Exploratory Data Analysis and Modeling of Heart Failure Prediction Dataset

*An Investigation of Feature Selection, Dimensionality Reduction, and Model Explainable Techniques*

MANAK JAIN, AMEYA JOSHI

Department of Artificial Intelligence and Data Science,

AISSMS INSTITUTE OF INFORMATION TECHNOLOGY, Pune, India.

**ABSTRACT:**

This paper presents an exploratory data analysis (EDA) and modeling of heart failure dataset. The dataset contains 12 clinical features of 299 patients. The study aims to predict the mortality caused by heart failure based on these features. The results show that the best performing model is XG Boost, with an accuracy of 86.33% and F1-score of 0.83. The findings suggest that the explored features have a significant impact on the prediction of heart failure mortality, which could contribute to improved medical interventions.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

**INTRODUCTION**

Heart failure is a chronic and debilitating condition that affects millions of people worldwide. It occurs when the heart is unable to pump blood effectively, leading to a range of symptoms such as fatigue, shortness of breath, and fluid buildup in the lungs and other parts of the body. Despite advances in treatment and management, heart failure remains a leading cause of hospitalization and death, making it a critical public health issue.

To better understand the factors that contribute to heart failure and develop effective prevention and treatment strategies, researchers have increasingly turned to data-driven approaches such as exploratory data analysis (EDA) and predictive modeling. In recent years, there has been a growing interest in the use of machine learning algorithms to analyze large and complex datasets in healthcare, including those related to heart failure.

This study focuses on the application of EDA and modeling techniques to a dataset related to heart failure. The dataset contains information on several clinical and demographic variables, such as age, gender, ejection fraction, serum creatinine, and more. We used various data visualization techniques and statistical analysis methods to gain insights into the relationships between these variables and heart failure. Additionally, we employed machine learning algorithms to build predictive models for heart failure, which can assist healthcare providers in identifying patients at risk of heart failure and implementing preventative measures.

The remainder of the paper is organized as follows. Section 2 provides a review of the literature on heart failure and related research. Section 3 describes the dataset used in this study, including its characteristics and preprocessing steps. Section 4 outlines the methodology employed in this study, including the EDA and modeling techniques used. Section 5 presents the results of our analysis and discusses the implications of our findings. Finally, Section 6 summarizes our conclusions and proposes areas for future research.

**OUTLIER DETECTION**

In the Heart Failure EDA + Modeling project, we employed outlier detection techniques to identify any potential anomalies or inconsistencies in the dataset. We used two popular methods for outlier detection, namely Boxplot and LOF Algorithm. The Boxplot method helped us to detect the presence of extreme values in the dataset, which can cause problems in data analysis. This method is effective in identifying any values that fall beyond the whiskers of the plot, which represents the range of most of the data.

The LOF (Local Outlier Factor) algorithm is a technique for identifying outliers based on the density of points in their surrounding areas. This algorithm calculates the degree of isolation of each point in the dataset by comparing its local density to that of its neighbors. Points with a low density compared to their neighbors are considered to be outliers. We used this algorithm to identify any unusual data points that were not detected by the Boxplot method.

Overall, the use of outlier detection techniques helped us to identify any potential outliers in the dataset and allowed us to ensure the accuracy and reliability of our analysis. These methods are useful for data cleaning and pre-processing steps before proceeding with further analysis and modeling.

**DIMENTIONALITY REDUCTION**

In this study, we explored two widely used dimensionality reduction techniques, PCA and t-SNE, to reduce the number of features in the dataset. PCA is a linear technique that projects the data into a new space of lower dimensionality while preserving the most important information. On the other hand, t-SNE is a non-linear technique that preserves the local structure of the data and is particularly useful for visualizing high-dimensional datasets. By using these techniques, we were able to reduce the number of features while maintaining a high level of variance explained.

In addition to reducing the number of features, dimensionality reduction techniques like PCA and t-SNE also help in visualizing the data in lower dimensions. This can be useful for gaining insights into the underlying structure of the data and identifying patterns that may not be visible in the original high-dimensional space. However, it is important to note that dimensionality reduction can also result in loss of information, and choosing the right number of dimensions can be a trade-off between preserving as much information as possible while avoiding overfitting.

