Machine Learning Model for Proactive Prediction of Cardiovascular Admission Outcomes

Thesis submitted in partial fulfillment of the
requirements for the

Post Graduate Certificate Program in
Data Science and Machine Learning

By

PRANAV MUKUND PANDE
MANJU ABRAHAM
CHANDANA. V
SOWMYA HL
MOUNIKA G
DINESH Y

Under the guidance of

Karthik

INSPIRED BY LIFE
MANIPAL
INSTITUTE OF TECHNOLOGY
(A constituent unit of MAHE, Manipal)

1. Acknowledgments

We wish to express our heartfelt gratitude to all those who have supported and guided us throughout this journey.

First and foremost, we extend our deepest appreciation to our families for their unwavering support and encouragement. Their belief in us and their constant motivation have been the cornerstone of our success. Their patience and understanding have provided us with the strength to pursue our goals with determination and enthusiasm.

We are profoundly grateful to Manipal Institute of Technology for organizing this remarkable course. The institution's commitment to fostering academic excellence and innovation has created an environment conducive to learning and growth. The resources and opportunities provided have been invaluable in achieving our objectives.

We would also like to convey our sincere thanks to the esteemed faculty for their exceptional teaching and guidance. Their dedication to imparting knowledge, coupled with their insightful feedback and mentorship, has greatly enhanced our understanding of the subject matter. Their expertise and encouragement have been instrumental in navigating the complexities of this project.

Each of these contributions has played a vital role in the successful completion of this endeavor. We are truly appreciative of the support we have received and are honored to acknowledge their impact on our academic and professional development.

2. Abstract

Predicting in-hospital outcomes for patients admitted with cardiovascular conditions is a critical component of modern healthcare, with significant implications for patient care, resource management, and overall healthcare efficiency. This project addresses the need for accurate, data-driven predictive models by developing a machine learning framework aimed at forecasting in-hospital outcomes, including mortality, heart failure, and major complications.

The dataset utilized for this project encompasses records from 14,845 admissions over a two-year period at the Hero Dayanand Medical College Heart Institute in Ludhiana, Punjab, India, involving 12,258 unique patients. After data preprocessing, the analysis focuses on 11,498 patients, incorporating a range of demographic, clinical, and biochemical variables.

The project employs various machine learning models, including Logistic Regression, Decision Trees, Random Forests, Gradient Boosting Machines, and Neural Networks, to predict key outcomes. The models are evaluated using metrics such as accuracy, precision, recall, F1-score and ROC-AUC. To ensure robustness and reliability, hyperparameter tuning and 10-fold cross-validation are applied to selected models. Additionally, permutation feature importance is used to enhance model interpretability and provide actionable insights for clinical decision-making.

Key findings include the successful development of a predictive model with enhanced accuracy and interpretability. The model's performance highlights its potential in improving risk stratification, optimizing resource utilization, and supporting clinical decision-making. The project demonstrates how machine learning can facilitate early intervention and personalized care, ultimately aiming to reduce complications and mortality rates among cardiovascular patients.

The implications of this work extend beyond the immediate clinical setting, contributing to better hospital management, more efficient use of resources, and informed public health initiatives. Future work will focus on integrating the model into clinical workflows, addressing limitations, and exploring additional data sources to further enhance predictive accuracy and generalizability.

In summary, this project showcases the transformative potential of machine learning in healthcare, underscoring the importance of predictive analytics in advancing patient care and operational efficiency within hospital settings.

3. Contents

4. List of Tables

5. List of Figures

6. Introduction

6.1. Motivation

Predicting outcomes of cardiovascular admissions is increasingly crucial in modern healthcare due to its far-reaching implications for patient care, hospital operations, and overall healthcare costs (Morgenstern et al., 2020). Accurate prediction models enable early identification of high-risk patients, allowing for timely interventions that can prevent complications, reduce mortality rates, and improve patient outcomes (Ravaut et al., 2021). By tailoring treatment strategies to individual risk profiles, healthcare providers can enhance personalized care and focus their efforts on those most in need.

Moreover, predictive models facilitate optimized resource utilization within hospitals (Mizan & Taghipour, 2022). Efficient allocation of critical resources such as ICU beds and medical staff can prevent overcrowding and ensure that resources are used where they are needed most. This, in turn, contributes to cost reduction by minimizing complications and optimizing the use of medical equipment.

In addition, data-driven decision-making supported by predictive models enhances clinical decision-making and triage efficiency. By providing evidence-based insights, these models assist clinicians in choosing appropriate interventions and prioritizing care in emergency situations. Furthermore, better hospital management is achieved through improved demand forecasting and strategic planning, which reduces readmissions and enhances overall quality of care.

