Comparison of Logistic Regression and KNN Prediction Model for Heart Failure

Davina Febryanthi Kurniawan

Information Systems Department

Faculty of Engineering and Informatics

Multimedia Nusantara University, Tangerang, Indonesia

[davina.kurniawan@student.umn.ac.id](mailto:davina.kurniawan@student.umn.ac.id)

*Abstract*—One of the roles managed by the heart is human survival, which necessitates the heart's protection and awareness of its harm. The last stage of any coronary ailment is cardiovascular failure. The problem of heart failure-related fatalities necessitates the development of a survival prediction tool. The patient electronic clinical record apparatus may be used to assess symptoms, body features, and clinical research center test esteems that can be used to do biostatistical dissects aimed at highlighting instances and connections not recognized by clinical experts. AI is a solution for predicting tolerant endurance from the information generated and for recognizing the most important components among those recalled for their clinical records. There are several shortcomings of the currently available risk prediction models for Heart Failur. Most previous models are developed using traditional statistical approaches, one of them is regression modeling. With data mining techniques used in the available history data, namely the Heart Failure Clinical Records dataset of 299 instances on 13 features a logistic regression algorithm will be used to see the interrelationship between the dependent variable and the independent factors involved in predicting cardiovascular disease. It is hoped that this study will help readers have a better understanding of how to assess cardiovascular disease using logistic regression algorithms and the primary components that cause heart failure.

Index Terms—Heart Failure; Cardiovascular Disease, Logistic Regression, K-Nearest Neighbor

# Introduction

The continuity of human life can never be separated from the organs that work continuously. The heart is the main organ of the human body because of its very important task, namely pumping blood and distributing it throughout the body and later it is blood that carries oxygen and nutrients needed by the human body. According to the World Health Organization (WHO), 17.5 million people will die from cardiovascular diseases (CVDs) in 2019, accounting for 30% of all deaths globally. CVDs are the main cause of mortality worldwide, with more people dying each year from CVDs than from any other cause. Heart failure (HF) is becoming more well recognized as a clinical and public health concern linked to high mortality, morbidity, and health-care costs, particularly among people aged 65 and older, as the world population ages [1].

Heart disease is a disorder that occurs in the large blood vessel system, causing the heart and blood circulation to not function properly. Diseases related to the heart and blood vessels include: heart failure, coronary heart disease, and rheumatic heart disease [2]. Heart failure is the final stage of all heart disease and is the cause of increased morbidity and mortality in cardiac patients. The incidence of heart failure will increase in the future due to increasing life expectancy and the development of myocardial infarction treatment therapy resulting in improved life expectancy of patients with decreased heart function [3]. Heart failure is often the result of a number of problems affecting the heart at the same time. Coronary heart disease, high blood pressure, anemia, drinking too much alcohol, could lead to heart failure

Among modern methods for computer-aided detection, machine learning (ML) is an emerging technology for clinical data analysis and prediction generation in the context of early detection of diseases. Machine Learning specifically, can anticipate tolerant endurance from its information and can recognize the most significant components remembered for their clinical records [4]. In this study we conducted a classification to predict the survival of heart failure patients using several classification algorithms in order to compare which algorithm is more suitable for use in the heart failure dataset.

Chicco and Jurman [4] used 10 machine learning methods to predict the survival of heart failure patients with the attributes used, namely: from serum creatinine and ejection fraction, the highest accuracy is by using Random Forest which is 74%. Dwivedi [5] tested six ML techniques for heart disease prediction. They reported the highest accuracy (85%) with logistic regression on the Statlog dataset. From a similar perspective, Belavagi and Muniyal [6] used historical medical data to predict coronary heart disease with the South African Heart Disease dataset using three MLAs to discover correlations in the data to improve coronary heart disease prediction rate. The results showed that the Naive Bayes algorithm was promising for heart disease detection.

Therefore, the goals of this study is to examine a clinical data set on heart failure and determine the key factors that influence mortality. In order to examine which method is better suited for usage in the heart failure dataset, we ran a classification to predict the survival of heart failure patients using several classification algorithms. We analyze the performance of two machine learning algorithms (MLAs), such as Logistic Regression and K-Nearest Neighbor using age, anemia, creatinine, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, and time were all utilized with the class death event.

# Literature Review

## Regression

Regression is a method that is used to test two ideas. First, regression analyses are commonly employed for forecasting and prediction, and their use overlaps heavily with machine learning. Second, regression analysis may be used to discover causal relationships between independent and dependent variables in particular situations [7]. Importantly, regressions alone reveal only relationships between a dependent variable and a set of fixed factors in a dataset. According to the regression models, the independent variables predict the dependent variables. Regression analysis estimates dependent 'y' variable value due to the range of independent variable values 'x' [8].