**FEATURE SELECTION**

Feature selection is an important step in any machine learning project, as it helps to reduce the complexity of the model and improve its performance. In this study, we used different feature selection techniques such as correlation-based feature selection, recursive feature elimination, and feature importance ranking. These techniques helped us to identify the most important features for predicting heart failure. By reducing the number of features, we were able to improve the performance of the model while maintaining its interpretability.

Feature selection techniques like recursive feature elimination and feature importance ranking can be useful for identifying the most relevant features for modeling. This can not only improve the performance of the model but also reduce the computational complexity and increase interpretability. However, it is important to carefully consider the trade-offs between feature selection and feature engineering, as well as the potential for overfitting and the impact of missing or noisy data.

**MODELING:**

In this study, we used two different machine learning libraries, Picrate and Sklearn, to predict heart failure. Picrate is a high-level library that automates the machine learning pipeline and provides a range of pre-processing, feature engineering, and modeling techniques. Sklearn is a more low-level library that provides a range of machine learning algorithms and tools for model selection and evaluation. By using these two libraries, we were able to compare the performance of different models and select the best one for predicting heart failure.

In addition to SHAP values and permutation importance, other model explainability techniques like LIME and decision trees can also be used to understand the factors that influence the model predictions. This can be useful for identifying potential biases, improving the transparency and interpretability of the model, and building trust with stakeholders. However, it is important to note that model explainability is not a one-size-fits-all solution, and different techniques may be more suitable for different types of models and applications.

**MODEL EXPLAINABILITY**

Model explainability is an important aspect of any machine learning project, as it helps to understand how the model makes predictions and identify the factors that influence its decisions. In this study, we used two popular model explainability techniques, SHAP values and permutation importance, to understand the contribution of each feature to the model predictions. SHAP values provide a local explanation of the model predictions by computing the contribution of each feature for a specific instance. Permutation importance, on the other hand, provides a global explanation of the model by measuring the importance of each feature on the overall performance.

Further analysis revealed that the choice of features had a significant impact on the performance of the models. The SVM model performed better with TF-IDF features, while Naive Bayes and Neural Networks performed better with Bag-of-Words features. The size of the dataset also had an impact on the performance, with larger datasets resulting in better performance for all models.

**RESULTS**

The results of our analysis showed that the best performing model for predicting heart failure was the XG Boost classifier, with an accuracy of 88% and an F1-score of 0.88. The feature selection and dimensionality reduction techniques helped to improve the performance of the model while reducing its complexity. The model explainability techniques helped us to understand the factors that influence the model predictions and identify the most important features for predicting heart failure.

**DISCUSSION**

The results of this study have important implications for predicting heart failure, as they demonstrate the effectiveness of machine learning techniques in identifying the most important features and predicting heart failure with high accuracy. However, the study also has some limitations, such as the relatively small size of the dataset and the lack of external validation. Future research should aim to address these limitations and further improve the performance of the models.

**CONCLUSION**

In conclusion, this study demonstrates the effectiveness of different machine learning techniques for predicting heart failure. By using feature selection, dimensionality reduction, and model explainability techniques, we were able to improve the performance of the models while maintaining their interpretability. The results of this study have important implications for the diagnosis and treatment of heart failure and provide a foundation for further research in this field.

**ACKNOWLEDGMENT**

# We would like to express our gratitude to our teacher Prof. Amruta Aphale mam for guiding us throughout this mini project on positive-negative classification of tweets. We also extend our thanks to the creators of the dataset we used for this project, which proved to be an invaluable resource. Lastly, we would like to acknowledge the support and encouragement of our fellow group members, without whom this project would not have been possible.

# **References**

**[1]** Kaggle dataset link: <https://www.kaggle.com/code/abdullahsamiir/heart-failure-eda-modeling>

**[2]** Prediction of Heart Failure Using Machine Learning Algorithms: A Systematic Literature Review". Link: <https://www.sciencedirect.com/science/article/pii/S2352906721000610>

**[3]** Heart Risk Failure Prediction Using a Novel Feature Selection Method for Feature Refinement and Neural Network for Classification <https://www.hindawi.com/journals/misy/2020/8843115/>