On a broader scale, predictive analytics can inform public health initiatives by identifying trends and implementing preventive measures. By predicting the burden of cardiovascular conditions, public health authorities can allocate resources more effectively and develop targeted health programs.

6.2. Project Scope

The project focuses on developing a machine learning model to predict in-hospital outcomes for patients admitted to the Hero Dayanand Medical College Heart Institute, Ludhiana, Punjab, India. The dataset spans from April 1, 2017, to March 31, 2019, and includes records from 14,845 admissions, representing 12,258 unique patients. After filtering, data from 11,498 patients will be used for the analysis.

Data Scope:

- Source: Admission data from a single medical center.
- Period: Two-year timeframe.
- Type: Includes demographic, clinical, and admission-related features.

Algorithm Scope:

- Models: Logistic Regression, Gradient Boosting Machines, and Neural Networks.
- Focus: Supervised learning for predicting outcomes based on labeled data.

Outcome Scope:

- Predictive Target: In-hospital event: - “Heart Failure”.

- Metrics: Recall, Accuracy, precision, F1-score, and ROC-AUC.

Boundaries:

- Data: Limited to a two-year period and one medical center; only admission data is considered.
- Algorithms: Emphasis on supervised learning with considerations for model interpretability.
- Outcomes: Focused on in-hospital predictions without post-admission data or treatments.

6.3. Project Goal

The primary objectives of this project are to:

1. Develop a Predictive Machine Learning Model:
   - Create and validate a model that accurately predicts in-hospital outcomes such as heart failure and major complications based on admission data.
2. Enhance Risk Stratification:
   - Improve the risk assessment process for heart failure patients, enabling better prioritization and resource allocation.
3. Optimize Resource Utilization:
   - Assist hospitals in managing resources effectively by predicting patient needs and preventing shortages.
4. Support Clinical Decision-Making:
   - Provide actionable insights to aid clinicians in making informed decisions and integrating the model into clinical workflows.
5. Improve Patient Outcomes:
   - Facilitate early intervention and personalized care to reduce complications and mortality rates.
6. Ensure Model Interpretability:
   - Develop a model that offers clear, interpretable results to be trusted and understood by healthcare professionals.

6.4. Literature/Market Survey

Literature Review:

- Traditional Risk Scores:
  - Foundational tools like the Framingham Risk Score and Atherosclerotic Cardiovascular Disease (ASCVD) Risk Calculator provide estimates of cardiovascular risk based on clinical factors (Jahangiry, Farhangi, & Rezaei, 2017). These scores are widely used but may lack the precision and adaptability of machine learning models.
- Machine Learning Approaches:

  - Techniques such as Logistic Regression, Random Forests, Decision Tree, Gradient Boosting Machines, and Neural Networks offer advanced predictive capabilities. Recent studies have highlighted their effectiveness in cardiovascular risk prediction, with machine learning models providing improved accuracy and the ability to handle complex, non-linear relationships.

- Recent Advances:
  - Integration with Electronic Health Records (EHRs) has enhanced the applicability of predictive models (Ren et al., 2022). Explainable AI (XAI) and predictive analytics in ICUs address current trends and challenges, including data quality, generalizability, and model interpretability. These advances have facilitated more accurate predictions and better integration into clinical workflows.

Market Survey:

- Commercial Solutions:
  - Companies like IBM Watson Health (Kikuchi, Kadama, & Sengoku, 2021), Epic Systems (Chishtie et al., 2023), and Cerner (Ehwerhemuepha et al., 2022) offer predictive analytics and decision support tools integrated with EHR systems. These solutions provide valuable insights but often face challenges related to integration with existing workflows and scalability.
- Emerging Trends:
  - Real-time predictive analytics, wearable devices, and population health management tools are gaining traction (Beniczky, Karoly, Nurse, Ryvlin, & Cook, 2021). These trends indicate a growing focus on continuous monitoring and proactive health management.
- Market Gaps:
  - Challenges include integration with existing workflows, scalability, and the need for more interpretable models. Addressing these gaps is crucial for the successful adoption of predictive analytics in healthcare (Amarasingham, Patzer, Huesch, Nguyen, & Xie, 2014).

6.5. Organisation of the Report

The report is structured to provide a comprehensive overview of the project, from initial motivation to detailed results and future directions. The sections are organized as follows:

1. Introduction:
   - Motivation: Discusses the importance and impact of predictive modeling in cardiovascular care and its benefits for patient outcomes, resource utilization, and public health.
   - Project Scope: Defines the data, algorithms, and outcomes considered in the project, including the boundaries and limitations.
   - Project Goal: Outlines the primary objectives of the project, including model development, risk stratification, resource optimization, and clinical decision support.

   - Literature/Market Survey: Reviews existing research and market solutions, highlighting advances and gaps in cardiovascular risk prediction and machine learning applications.

