Project Overview

Project Title: Improving Heart Failure Prediction Accuracy with PCA and Information Gain in Machine Learning Models

**Short Summary of the Project Topic:** Heart failure is one of the severe cardiovascular diseases that impose disability and mortality in the global population. The use of effective and efficient features for early detection is key to enhancing the patient’s prognosis and treatment effectiveness, while minimizing health care costs; however, identifying the most significant features and optimal model type remains unresolved (Ishaq et al, 2021). Research shows that datasets with many features containing noise or less important features lead to inaccurate models that don’t give reliable predictions (Reddy et al, 2021). Based on the procedure used in Kakoly et al, (2023), who employed Principal Component Analysis (PCA) and Information Gain (IG) for diabetes risk prediction, this study adapts the same to predict heart failure risk. PCA and IG were used in previous studies to improve classification performance and feature importance, making them applicable to the current study. This research analyses the performance of algorithms, such as Logistic Regression, Naïve Bayes, Random Forest, K-NN, and Decision Tree, with or without applying PCA and IG feature selection. The study compares the performances of the selected models to determine the influence of feature selection on accuracy. Furthermore, it investigates whether it is possible to forecast the likelihood of a patient having heart failure through machine learning for both medical and non-medical dataset attributes. PCA and IG can enhance model performance as PCA reduces dimensionality, while IG selects the most salient features.

Research Questions

1. Which feature selection method (Principal Component Analysis (PCA) or Information Gain (IG)) enhances the effectiveness of heart failure prediction models?
2. Which combination of machine learning algorithms and feature selection techniques (PCA and IG) leads to a more reliable and early prediction of heart failure risk?

Project Objectives

* To collect and pre-process heart failure datasets from sources from Kaggle.
* To apply PCA and Information Gain for feature selection on pre-processed dataset.
* To design machine learning algorithms such as Logistic Regression, Naive Bayes, Random Forest, K-Nearest Neighbor, and Decision Tree for heart failure prediction and compare its performance with and without feature selection.
* To determine the impact of PCA and IG on model performance in heart failure prediction, focusing on accuracy, precision, recall, F1-score, and AUC-ROC.
* To identify the optimal combination of feature selection method and machine learning model for heart failure prediction accuracy and reliability.

Reference List

Kakoly, I.J., Hoque, M.R. and Hasan, N., 2023. Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique. Sustainability, 15(6), p.4930.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V. and Nappi, M., 2021. Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques. *IEEE access*, *9*, pp.39707-39716.

Reddy, K.V.V., Elamvazuthi, I., Abd Aziz, A., Paramasivam, S. and Chua, H.N., 2021, July. Heart disease risk prediction using machine learning with principal component analysis. In *2020 8th international conference on intelligent and advanced systems (ICIAS)* (pp. 1-6). IEEE.

Project Plan

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Details | Start Date | End Date |
| Project Planning | Define objectives, select dataset, submit proposal. | 23/01/2025 | 27/01/2025 |
| Initial Supervision | Discuss research focus and confirm dataset. | 27/01/2025 | 02/02/2025 |
| Background Research | Review heart failure prediction studies and ML methods. | 03/02/2025 | 23/02/2025 |
| Project & Data Management Plan | Prepare PDM document and presentation. | 06/02/2025 | 10/02/2025 |
| Supervisory Feedback | Discuss project roadmap and improvements. | 10/02/2025 | 16/02/2025 |
| Ethical Compliance | Review ethical guidelines and prepare for assessment. | 17/02/2025 | 20/03/2025 |
| Literature Review Draft | Submit and refine literature review with feedback. | 24/02/2025 | 02/03/2025 |
| Data Processing | Clean, encode, and normalize data; perform EDA. | 03/03/2025 | 09/03/2025 |
| Feature Selection | Apply PCA and IG for dataset refinement. | 10/03/2025 | 16/03/2025 |
| Ethics Quiz Submission | submit the ethics assessment. | 20/03/2025 | 20/03/2025 |
| Pre-Implementation Review | Discuss preliminary implementation and challenges. | 24/03/2025 | 30/03/2025 |
| Model Implementation | Build models like Logistic Regression, Naïve Bayes, etc. | 31/03/2025 | 06/04/2025 |
| Training & Tuning | Train models, optimize hyperparameters, evaluate. | 07/04/2025 | 13/04/2025 |
| Draft Report Discussion | Submit draft report and refine analysis. | 14/04/2025 | 20/04/2025 |
| Model Validation & Interpretation | Compare models, analyze metrics, document findings. | 21/04/2025 | 27/04/2025 |
| Final Report Submission | Submit completed research report. | 28/04/2025 | 29/04/2025 |
| Presentation & Examination | Prepare and attend viva session. | 30/04/2025 | 13/05/2025 |

