individual diseases. The values are most concentrated at 135 mg/dl to 145 mg/dl in both CKD and HF-plots and past medical records revealed reasonably similar outcomes [34].
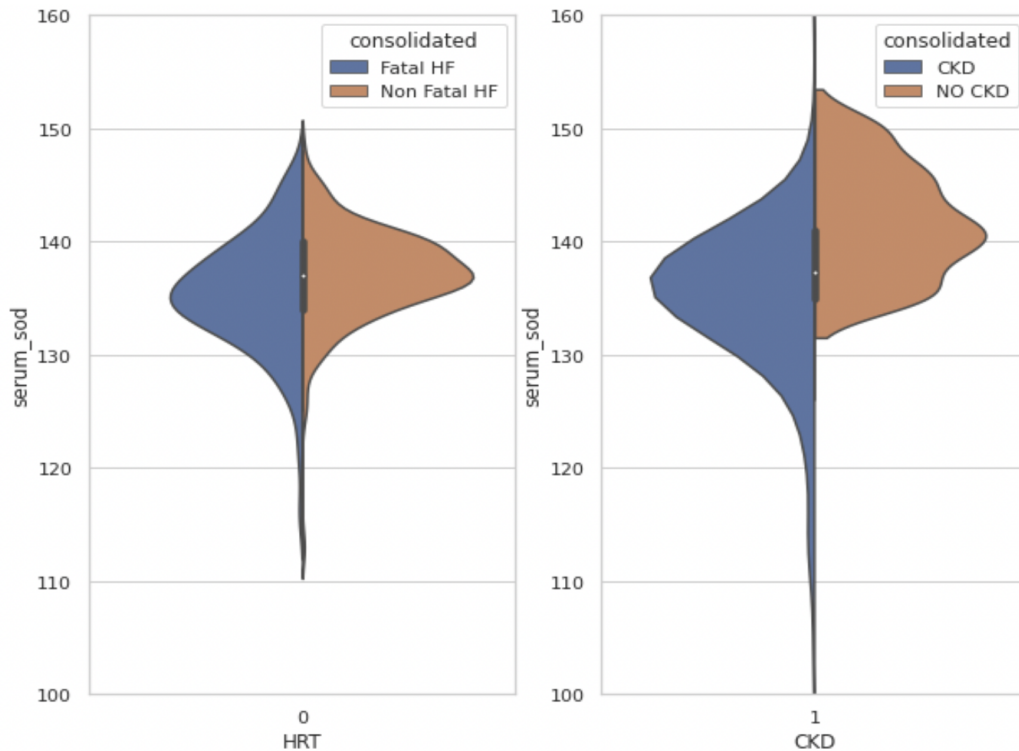

Figure 4.8: Violin Plot of Serum Sodium

Age

Here the violin represents the data points showing the age range of patients with heart failure is well distributed from 30 to above 90. It is comparatively a flatter curve indicating that congestive heart failure affects adults of all ages which coincides with the studies of the healthcare sector. The CKD dataset had some discrepancies and outliers that we plan to handle in the future. Due to the outliers in the dataset the violin of CKD appears to be skewed around the age range of 0 to 10 - there are data points on the negative axis as well, whereas we know age cannot be less than zero. Eliminating these outliers, the violin plot shows that the age of patients from the CKD dataset ranges from little above 40 to about 80. The slope of the curve is steeper and values are more concentrated around 60 to 65 in our plot. According to current estimates of medical research: CKD is more common in people aged 65 years or older (38%), than in people aged 45-64 years (12%) and 18-44 years (6%) [35]. Our finding agrees with the medical point of view on the fact heart failure is