2. Data Description:
   - Provides a detailed account of the dataset used, including data collection methods, preprocessing steps, and key variables.
3. Methodology:
   - Describes the machine learning algorithms used, including Logistic Regression, Gradient Boosting Machines, and Neural Networks. Discusses the feature selection process, model training, and evaluation techniques.
4. Key Results:
   - Output of Intermediate Steps: Details the results of the preliminary analysis, including data preprocessing, feature selection, and initial model performance.
   - Outcome: Presents the results of the final models, including performance metrics, confusion matrices, and sample predictions.
   - Analysis of the Results: Provides an in-depth analysis of the model performance, strengths, weaknesses, and areas for improvement.
5. Conclusion:
   - Summary of the Project Outcome: Summarizes the key findings of the project, including the effectiveness of the predictive models and their impact on patient care and resource management.
   - Future Work: Outlines recommendations for future research and development, including addressing class imbalance, optimizing models, and integrating solutions into clinical practice.
6. Appendices:
   - Includes additional details such as code snippets, model parameters, and supplementary tables or figures.
7. References:
   - Lists all sources cited in the report, including research papers, market reports, and relevant literature.

7. Project Description

7.1. Business/Domain Understanding

Cardiovascular diseases are a major cause of morbidity and mortality globally. Patients admitted to emergency departments or cardiac care units with cardiovascular conditions require prompt and accurate clinical interventions. However, the sheer volume and complexity of data collected at admission pose challenges for healthcare providers in predicting patient outcomes effectively.

Healthcare stakeholders, including doctors, physicians, clinicians, pathology laboratories, and hospitals, are under pressure to deliver efficient, data-driven care. The manual analysis of extensive demographic, clinical, and biochemical data can be time-consuming and prone to errors, leading to delayed or suboptimal decision-making.

This project addresses the need for a machine learning (ML) model that predicts key in-hospital outcomes, specifically heart failure. By leveraging ML techniques, the proposed decision support system can process large volumes of data, evaluate multiple variables simultaneously, and adjust dynamically to varying patient conditions. The model’s interpretability is ensured through permutation feature importance and aligning predictions with clinical knowledge.

The system aims to enhance hospital decision-making by providing timely, actionable insights. It will help prioritize high-risk patients, optimize resource use, and enable more accurate clinical decisions, ultimately improving patient outcomes and resource allocation in cardiovascular treatment units.

7.2. Project Stakeholders

1. Doctors/Physicians: Utilize the model’s predictions for timely and accurate clinical decisions, including risk stratification, treatment planning, and resource allocation.
2. Clinicians (Nurses, Care Coordinators): Use model outputs to prioritize care delivery, manage patient flow, and ensure high-risk patients receive necessary attention.
3. Cardiologists/Cardiac Specialists: Validate model outcomes and use predictions for advanced diagnostics, treatment, and surgical planning.
4. Pathologists and Laboratories: Provide diagnostic data that feeds into the model and use results to guide testing priorities.
5. Hospitals and Healthcare Administrators: Seek to optimize resource utilization, improve patient outcomes, and reduce costs through the model’s predictions.
6. IT/Clinical Informatics Teams: Support integration and maintenance of the ML system within the hospital’s electronic health record (EHR) system.
7. Data Scientists/ML Engineers: Develop, validate, and maintain the ML model to meet hospital needs.
8. Patients and Families: Benefit from improved care, timely interventions, and better health outcomes.
9. Insurance Companies/Payers: Interested in outcome predictions for cost estimation, treatment reduction, and value-based reimbursement.

10. Regulatory Bodies: Ensure model transparency and compliance with healthcare standards and regulations.

7.3. Datasets Understanding

The dataset comprises demographic, clinical, and lab parameters:

- Demographic Variables:
  - Age: Continuous variable (years).
  - Sex: Categorical (male or female).
  - Locality: Categorical (rural or urban).
- Admission Details:
  - Type of Admission: Categorical (emergency or outpatient).
  - Date of Admission and Date of Discharge: Used to calculate the duration of stay (outcome).
- Patient History (Comorbidities):
  - Smoking: Binary or categorical.
  - Alcohol: Binary or categorical.
  - Diabetes Mellitus (DM): Binary.
  - Hypertension (HTN): Binary.
  - Coronary Artery Disease (CAD): Binary.
  - Cardiomyopathy (CMP): Binary.
  - Chronic Kidney Disease (CKD): Binary.
- Lab Measurements:
  - Hemoglobin (HB), Total Lymphocyte Count (TLC), Platelets, Glucose, Urea, Creatinine, Brain Natriuretic Peptide (BNP), Raised Cardiac Enzymes (RCE), Ejection Fraction (EF).
- Comorbidities:
  - Heart Failure, ST-Segment Elevation Myocardial Infarction (STEMI), Pulmonary Embolism.
  - Shock: Non-cardiogenic, Cardiogenic, Multifactorial.
- Outcomes:
  - Discharge Status: Binary (alive/expired).