## Logistic Regression

Logistic regression is a predictive model used to evaluate the relationship between the dependent variable (target) which is categorical data with nominal or ordinal scale and the independent variable (predictor) which is categorical data with interval or ratio scale. The dependent variable in logistic regression is represented as a binary variable with a value of 1 (yes) or 0 (no) (no). When the dependent variable is dichotomous, logistic regression is used to perform the necessary regression analysis (binary) [9].

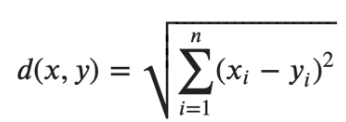


Logistic regression has several advantages and disadvantages. The benefits of logistic regression include the following. First, logistic regression can show a significant relationship between the dependent variable and the independent variable. Second, logistic regression analysis can also be used to compare the effect of variables measured at different scales including the effect of price changes and the number of promotional activities. This benefit helps market researchers or data analysts to eliminate and evaluate the best set of variables that will be used to build predictive models. Third, the logistic regression model is not only a classification model, but also provides information related to probability. To achieve a better result using Logistic Regression, first all independent variable must contain their valid value. Secondly, logistic regression works well for predicting categorical results and multinomial results. Third, there is no multicollinearity between variables in the dataset [10,11].

Logistic regression is a linear models. The coefficient ˇ enters the distribution of yn through a linear combination of xn. The difference is in the type of the response. In linear regression the response is real valued; in logistic regression the response is binary [12]. Linear and logistic regression are instances for a more general class of models, generalized linear models (GLMs) (McCullagh and Nelder, 1989). The idea is to use a general exponential family for the response distribution. In addition to real and binary responses, GLMs can handle categorical, positive real, positive integer, and ordinal responses. Logistic Regression is a special case of Generalized Linear Models. GLMs is a class of models, parametrized by a link function. If we choose logit link function, we will get Logistic Regression.

## K-Nearest Neighbor

K-Nearest Neighbor (KNN) algorithm is an approach used to classify data into existing data simply and efficiently. The concept of K-Nearest Neighbor algorithm is to find data with the closest distance between the data evaluated and the closest number of K (neighbors) to the training data [13]. The K-Nearest Neighbor algorithm has a goal, that is, new objects are classified based on attributes and training samples, then the results of the test are classified according to the majority of the categories in the KNN.



KNN is a non-parametric method which is used for classification and regression. Compared to other machine learning algorithm KNN is the simplest algorithm. This algorithm consists K-closet training examples in the feature space. In this algorithm K is a user defined constant. The test data are classified by assigning a constant value which is most chronic among the K-training samples nearest to the point. Literature shows the KNN has a strong consistency result [14].

# Methodology

Human cardiovascular system is examined in this study using some variables that affect its performance. As shown on Fig. 1, the process is started from retrieve data, analyze the correlation between variables, split data, prediction with Logistic Regression and K-Nearest Neighbor algorithm, and finished with data validation.

## Data Retrieval

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. In this study, the dataset used in this research is Heart Failure Clinical Records data taken from the Kaggle website (<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>)*.* The dataset was published in 2020 with 13 features, namely: age, anemia, creatinine, diabetes, ejection\_fraction, high\_blood\_pressure, platelets, serum\_creatinine, serum\_sodium, sex, smoking, time. In addition, the dataset contains the record of 299 patients, complete description of each attribute and the number of values for each attribute is shown in the Table I below:

|  |  |  |
| --- | --- | --- |
| No | Feature | Explanation |
| 1 | Age | Age of the patient (Years) |
| 2 | Anaemia | Decrease of red blood cells or hemoglobin (Boolean) |
| 3 | High blood pressure | If a patient has hypertension (Boolean) |
| 4 | Creatinine  phosphokinase | Level of the CPK enzyme in the blood (mcg/L) |
| 5 | Diabetes | If the patient has diabetes (Boolean) |
| 6 | Ejection fraction | Percentage of blood leaving the heart at each contraction (Percentage) |
| 7 | Sex | Gender: female or male (Binary) |
| 8 | Platelets | platelets in the blood (kiloplatelets/mL) |
| 9 | Serum creatinine | Level of creatinine in the blood (mg/dL) |
| 10 | Serum sodium | Level of sodium in the blood (mEq/L) |
| 11 | Smoking | If the patient smokes (Boolean) |
| 12 | Time | Follow-up period (Days) |
| 13 | Target (death event) | If the patient died during the follow-up period (Boolean) |

## The Correlation between Variables Analysis

Furthermore, to facilitate data analysis, all variables in the imported dataset will be represented in the form of a correlation plot to f data reading in general. Analyze the Correlation between Variables as part of the process; the correlation between variables is investigated to see whether the Logistic Regression or K-Nearest Neighbor model is the best model. A matrix will be displayed to show the relationships between variables in the provided dataset. This is also done to see whether there is any multicollinearity in the dataset's variables.