seen in individuals of varying age while kidney disease is more prevalent in patients of a specific age.
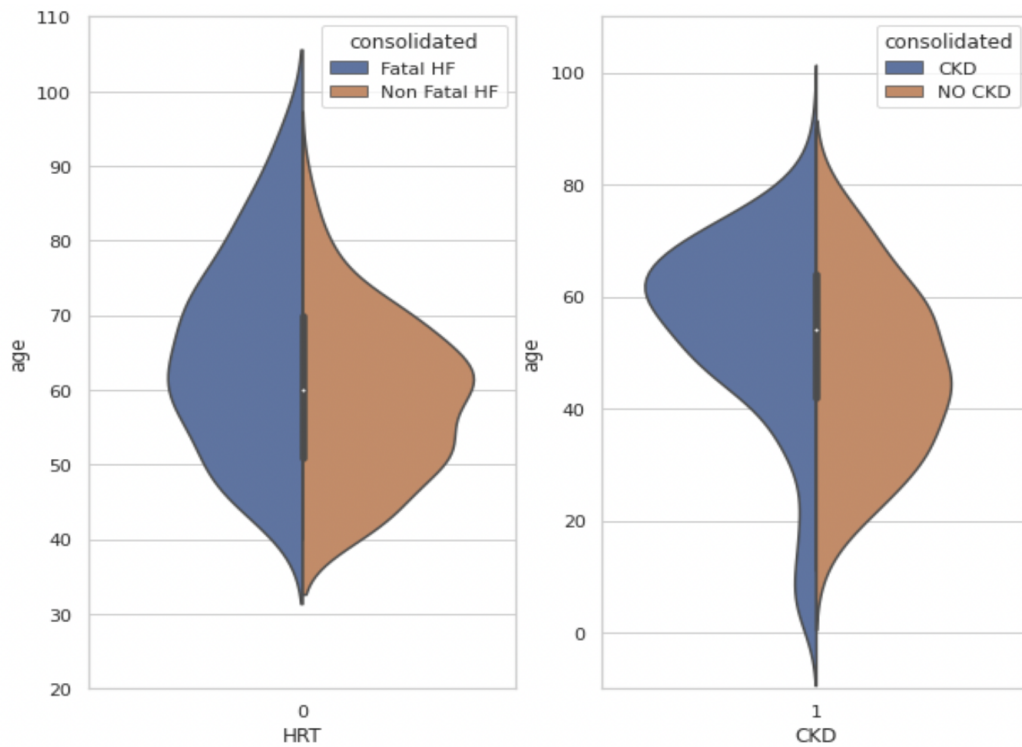
Figure 4.9: Violin Plot of Age

Serum Creatinine

The last common numerical data for violin plot representation is the serum creatinine level. From the two datasets, there is significant similarity in the sc-levels concentrated in the range of 1.5 to 4. In the event of death from heart failure, serum creatinine level has an increase at least 0.5 mg/dL from baseline and causes total loss of kidney function within 4 weeks or end-stage kidney disease by 3 months [36].

Figure 4.10: Violin Plot of Serum Creatinine

4.5 Evaluation

Heart failure and chronic kidney disease are associated and share many contributing factors. Leveraging actionable information extracted from the visualization and feat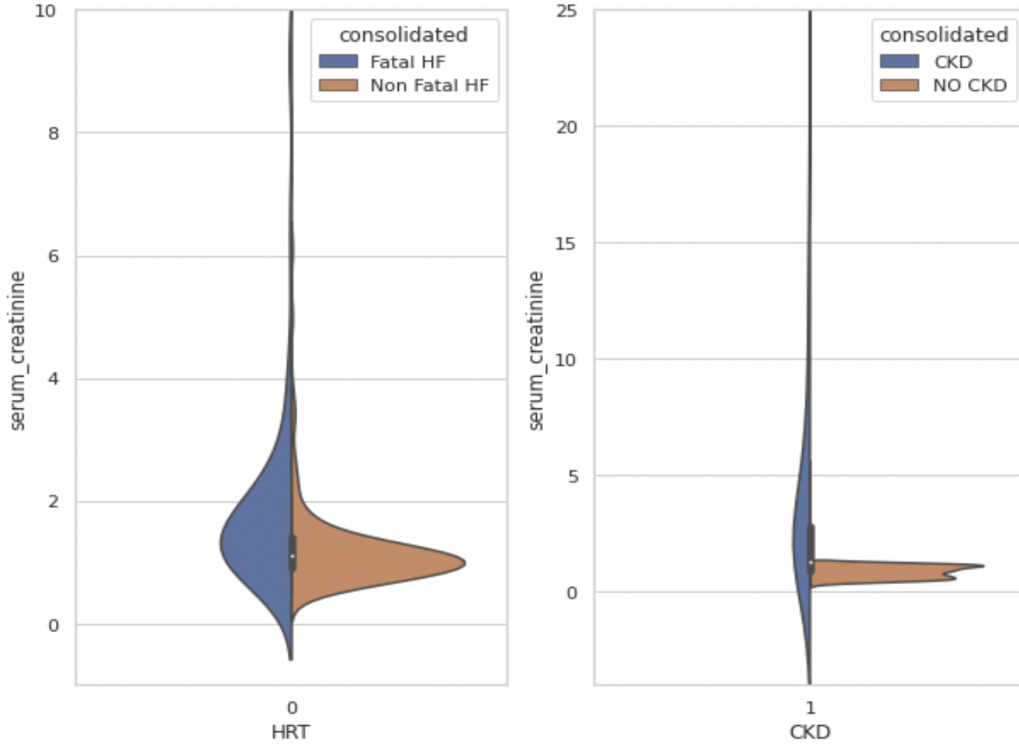ure importance, we arrive at the same conclusion as the medical hypothesis. It can be deduced from our research that machine learning can detect medical patterns in complex data, revealing the existence of a link between two diseases.