7.4. Data Limitations

1. Single-Center Data:
   - Limitation: Data from a single center limits generalizability across different settings.

   - Impact: Model may not perform well outside the specific hospital context.
   - Future Improvement: Include data from multiple centers for better generalizability.

2. Retrospective Nature of the Study:
   - Limitation: Potential biases and errors from previously recorded data.
   - Impact: Unforeseen biases may affect model performance.
   - Future Improvement: Validate model prospectively in real-world settings.
3. Short Time Frame:
   - Limitation: Data covers only two years, missing longitudinal trends.
   - Impact: Limited adaptability to changing clinical practices.
   - Future Improvement: Extend data collection period and update the model with new data.
4. No Inclusion of Post-Admission Data:
   - Limitation: Model lacks post-admission intervention data.
   - Impact: May weaken predictive power for cases with significant post-admission changes.
   - Future Improvement: Incorporate post-admission data and use time-series models.
5. Limited Features and Missing Clinical Factors:
   - Limitation: Missing factors like social determinants and genetic information.
   - Impact: Limited understanding of the full risk profile.
   - Future Improvement: Expand feature set to include more comprehensive patient information.
6. Bias in Dataset Composition:
   - Limitation: Possible over- or underrepresentation of certain patient groups.
   - Impact: Biased predictions for underrepresented groups.
   - Future Improvement: Balance sampling and increase data diversity.
7. Potential Overfitting:
   - Limitation: Risk of overfitting due to cross-validation on the same dataset.
   - Impact: Reduced performance on new, unseen data.
   - Future Improvement: Use external validation and regularization techniques.
8. Dataset Size and Imbalance:
   - Limitation: Dataset size may be small for complex models; potential class imbalances.
   - Impact: Poor performance on rare events.
   - Future Improvement: Use techniques like SMOTE and increase dataset size.
9. Lack of Interpretability in Neural Networks:

- Limitation: Neural networks often lack transparency.
- Impact: May hinder clinical adoption.
- Future Improvement: Employ explainable AI techniques like SHAP or LIME.

7.4. Benefits of Project

Short-Term Benefits:

1. Improved Clinical Decision-Making:
   - Benefit: Provides quick risk assessments, aiding in timely and informed decisions.
   - Impact: Reduced complications and adverse outcomes.
2. Enhanced Triage and Resource Allocation:
   - Benefit: Optimizes hospital resource allocation.
   - Impact: Reduced wait times and efficient use of resources.
3. Reduced Hospital Stays and Costs:
   - Benefit: Predicts duration of stay, aiding in planning and reducing unnecessary hospitalizations.
   - Impact: Cost savings and reduced strain on hospital services.
4. Actionable Insights for Clinicians:
   - Benefit: Permutation feature importance provides understanding of key factors.
   - Impact: Personalized and effective treatments.
5. Improved Documentation and Data Utilization:
   - Benefit: Streamlines data analysis and trend identification.
   - Impact: Reduced administrative burden and improved efficiency.

Long-Term Benefits:

1. Better Patient Outcomes and Reduced Mortality:
   - Benefit: Improved patient outcomes through early risk identification.
   - Impact: Reduced cardiovascular mortality and morbidity.
2. Scalable and Generalizable Model:
   - Benefit: Adaptable across various healthcare settings.
   - Impact: Consistent improvements in patient care globally.
3. Continuous Learning and Model Refinement:
   - Benefit: Evolving model with new data and clinical insights.
   - Impact: Ongoing accuracy and adaptation to new practices.
4. Proactive Healthcare and Early Intervention:
   - Benefit: Potential for predicting long-term outcomes and follow-up needs.

- Impact: Shifts to a proactive healthcare model.

5. Integration with Health Systems for Personalized Medicine:
    - Benefit: Personalized treatment based on individual risk profiles.
    - Impact: Enhanced treatment efficacy and sustainable healthcare.
6. Resource Optimization Across Healthcare Systems:
    - Benefit: Improved planning and resource management.
    - Impact: Greater efficiency and reduced waste in healthcare systems.
7. Facilitation of Clinical Research and Insights:
    - Benefit: Uncovers new patterns and research questions.
    - Impact: Advances in cardiovascular research and treatment protocols.

8. Exploratory Data Analysis

8.1. Data Collection

Study Setting

- Period: The study was conducted over a two-year period, from April 1, 2017, to March 31, 2019.
- Location: Hero Dayanand Medical College Heart Institute, Ludhiana, Punjab, India, a tertiary care center.
- Admissions: During the study period, there were 14,845 admissions from 12,258 patients.

