**EMERALD INTERNATIONAL COLLEGE**


**School Of Computing and Analytics**

**Data Science Program Unit**

**Business Intelligence Course**

**Instructur:- Dr.Genet Shanko**

# Machine Learning Approach for Heart Failure Prediction

**BY:**

| Name | ID |
|---|---|
| 1.Tekele Worku | DS0077/15 |
| 2. Shimelis Tesfaye | DS0073/15 |
| 3. Asnake Dagne | DS0055/15 |
| 4. Bekalu Metalign | DS0096/16 |
| 5. Habtamu Hailu | DS0023/15 |

**November, 2024**

## Abstract

Heart failure represents a significant global health challenge, characterized by high mortality rates and considerable healthcare costs. Traditional statistical models often struggle to detect subtle patterns in patient data, prompting the need for advanced predictive methods. This project develops a machine learning model to predict heart failure with high accuracy using historical clinical data. We evaluated five popular machine learning techniques Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, and K-Nearest Neighbors utilizing a publicly available dataset from Kaggle.com. These models were chosen due to their good performance in medical applications and the pressing need to improve heart failure management and survival rates. The performance of each model was measured using accuracy, precision, recall, F1-score, and an overall average score, based on 12 predictor features of heart failure. Logistic Regression and Random Forest demonstrated the highest accuracy at 87%, with precision scores of 83% and 85%, respectively. SVM achieved slightly lower accuracy (85%) but maintained competitive precision. Overall average scores corroborated these findings, with LR and RF scoring above 81. These results underscore the effectiveness of LR and RF in predicting heart failure, suggesting their potential for clinical application. Future work may focus on model enhancements and the incorporation of additional features to further improve predictive performance.

***Key words:*** *Heart Failure, Machine Learning, Predictive Modeling and Performance Evaluation*

## 1. Introduction

Heart failure (HF) is a significant global health issue, leading to high mortality rates and healthcare costs. It is a serious problem which has a huge impact on people's life. With the accelerated pace of life, increased portion sizes and inactivity, most people always ignore their health. Moreover, because of the environmental deterioration, those factors can lead to the issue of heart failure which can become more and more common in the future. If people did not pay attention to the issue of heart failure, it would finally cause the death (Wang, J. ,2021). Despite advancements in medical technology, accurately predicting heart failure events remains a challenge. Current predictive models often lack precision and fail to account for the multifactorial nature of HF. This project addresses the need for a robust machine learning-based system that can analyze diverse patient data to identify individuals at high risk of developing heart failure. By utilizing ML algorithms, we aim to enhance prediction capabilities, ultimately improving patient outcomes and reducing healthcare costs. Despite advancements in medical technology, accurately predicting heart failure events remains a challenge. Current predictive models often lack precision and fail to account for the multifactorial nature of HF. This this project addresses the need for a robust machine learning-based system that can analyze diverse patient data to identify individuals at high risk of developing heart failure.

### 2.Related Work

Traditional methods for predicting heart failure often rely on clinical assessments and static risk factors, such as age, gender, and medical history. However, these approaches may not capture the dynamic nature of patient health and may lack precision in identifying individuals at high risk of developing heart failure.Machine learning (ML) has emerged as a promising approach for predicting heart failure events. By leveraging large datasets and advanced algorithms, ML models can uncover complex patterns and relationships in patient data, leading to more accurate predictions. Several studies have explored the application of ML in heart failure prediction, demonstrating its potential to enhance clinical decision-making and patient management (Raghupathi & Raghupathi, 2014) conducted a review of big data analytics in healthcare, highlighting the potential of ML in predicting and managing chronic diseases like heart failure. The authors emphasized the importance of integrating ML into clinical workflows to enhance patient care. For instance, (Chicco & Jurman, 2020) demonstrated that models like Random Forest

and XGBoost achieved over 80% accuracy. Similarly (Alizadehsani *et al.* 2019) highlighted the importance of feature selection in enhancing model performance. This project demonstrates the growing interest and potential of machine learning in heart failure prediction.

**General objective**

- ✓ The primary objective of this project is to develop a machine learning-based predictive model for heart failure using Kaggle data.

**Specific objective**

- ✓ Collect and preprocess relevant patient data from existing datasets.
- ✓ Implement and evaluate multiple Machine learning algorithms, including Logistic Regression, Random Forest, supervector machine, Descension Tree and K-Nearest Neighbors.
- ✓ Compare the performance of these algorithms to identify the most effective approach for heart failure prediction**.**

## 2.1 Data set and Features

The dataset is obtained from the Kaggle heart disease dataset consisting of 299 patients (https://www.kaggle.com). There are 13 variables in this dataset: age, anemia creatinine_phosphokinase, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, time, Death event (figure.1).

**Figure1**: **Example of Dataset for Heart Disease Prediction**

| age | anaemia | creatinine_phosphokinase | diabetes | ejection_fraction | high_blood_pressure | platelets | serum_creatinine | serum_sodium | sex | smoking | time | DEATH_EVENT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 75.0 | 0 | 582 | 0 | 20 | 1 | 265000.00 | 1.9 | 130 | 1 | 0 | 4 | 1 |
| 55.0 | 0 | 7861 | 0 | 38 | 0 | 263358.03 | 1.1 | 136 | 1 | 0 | 6 | 1 |
| 65.0 | 0 | 146 | 0 | 20 | 0 | 162000.00 | 1.3 | 129 | 1 | 1 | 7 | 1 |
| 50.0 | 1 | 111 | 0 | 20 | 0 | 210000.00 | 1.9 | 137 | 1 | 0 | 7 | 1 |
| 65.0 | 1 | 160 | 1 | 20 | 0 | 327000.00 | 2.7 | 116 | 0 | 0 | 8 | 1 |
| 90.0 | 1 | 47 | 0 | 40 | 1 | 204000.00 | 2.1 | 132 | 1 | 1 | 8 | 1 |
| 75.0 | 1 | 246 | 0 | 15 | 0 | 127000.00 | 1.2 | 137 | 1 | 0 | 10 | 1 |
| 60.0 | 1 | 315 | 1 | 60 | 0 | 454000.00 | 1.1 | 131 | 1 | 1 | 10 | 1 |