From the analysis of chronic kidney disease dataset, the following features had the highest importance:

$$RF \cap CatBoost \cap XGBoost = \{hemoglobin, albumin, serum\ creatinine,$$
$$pcv, sg, hypertension, serum\ sodium, bgr, bu\}$$

Analysis of the heart failure dataset yielded the following importance scores:

$$RF \cap CatBoost \cap XGBoost = \{ejection\ fraction, serum\ creatinine, age,$$
$$platelets, creatine\ phosphokinase, serum\ sodium, hypertension, diabetes\ mellitus\}$$

The aforementioned relationships have been thoroughly documented in previous

literature. In general, among both the diseases we found that serum creatinine, hypertension, serum sodium and diabetes mellitus were of high importance and the relationship may be denoted in the following manner:

$$CKD \cap HF = \{serum\ creatinine, hypertension, serum\ sodium, diabetes\ mellitus\}$$

4.6 Conclusion

The findings of this study can be applied to the domain of healthcare and artificial intelligence applications in the medical field. This paper aims to investigate two chronic conditions - chronic kidney disease and heart failure - and their relationship in order to demonstrate that developing one of these chronic illnesses significantly raises your risk of developing the other. Asymptomatic, chronic kidney disease can go undiagnosed until it has progressed to serious cardiovascular complications, including fatal heart failure. Again, the primitive stage of heart failure is a health crisis, and its presence alone confers a disproportionately high risk of developing CKD. Thus, in this research, we used an analysis of the physiological attributes of people to predict the probability of heart failure in patients diagnosed with chronic kidney disease (CKD) or vice versa, combining machine learning systems and conventional methods. We get sufficient comprehension of the attributes through the use of models and conclude that machine learning is capable of collecting information and patterns, as well as establishing the relationship between the two circumstances. Early diagnosis of these chronic diseases can enable clinicians in strengthening and implementing preventative and treatment remedial strategies aimed at reducing high morbidity and death. There is research employing imaging methods to detect risk factors for developing HF in people with chronic kidney disease [37]. Future studies may combine the image processing technology with data mining to more precisely determine the relevance of the features.

The findings of any research study is universal in nature. As a result, this extensive research builds a solid basis for the medical theory associating chronic kidney disease with heart failure.

Bibliography

1. Oishi, Y., & Manabe, I. (2020). Organ System Crosstalk in Cardiometabolic Disease in the Age of Multimorbidity. Frontiers in Cardiovascular Medicine. DOI: https://doi.org/10.3389/fcvm.2020.00064

2. Pickup, L. C., Law, J. P., Townend, J. N., & Ferro, C. J. (2021). Sudden cardiac death in chronic renal disease: aetiology and risk reduction strategies. Nephrology Dialysis Transplantation, 36(8). DOI: https://doi.org/10.1093/ndt/gfz232

3. Wright, J., & Hutchison, A. (2009). Cardiovascular disease in patients with chronic kidney disease. Vascular Health and Risk Management, 5, 713–722. DOI: https://doi.org/10.2147/VHRM.S6206

4. Shiba, N., & Shimokawa, H. (2011). Chronic kidney disease and heart failure—Bidirectional close link and common therapeutic goal. Journal of Cardiology Home, 57(1), 8-17. DOI: https://doi.org/10.1016/j.jjcc.2010.09.004

5. Heart disease & chronic kidney disease (CKD), American Kidney Fund. Available online:
https://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/complications/heart-disease/

6. Lin, L., Zhou, X., Dekkers, I. A., & Lamb, H. J. (2021). Cardiorenal Syndrome: Emerging Role of Medical Imaging for Clinical Diagnosis and Management. Journal of Personalized Medicine, 11(8), 734. DOI: https://doi.org/10.3390/jpm11080734