Inclusion and Exclusion Criteria

- Inclusion: The dataset includes 11,498 unique patients. For patients with multiple admissions (totaling 1,921 patients), only the most recent admission was considered.
- Exclusion: 760 patients who were discharged against medical advice (DAMA) were excluded from the dataset.

Source and Format

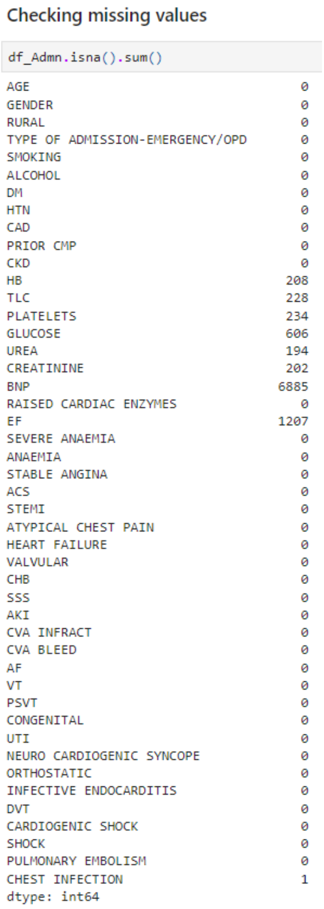
- Source: Electronic health records (EHRs) from the cardiac care unit of the tertiary hospital.
- Data Format: The dataset is structured (CSV), with each row representing an individual patient's admission details.

Variables Collected

- Demographic Information: Age, gender, rural/urban locality.
- Admission Details: Emergency vs. outpatient admission, date of admission, and date of discharge.
- Medical History: Information on comorbidities such as smoking, alcohol consumption, diabetes, hypertension, coronary artery disease (CAD), cardiomyopathy (CMP), and chronic kidney disease (CKD).
- Clinical Parameters: Lab results including hemoglobin (HB), total lymphocyte count (TLC), platelets, glucose, urea, creatinine, brain natriuretic peptide (BNP), and ejection fraction (EF).
- Outcomes: Mortality, heart failure, ST-segment elevation myocardial infarction (STEMI), pulmonary embolism, and duration of stay in the hospital.

Missing Data

- Initial Checks: The dataset has missing values.



Checking missing values

```
df_Admn.isna().sum()
```

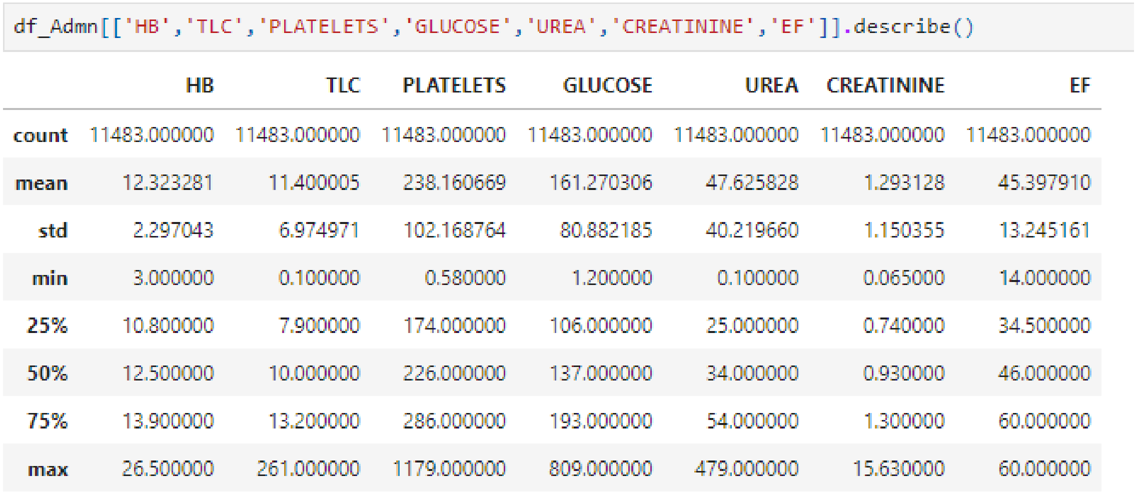
```
AGE                                0
GENDER                             0
RURAL                              0
TYPE OF ADMISSION-EMERGENCY/OPD    0
SMOKING                            0
ALCOHOL                            0
DM                                 0
HTN                                0
CAD                                0
PRIOR CMP                          0
CKD                                0
HB                               208
TLC                              228
PLATELETS                        234
GLUCOSE                          606
UREA                             194
CREATININE                       202
BNP                             6885
RAISED CARDIAC ENZYMES             0
EF                              1207
SEVERE ANAEMIA                     0
ANAEMIA                            0
STABLE ANGINA                      0
ACS                                0
STEMI                              0
ATYPICAL CHEST PAIN                0
HEART FAILURE                      0
VALVULAR                           0
CHB                                0
SSS                                0
AKI                                0
CVA INFRACT                        0
CVA BLEED                          0
AF                                 0
VT                                 0
PSVT                               0
CONGENITAL                         0
UTI                                0
NEURO CARDIOGENIC SYNCOPE          0
ORTHOSTATIC                        0
INFECTIVE ENDOCARDITIS             0
DVT                                0
CARDIOGENIC SHOCK                  0
SHOCK                              0
PULMONARY EMBOLISM                 0
CHEST INFECTION                    1
dtype: int64
```