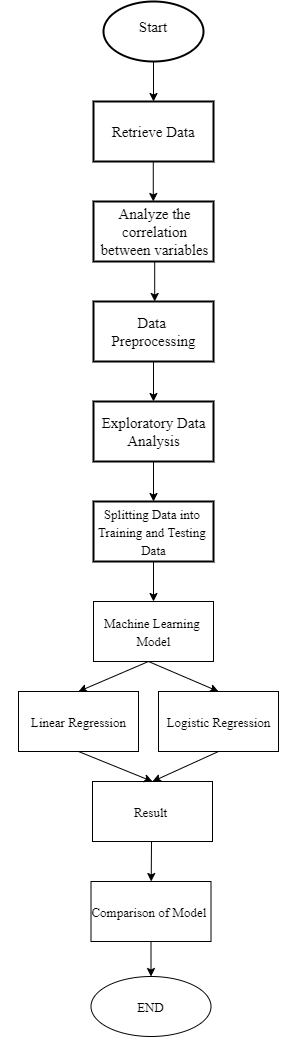


Figure 1 Framework

## Data Preprocessing

Data preprocessing is an essential step use to clean the data and make it useful for any experiment associated with machine learning or data mining [24]. Machine learning applied to medical records can be useful to predict the survival of a patient, highlighting patterns and even ranking the features to understand which are risk factors, possibly undetectable by doctors. In this study the analysis will be done starting from an EDA to understand the dataset and applying some preprocessing to be able to learn properly from it. Then will follow a number of machine learning models trained on the preprocessed dataset, aiming to predict the survival of patients that suffered heart failure.

The column DEATH\_EVENT is our dependent variable that we are going to predict, and it is consisted of binary values (0= alive, 1=death). Binary values of dependent variable oindicate that we can use a logistic regression to build a model. Binary variables consisting of anaemia, diabetes, high blood pressure, sex, smoking and will be converted into factor.

## Data Preparation

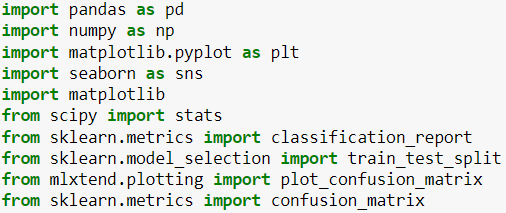
The dataset imported in Rstudio will be divided into two section, namely training data and testing data. Training data is used as a basis for building models. Meanwhile, testing data is used as a basis for testing or validating the model. In this data preparation process, 299 data will be sampled. Then the data will be partitioned into train data and test data.

## Prediction with Machine Learning Model

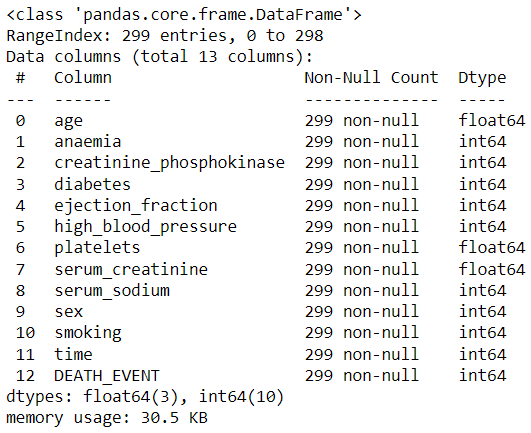
The data that has been partitioned in the previous process will be used by the data analysis. Prediction using two model, namely Logistic Regression and K-Nearest Neighbor model will produce several data or result that can be useful as a basis for concluding to make predictions and examine which model is better and has higher accuracy by comparing the two models.