7. Ahmed, A., Rich, M. W., Sanders, P. W., Perry, G. J., Bakris, G. L., Zile, M. R., Love, T. E., Aban, I. B., & Shlipak, M. G. (2007). Chronic Kidney Disease Associated Mortality in Diastolic Versus Systolic Heart Failure: A Propensity Matched Study. The American Journal of Cardiology, 99(3), 393-398. DOI: https://doi.org/10.1016/j.amjcard.2006.08.042

8. McAlister, F. A., Ezekowitz, J., Tonelli, M., & Armstrong, P. W. (2004). Renal Insufficiency and Heart Failure Prognostic and Therapeutic Implications From a Prospective Cohort Study. Circulation, 109(8), 1004–1009. DOI: https://doi.org/10.1161/01.CIR.0000116764.53225.A9

9. Shlipak, M. G., Smith, G. L., Rathore, S. S., Massie, B. M., & Krumholz, H. M. (2004). Renal Function, Digoxin Therapy, and Heart Failure Outcomes: Evidence from the Digoxin Intervention Group Trial. Journal of the American Society of Nephrology, 15(8), 2195-2203. DOI: https://doi.org/10.1097/01.ASN.0000135121.81744.75

10. Tomaselli, G. F., & Zipes, D. P. (2004). What Causes Sudden Death in Heart Failure? Circulation Research, 95, 754–763. DOI: https://doi.org/10.1161/01.RES.0000145047.14691.db

11. Hobbs, F. D. R., Davis, R. C., & Lip, G. Y. H. (2000). Heart failure in general practice. BMJ, 320. DOI: https://doi.org/10.1136/bmj.320.7235.626

12. Abdar, M., Książek, W., Acharya, U. R., Tan, R., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 179. DOI: https://doi.org/10.1016/j.cmpb.2019.104992

13. Muhammad, L. J., Al-Shourbaji, I., Haruna, A. A., Mohammed, I. A., Ahmad, A., & Jibrin, M. B. (2021). Machine Learning Predictive Models for Coronary Artery Disease. SN Computer Science. DOI: https://doi.org/10.1007/s42979-021-00731-4

14. Pujari, R. M., & Hajare, V. D. (2014). Analysis of ultrasound images for identification of Chronic Kidney Disease stages. First International Conference on Networks & Soft Computing. DOI: https://doi.org/10.1109/CNSC.2014.6906704

15. Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., Alrashed, S., & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in Biology and Medicine, 109, 101-111. DOI: https://doi.org/10.1016/j.compbiomed.2019.04.017

16. Kunwar, V., Chandel, K., Sabita, A. S., & Bansal, A. (2016). Chronic Kidney Disease analysis using data mining classification techniques. International Conference on Confluence The Next Generation Information Technology Summit (Confluence). DOI: https://doi.org/10.1109/CONFLUENCE.2016.7508132

17. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making. DOI: https://doi.org/10.1186/s12911-019-0918-5

18. Chan, L., Nadkarni, G. N., Fleming, F., McCullough, J. R., Connolly, P., Mosoyan, G., Salem, F. E., Kattan, M. W., Vassalotti, J. A., Murphy, B., Donovan, M. J., Coca, S. G., & Damrauer, S. M. (2021). Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease. Diabetologia, 64, 1504–1515. DOI: https://doi.org/10.1007/s00125-021-05444-0

19. Pasha, S. M. (2020). Diabetes and Heart Disease Prediction Using Machine Learning Algorithms. International Journal of Emerging Trends in Engineering Research, 8(7). DOI: https://doi.org/10.30534/ijeter/2020/60872020

20. Chaudhuri, A. K., Ray, A., Das, A., Chakrabarti, P., & Banerjee, D. K. (2020). Early Detection of Cardiovascular Disease in Patients with Chronic Kidney Disease using Data Mining Techniques. Asian Journal of Convergence in Technology, 6(3). DOI: https://doi.org/10.33130/AJCT.2020v06i03.011

21. Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M. & Raz, M. A. (2017). Heart Failure Clinical Records Dataset, UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

22. Soundarapandian, P., Rubini, L. J., & Eswaran, P. (2015).Chronic_Kidney_ Disease Data Set. UCI Machine Learning Repository. Available online: https: //archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