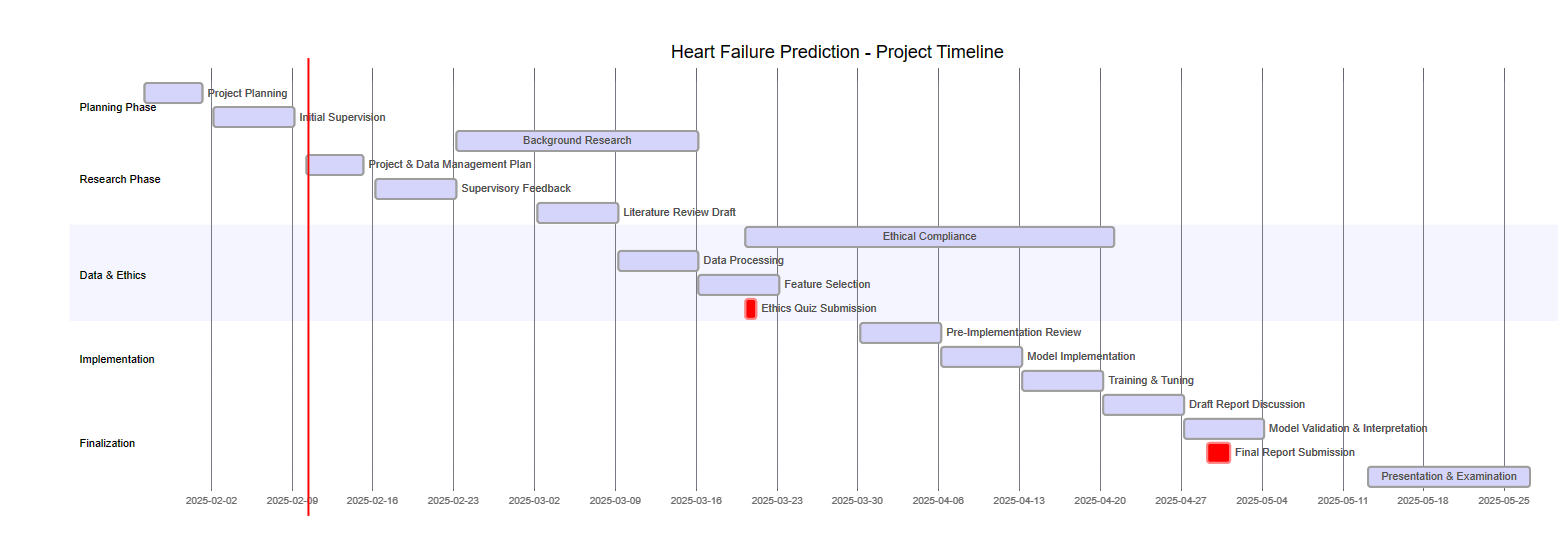


Figure 1 Gantt Chart

Data Management Plan

**Overview of the Dataset:** The heart failure prediction dataset, compiled from five medical datasets, consists of 918 unique observations with 11 clinical features for predicting cardiovascular disease events. The data originates from various medical institutions, including hospitals in Hungary, Switzerland, and the United States, and was curated by fedesoriano in September 2021.

**Data Collection:** The dataset will be downloaded from Kaggle (<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction> )

**Metadata:** The dataset is provided as a 35.92 kB CSV file, containing 11 clinical attributes related to cardiovascular disease risk and one target variable for classification.

**Document Control:** A GitHub repository (https://github.com/Aleena3443/Improving-Heart-Failure-Prediction-Accuracy-with-PCA-and-Information-Gain-in-Machine-Learning-Models.git) will be used to store the code, with weekly commits documenting changes. A standard file naming convention will be followed, and each commit will include a description of modifications.

**ReadMe File:** The ReadMe file will include a project description, instructions on dataset usage, installation steps, and dependencies. It will be structured to assist future users in understanding and utilizing the dataset and code.

**Security and Storage:** The dataset and code will be stored on GitHub, with regular backups to an online cloud storage system. Access to the repository will be restricted to project staff and authorized reviewers only.

Ethical Requirements: The dataset adheres to ethical standards through fully anonymized medical data, complying with GDPR requirements. It originates from reputable hospitals (Hungarian Institute of Cardiology, University Hospitals of Zurich and Basel, Cleveland Clinic). The Open Database License permits free research use, while credited medical professionals ensure ethical collection standards.