# Result and Discussion

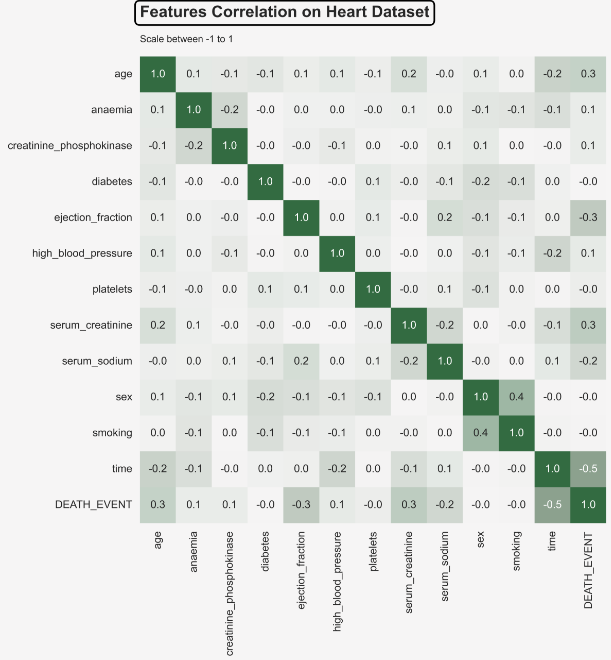
## Identification Phase



Picture 4.1 Load *library package*

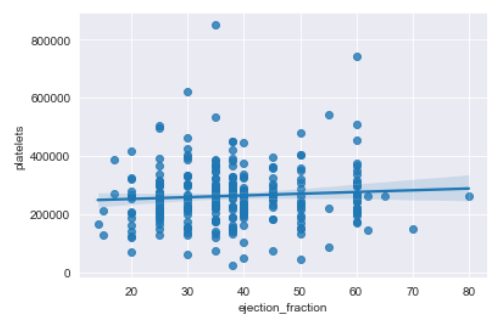


Picture 4.2 Data Structure and Summary



Picture 4.3 Correlation Plot

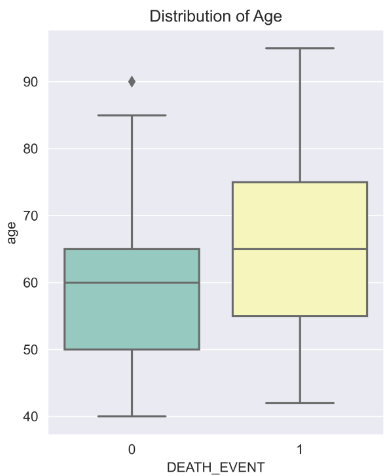
The possible range of values for the correlation coefficient is -1.0 to 1.0. In other words, the values cannot exceed 1.0 or be less than -1.0. A correlation of -1.0 indicates a perfect negative correlation, and a correlation of 1.0 indicates a perfect positive correlation. If the correlation coefficient is greater than zero, it is a positive relationship. Conversely, if the value is less than zero, it is a negative relationship. A value of zero indicates that there is no relationship between the two variables. From the correlation matrix, we can see Death Event is highly correlated with serum creatinine, age, serum sodium, ejection fraction.



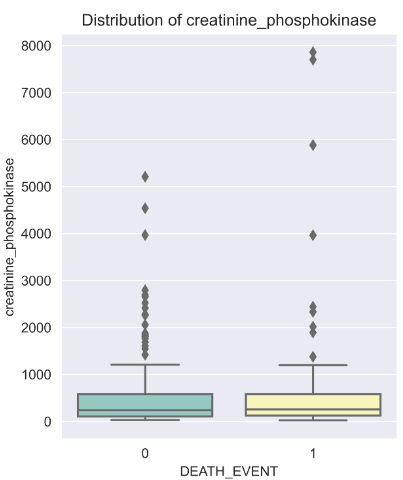
Picture 4.4 Correlation between platelets and ejection\_fraction

Two main functions in seaborn are used to visualize a linear relationship as determined through regression. The analysis use regplot() to explore the correlation between platelets and ejection fraction. The regplot consist two symbol, the dots (scatter plot) and linear regression model fit.