23. Rocca, H. B., Fleischhacker, L., Golubnitschaja, O., Heemskerk, F., Helms, T., Hoedemakers, T., Allianses, S. H., Jaarsma, T., Kinkorova, J., Ramaekers, J., Ruff, P., Schnur, I., Vanoli, E., Verdu, J., & Zippel-Schultz, B. (2016). Challenges in personalised management of chronic diseases—heart failure as prominent example to advance the care process. EPMA Journal, 7(1). DOI: https: //dx.doi.org/10.1186/s13167-016-0051-9

24. Segall, L., Nistor, I., Covic, A. (2014). Heart Failure in Patients with Chronic Kidney Disease: A Systematic Integrative Review. BioMed Research International. DOI: https://dx.doi.org/10.1155/2014/937398

25. Mukaka, M. M. (2012). A guide to appropriate use of Correlation coefficient in medical research. Malawi Med Journal, 24(3), 69–71.

26. Hung, M., Bounsanga, J., & Voss, M. W. (2017). Interpretation of correlations in clinical research. Postgraduate medicine, 129(8), 902–906. DOI: https: //doi.org/10.1080/00325481.2017.1383820
27. Weisberg, S. (2001). Yeo-Johnson Power Transformations. Available online: https://www.stat.umn.edu/arc/yjpower.pdf

28. Ram, S. (2020). Feature Importance — Everything you need to know. Available online:
https://medium.com/swlh/feature-importance-hows-and-why-s-3678ede1e58f

29. Rangaswami, J., Bhalla, V., Blair, E. A. J., Chang T. I., Costa S., Lentine K. L., Lerma E. V., Mezue K., Molitch M., Mullens W., Ronco C., Tang W. H. W., & McCullough P. A. (2019). Cardiorenal Syndrome: Classification, Pathophysiology, Diagnosis, and Treatment Strategies: A Scientific Statement. Circulation, 139(16). DOI: https://doi.org/10.1161/CIR.0000000000000664

30. Ardhanari, S., Alpert, M. A., & Aggarwal, K. (2014). Cardiovascular Disease in Chronic Kidney Disease: Risk Factors, Pathogenesis, and Prevention. Advances in Peritoneal Dialysis, 30. Available online: https://www.advancesinpd.com/adv14/ 40-53_Ardhanari.pdf

31. Sun, L., Hou, Y., Xiao, Q., & Du, Y. (2017). Association of serum sodium and risk of all-cause mortality in patients with chronic kidney disease: A meta-analysis and sysematic review. Scientific Reports, 7. DOI: https://doi.org/10.1038/ s41598-017-16242-3

32. Mavrakanas, T. A., Khattak, A., Wang, W., Singh, K., & Charytan, D. M. (2019). Association of Chronic Kidney Disease with Preserved Ejection Fraction Heart Failure Is Independent of Baseline Cardiac Function. Kidney and Blood Pressure Research, 44, 1247–1258. DOI: https://doi.org/10.1159/000502874

33. Shah, R., & Agarwal, A. (2013). Anemia associated with chronic heart failure: current concepts. Clinical Interventions in Aging, 8, 111—122. DOI: https://doi.org/10.2147/CIA.S27105

34. Singh, P., Deshwali, S., Potey, G. G., & Sharma, A. (2017). Study of serum sodium and serum potassium level in chronic renal failure. International Journal of Advanced Research in Biological Sciences, 4(11), 103-111. DOI: http://dx.doi.org/10.22192/ijarbs.2017.04.11.013

35. Centers for Disease Control and Prevention (2021). Chronic Kidney Disease in the United States. Centers for Disease Control and Prevention. Available online: https://www.cdc.gov/kidneydisease/publications-resources/CKD-national-facts.html

36. Metra, M., Cotter, G., Gheorghiade, M., Cas, L. D., & Voors, A. A. (2012). The role of the kidney in heart failure. European Heart Journal, 33(17), 2135–2142. DOI: https://doi.org/10.1093/eurheartj/ehs205

37. Nanasato, M., Matsuo, S., Nakajima, K., Nishimura, S., & Nishimura, T. (2020). Predictive value of electrocardiography-gated myocardial perfusion imaging to new-onset heart failure in patients with chronic kidney disease: findings from the J-ACCESS 3 study. The International Journal of Cardiovascular Imaging, 36, 749–755. DOI: https://doi.org/10.1007/s10554-019-01761-z