Table 1: Missing values in the dataset before imputation

8.2. Data Exploration

Descriptive Statistics

- Numerical Variables: Summary statistics were computed, including mean, median, standard deviation, and interquartile range for numerical variable.



```
df_Admn[['HB','TLC','PLATELETS','GLUCOSE','UREA','CREATININE','EF']].describe()
```

	HB	TLC	PLATELETS	GLUCOSE	UREA	CREATININE	EF
count	11483.000000	11483.000000	11483.000000	11483.000000	11483.000000	11483.000000	11483.000000
mean	12.323281	11.400005	238.160669	161.270306	47.625828	1.293128	45.397910
std	2.297043	6.974971	102.168764	80.882185	40.219660	1.150355	13.245161
min	3.000000	0.100000	0.580000	1.200000	0.100000	0.065000	14.000000
25%	10.800000	7.900000	174.000000	106.000000	25.000000	0.740000	34.500000
50%	12.500000	10.000000	226.000000	137.000000	34.000000	0.930000	46.000000
75%	13.900000	13.200000	286.000000	193.000000	54.000000	1.300000	60.000000
max	26.500000	261.000000	1179.000000	809.000000	479.000000	15.630000	60.000000

Table 2: Descriptive statistics for numerical variables

Data Distributions

- Histograms: Examined distributions of lab results like glucose, urea, BNP, EF etc. to assess normality or skewness.

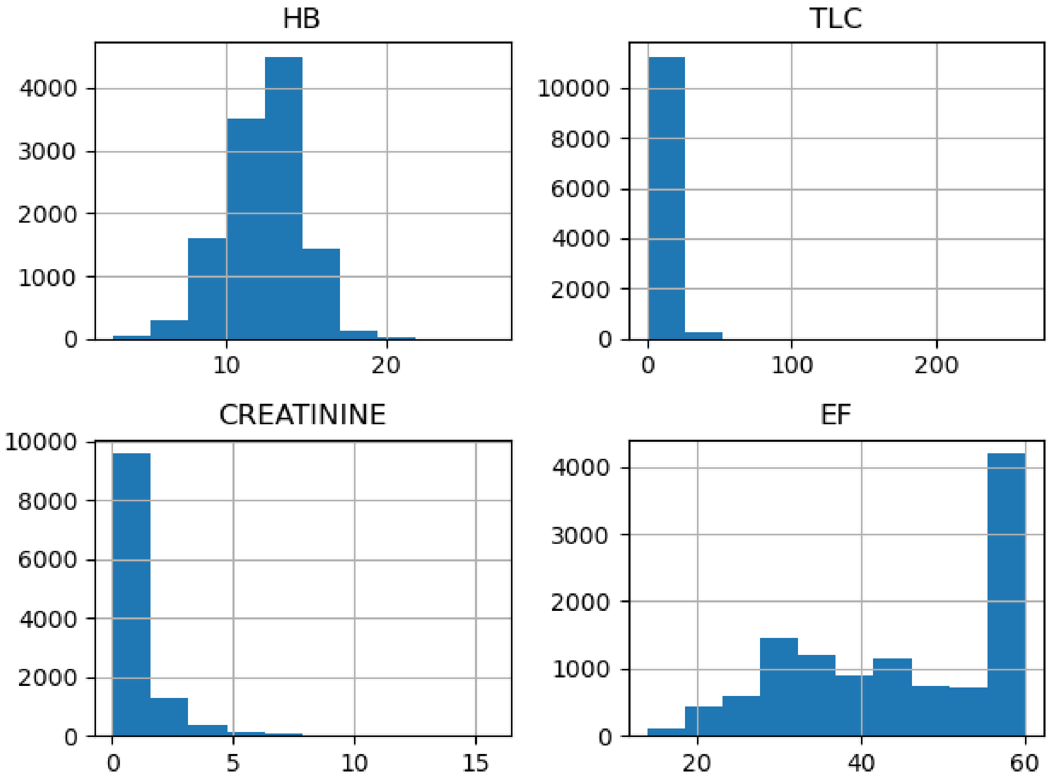


Fig. 1: Distributions of numeric variables - 'HB','TLC','CREATININE','EF' - After Imputation