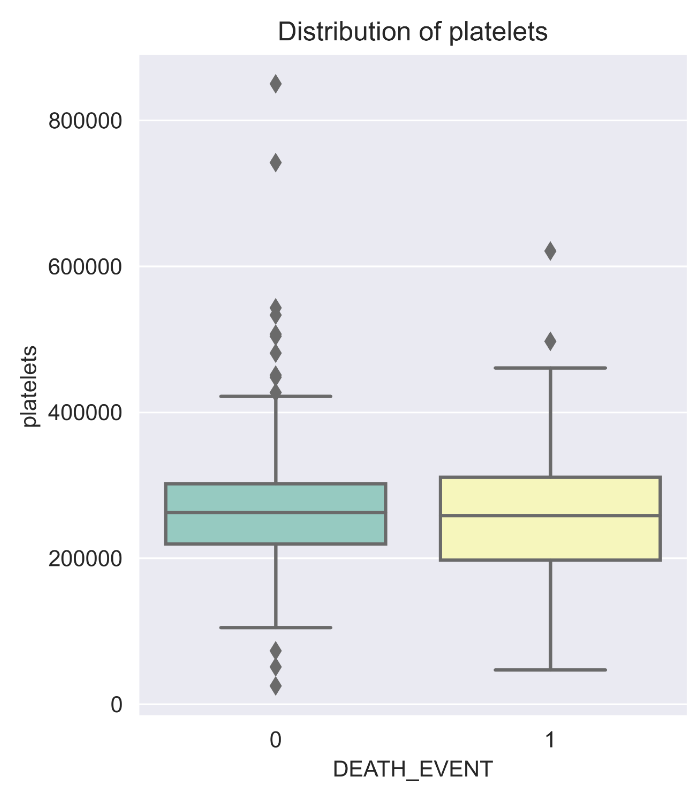
Picture 4.5 data preprocessing



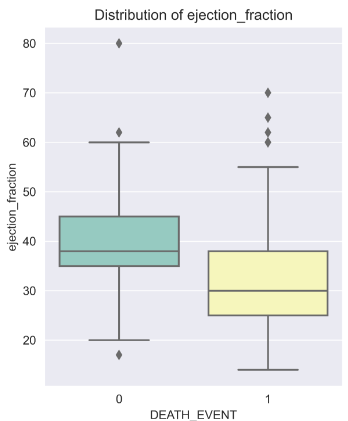
Picture 4.6 Boxplot of age with death event status

****

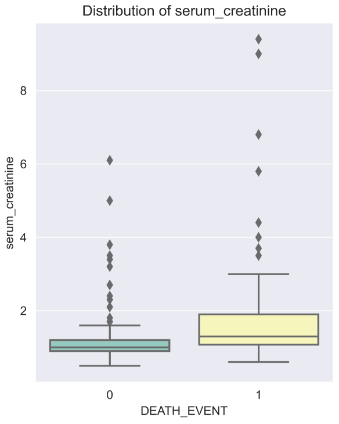
Picture 4.7 Boxplot of CP with death event status

****

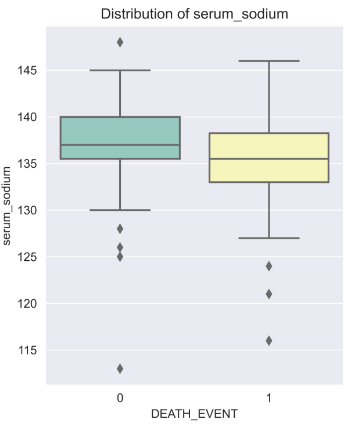
Picture 4.8 Boxplot of platelets with death event status

****

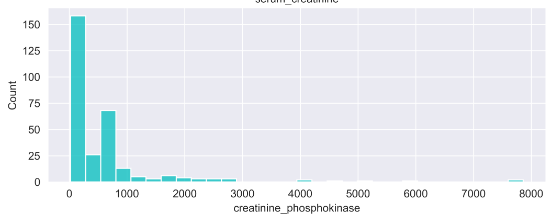
Picture 4.9 Boxplot of EF with death event status

****

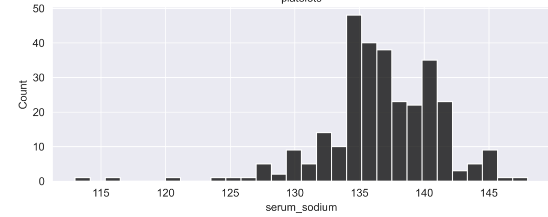
Picture 4.10 Boxplot of SC with death event status

****

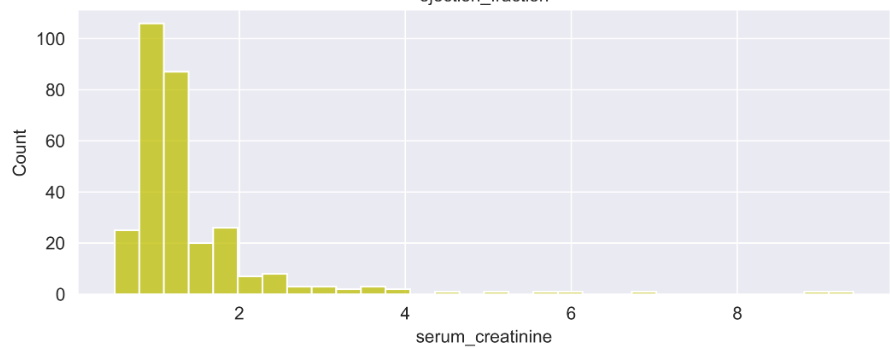
Picture 4.11 Boxplot of SS with death event status