### 2.2. Machine Learning Techniques

In this project to predict heart failure, we employed various machine learning techniques those were:

**Logistic Regression**: This model is useful for binary classification problems. It estimates the probability of a binary outcome based on one or more predictor variables. Its simplicity allows for easy interpretation of the coefficients, providing insights into how each feature influences the likelihood of heart failure.

**Support Vector Machine (SVM)**: SVM is effective for both linear and non-linear classification tasks. By finding the optimal hyperplane that separates different classes in the feature space, it can handle high-dimensional data well. Kernel functions can be applied to manage complex relationships between variables.

**Decision Tree**: This model builds a flowchart-like structure where each node represents a decision based on the value of a feature. It's intuitive and easy to visualize, making it suitable for understanding decision paths. However, it can be prone to overfitting if not properly managed.

**Random Forest**: An ensemble method that combines multiple decision trees to improve accuracy and control overfitting. Each tree is trained on a random subset of the data and features, and their predictions are aggregated. This approach enhances robustness and performance, particularly with large datasets.

**K-Nearest Neighbors (KNN)**: A non-parametric method that classifies data points based on the majority class among the k-nearest neighbors. It's simple and effective for small datasets, but its performance can diminish with larger datasets and higher dimensionality due to the "curse of dimensionality."

## 3. Implementation

In this Project , the dataset was first prepared by splitting it into features (X) and the target variable (y), where the target variable is DEATH_EVENT and the features encompass all other columns in the DataFrame.

```
X=df.drop('DEATH_EVENT',axis=1)
```

Y=df['DEATH_EVENT']

In the analysis of heart failure outcomes, the distribution of the target variable, DEATH_EVENT, revealed that 32.1% of patients experienced a death event, while 67.9% did not. This distribution underscores a notable proportion of cases associated with mortality, highlighting the importance of predictive modeling in this context.

Subsequently, the data was divided into training and testing sets, allocating 30% of the data for testing purposes. To facilitate the implementation of various machine learning models, a dictionary was created to include logistic regression, support vector machine (SVM), decision tree, random forest, and K-nearest neighbors (KNN). Each model was then fitted to the training data, and predictions were made on the test set. The accuracy of each model, along with a classification report that includes precision, recall, and F1-score, was calculated and stored in a results dictionary. Finally, the accuracy and classification report for each model were printed to provide a comprehensive evaluation of their performance.

## 4.Result and discussion

The evaluation of the machine learning models—Logistic Regression (LR), Random Forest (Rf), Support Vector Machine (SVM) yielded insightful results regarding their performance in predicting heart failure outcomes. Both Logistic Regression achieved the highest accuracy, scoring 87%, indicating strong predictive capabilities in correctly identifying both death events and non-death events. The Random Forest model closely followed with an accuracy of 86%, while the SVM model exhibited lower accuracy at 82%.

In terms of precision, Random Forest and SVM demonstrated strong precision with scores of 82%, while Logistic Regression achieved a precision of 83%.

When assessing recall, Logistic Regression led with a score of 77%, indicating its effectiveness in capturing actual positive cases. Random Forest followed with a recall of 74%, while SVM 71%. The F1-scores, which provide a balance between precision and recall, revealed that Logistic Regression had the highest score at 80%, closely followed by Random Forest at 79%. SVM scored 77%,.

Overall, the average scores across all metrics indicated that Logistic Regression was the most effective model, with an average score of 81.75, followed by Random Forest at 81.25 and SVM at 79.5. These findings emphasize the superiority of Logistic Regression and Random Forest in accurately predicting heart failure outcomes.

**Table 2: Result of Performance score of LR, RF, SVM.**

| **Model evaluation metrics** | **Logistic Regression** | **Random forest** | **Support Vector Machine** |
|---|---|---|---|
| Accuracy | 87% | 86% | 82% |
| Precision | 83% | 85% | 85% |
| Recall | 77% | 74% | 71% |
| F1-score | 80% | 79% | 77% |
| **Average** | **81.75%** | **81.25%** | **79.5%** |

## 5. Conclusion

In conclusion, this project successfully demonstrated the potential of machine learning techniques in predicting heart failure outcomes using a Kaggle dataset. By implementing and evaluating multiple algorithms, including Logistic Regression and Random Forest, the study identified that both models achieved the highest accuracy of 87%.. Overall, the findings emphasize the importance of selecting robust predictive models to enhance clinical decision-making, ultimately improving patient outcomes and reducing healthcare costs associated with heart failure management. This research underscores the need for further exploration of machine learning applications in healthcare to address complex health issues effectively.

## 6.References

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*.

Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*.

Dimitris, B., Paliouras, G., & Dimitris, A. (2018). Deep learning for time-series based heart failure prediction. *Journal of Biomedical Informatics*.

Wang, J. (2021, September). Heart failure prediction with machine learning: a comparative study. In *Journal of Physics: Conference Series* (Vol. 2031, No. 1, p. 012068). IOP Publishing.