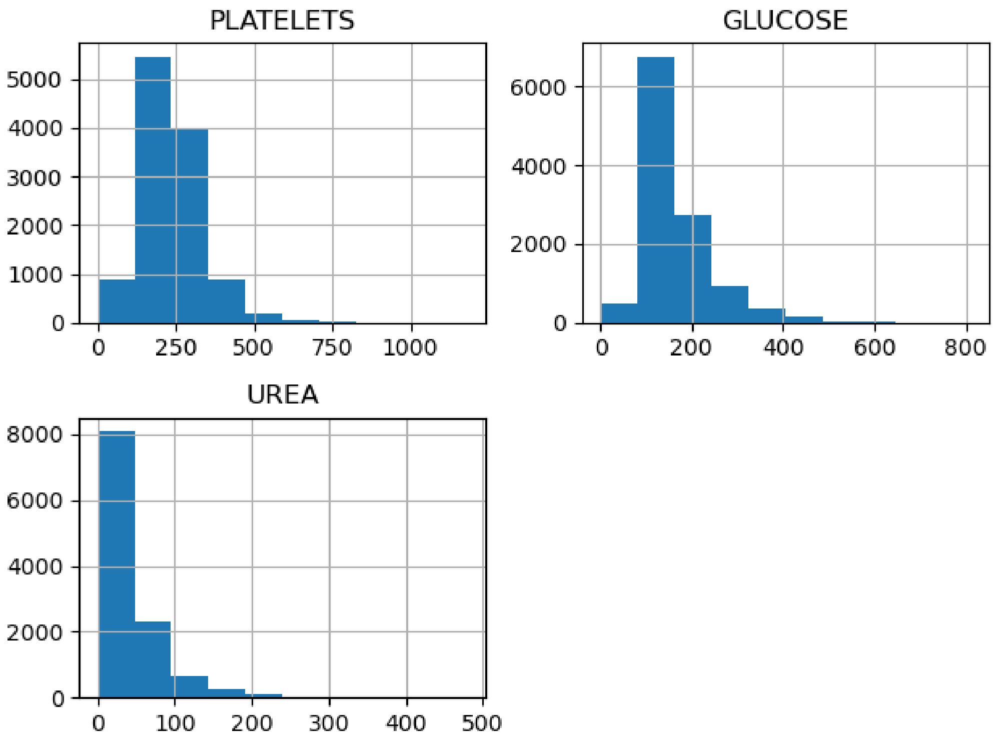


Fig.2: Distributions of numeric variables - 'PLATELETS','GLUCOSE','UREA' - After Imputation

Outliers (more details under section ‘Variable-Specific Outlier Analysis’)

- Identification: Detected extreme values in variables.

Correlations

- Correlation Matrix: Revealed relationships between numerical variables.