****

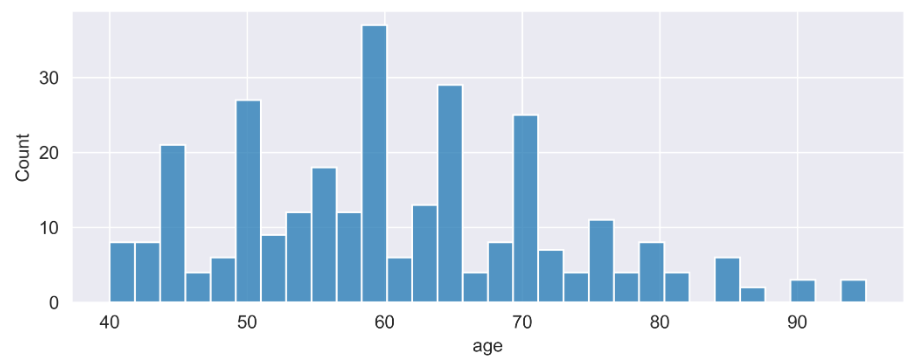
Picture 4.12 Histogram of Creatinine P.

****

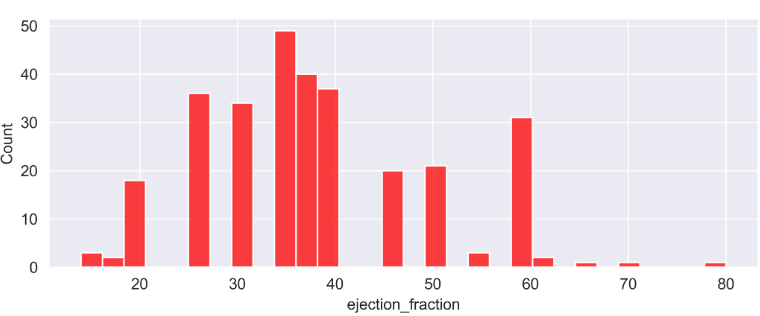
Picture 4.13 Histogram of S. Sodum

****

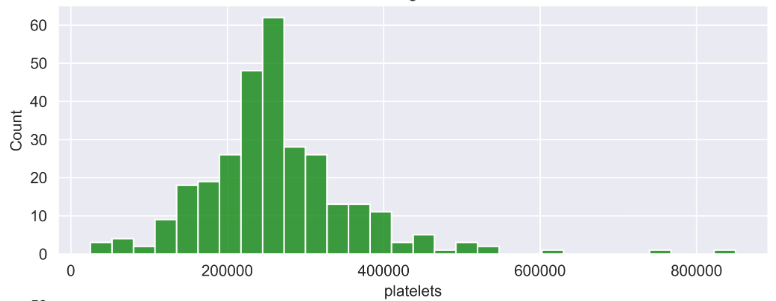
Picture 4.14 Histogram of S. Creatinine

****

Picture 4.15 Histogram of Age Distribution

****

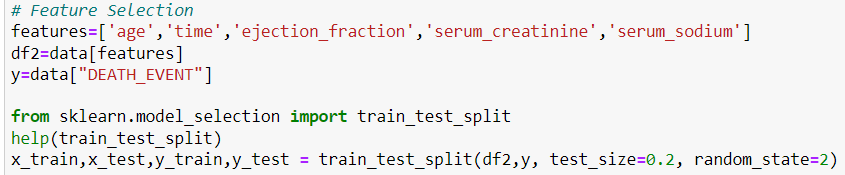
Picture 4.16 Histogram of EF Distribution

****

Picture 4.17 Histogram of Platelets Distribution

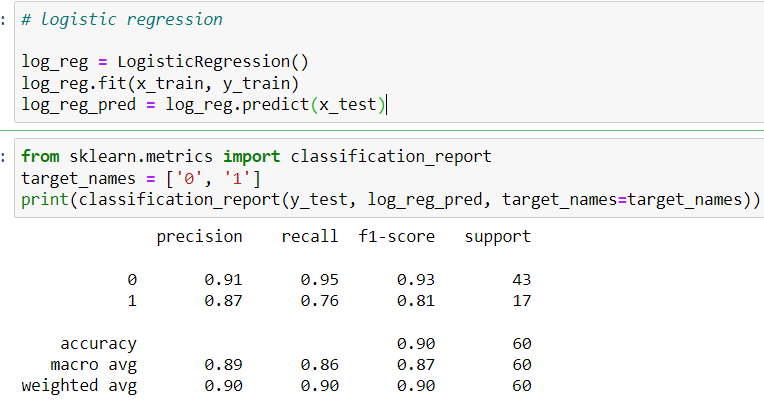
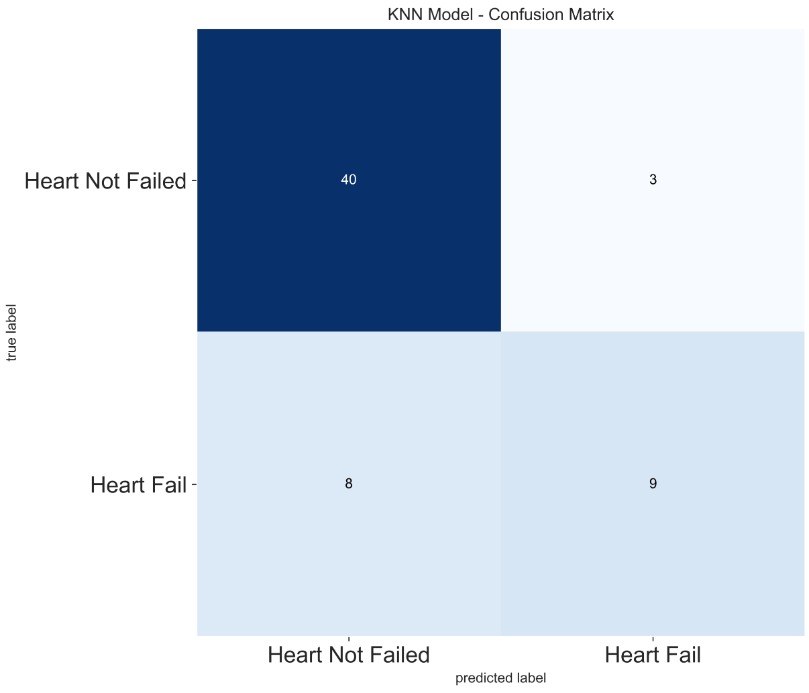
## Interpretation and Testing Phase

The data will be split in the interpretation and testing section. Then the data will be divided into two parts, training set and testing set. The training dataset will be stored in an object named "training" and the testing dataset will be stored in an object called "testing".

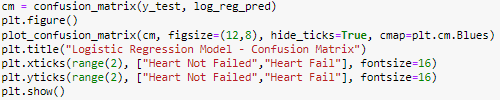


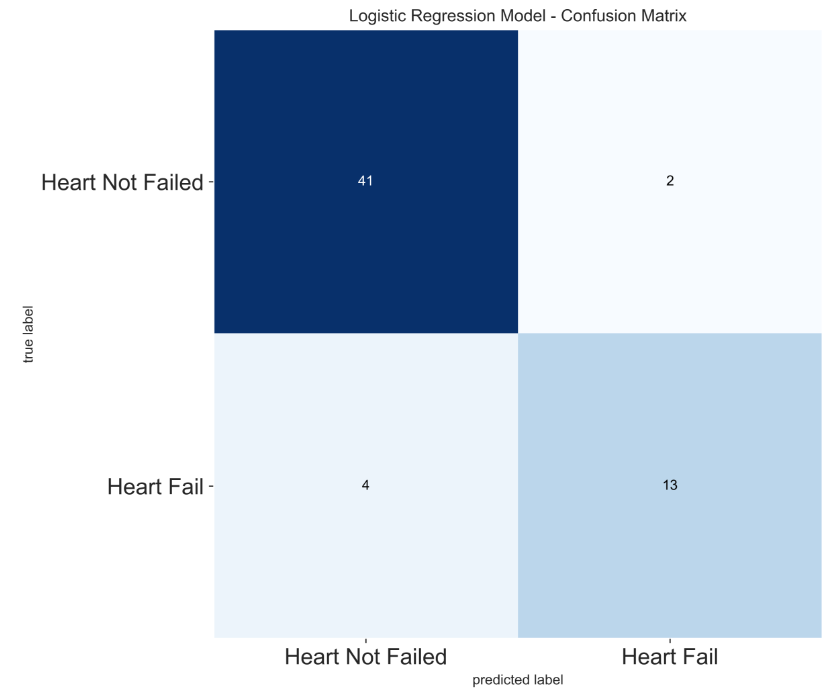
Picture 4.17 Splitng data

* Logistic Regression



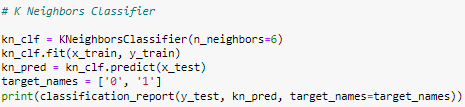
Picture 4.18 Logistic Regression Accuracy

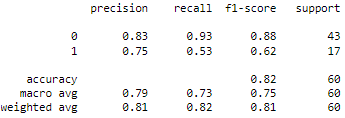




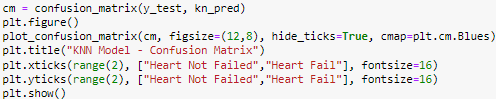
Picture 4.19 Logistic Regression Confussion Matrix

* K-Nearest-Neighbor





Picture 4.20 K-Nearest Neighbor Accuracy



Picture 4.21 K-Nearest-Neighbor Accuracy

A logistic regression model was built using thirteen attributes that affect cardiovascular function as variables in this study. There is no evidence of a significant association between variables among the variables. As a result, the likelihood of multicollinearity in this study is lower. As a solution to the problem, this study employs a logistic regression approach. The logistic regression algorithm was classed as an effective and efficient algorithm in predicting the primary causes causing cardiovascular disease, as the problem presented in this study, employing the algorithm. Picture 4.18 and 4.19 depicts a confusion matrix. The logistic regression approach may be claimed to be successful in predicting parameters that have a major impact on cardiovascular function, with an accuracy of 90%. It is possible to calculate the probability of a person's potential for cardiovascular disease, especially when utilizing particular predicted numbers.

# Conclusion

The heart is one of the most essential organs in the body, and it is never separated from the continuity of the body organs that operate well together. Cardiac failure is one of the most fatal heart disorders and the ultimate stage of heart disease. Data mining techniques can aid in forecasting the survival of heart failure patients. The suggested strategy in this work is to apply a generalized linear model and multiple linear regression using preprocessing techniques on the Failure Clinical Records dataset, which has 12 features and one class, and has been shown to have the maximum accuracy when compared to logistic regression. It was found that the use of the logistic regression algorithm is effective and efficient in predicting cardiovascular disease where based on the results of data validation it is found that the accuracy of the prediction results with the algorithm reaches 90%. It proves that this algorithm is suitable for use as a prediction algorithm in this study. Heart failure is substantially connected with two variables according to the results of heart failure predictions. An increase in the value of these factors will have an effect on total cardiovascular performance, causing cardiovascular function to decline while the risk of cardiovascular disease to rise. The logistic regression technique is effective in predicting the primary causes of cardiovascular disease.

In this study, the fact that the traditional biostatistics analysis selected ejection fraction and serum creatinine as the two most relevant features confirmed the relevance of the feature ranking executed with machine learning. Moreover, our approach showed that machine learning can be used effectively for binary classification of electronic health records of patients with cardiovascular hearth diseases

##### References

1. 1. Mozaffarian D, Benjamin EJ, Go AS, et al; Writing Group Members; American Heart Association Statistics Committee; Stroke Statistics Subcommittee. Executive summary: heart disease and stroke statistics—2016 update: a report from the American Heart Association. Circulation. 2016;133(4):447-454.
2. Aeni, W. N., Santosa, S., & Supriyanto, C. (2014). Algoritma Klasifikasi data mining naïve bayes berbasis Particle Swarm Optimization untuk deteksi penyakit jantung. Jurnal Pseudocode, 1(1), 11–14.I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
3. Rozie, F., Hadary, F., & Wigyarianto, F. T. P. (2016). Rancang Bangun Alat Monitoring Jumlah Denyut Nadi / Jantung Berbasis Android. Jurnal Teknik Elektro Universitas Tanjungpura, 1(1), 1–10.
4. Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC medical informatics and decision making, 20(1), 1-16.
5. Dwivedi, A.K. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput. Appl. 2019, 29, 685–693.
6. Belavagi, M.C.; Muniyal, B. Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection. Procedia Comput. Sci. 2016, 89, 117–123.
7. J. Wu, C. Liu, W. Cui, and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," in 2019 IEEE International Conference on Power Data Science (ICPDS), 2019, pp. 139-142.
8. J. Wu, C. Liu, W. Cui, and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," in 2019 IEEE International Conference on Power Data Science (ICPDS), 2019, pp. 139-142.
9. M. S. Acharya, A. Armaan, and A. S. Antony, "A comparison of regression models for prediction of graduate admissions," in 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-5.
10. Schober, Patrick MD, PhD, MMedStat\*; Vetter, Thomas R. MD, MPH† Logistic Regression in Medical Research, Anesthesia & Analgesia: February 2021 - Volume 132 - Issue 2 - p 365-366 doi: 10.1213/ANE.
11. A. L’Heureux, K. Grolinger, H. El Yamany, and M. Capretz, ―Machine Learning with Big Data: Challenges and Approaches,‖ IEEE Access, 2017.
12. Iqbal, M. A. Article 4 Application of Regression Techniques with their Advantages and Disadvantages.
13. D. A. Anggoro and P. I. Rahmatullah, The implementation of subspace outlier detection in k-nearest neighbors to improve accuracy in bank marketing data. International Journal of Emerging Trends in Engineering Research, vol. 8, pp. 545–550, February2020.
14. Beant Kaur h, Williamjeet Singh, “Review on Heart Disease Prediction System using Data Mining Techniques”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, pp.3003-08,October 2014.