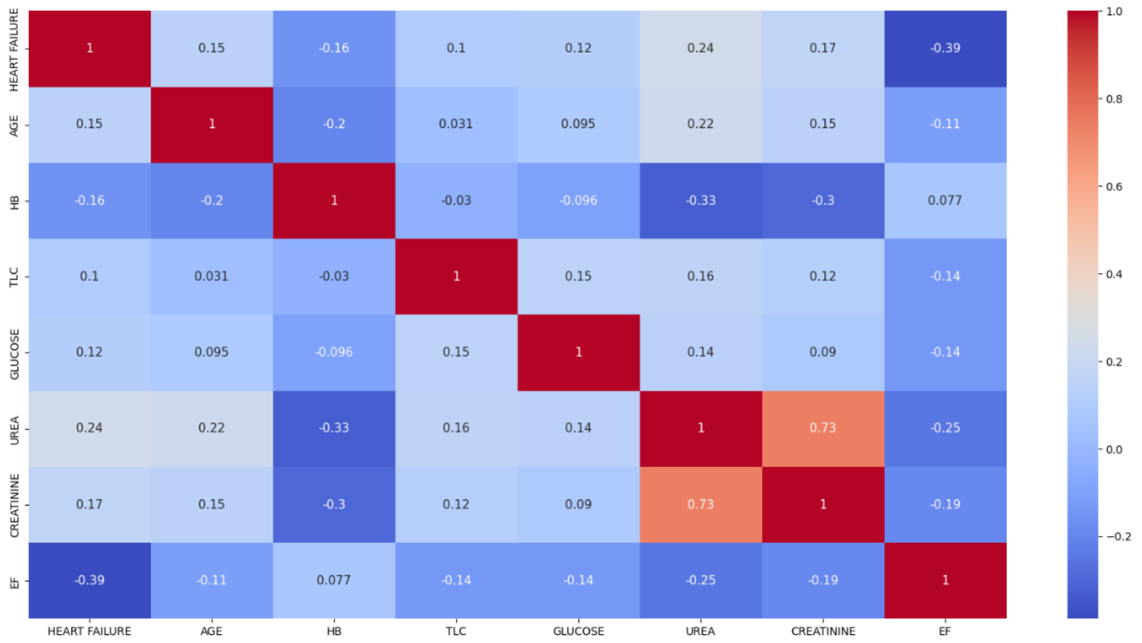


Fig.3: Correlation between heart failure and other lab parameters (1)

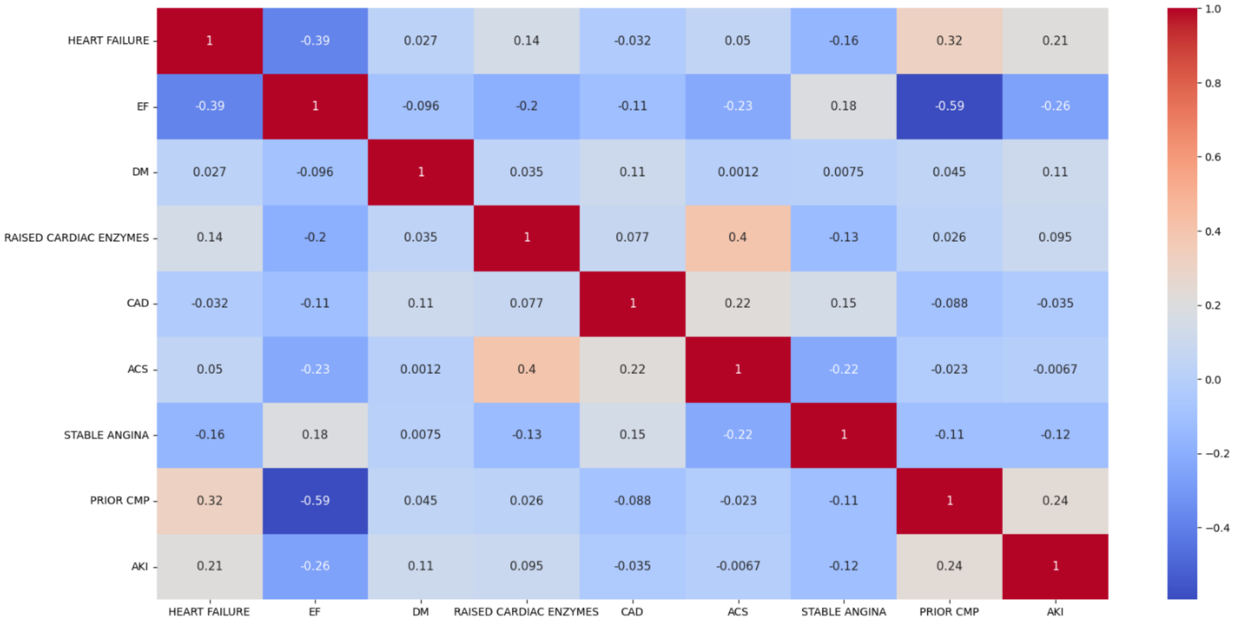


Fig.4: Correlation between heart failure and other lab parameters (2)

Example

- A correlation matrix may indicate a high correlation between urea and creatinine, suggesting that one of these variables might be redundant and could be excluded during feature selection.

8.3. Complexity of the Data

High Dimensionality

- Feature Count: The dataset includes 51 variables, which could be numerous and might contain redundant or highly correlated features (e.g., different lab tests showing similar patterns).

Multicollinearity

- Detection: Strong correlations between clinical parameters (e.g., urea and creatinine) or between multiple comorbidities (e.g., diabetes and hypertension) could affect model interpretability.
- VIF Analysis: Variance Inflation Factor (VIF) analysis was used to detect and mitigate multicollinearity.

Non-linearity

- Relationships: The relationships between features and outcomes such as heart failure or mortality may be non-linear, necessitating the use of advanced modeling techniques or transformations.

Class Imbalance

- Binary Classification: Imbalance in outcomes like pulmonary embolism or STEMI may skew model performance, requiring techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or class weighting.

Feature Interactions

- Complex Interactions: Variables may interact in complex ways (e.g., age combined with chronic kidney disease might influence outcomes differently than each factor alone), necessitating the assessment and modeling of these interactions.

Time Dependency

- Snapshot Data: The data is based on snapshots at admission time and may not account for changes due to treatments or interventions. Future time-series modeling could address this limitation.

8.4. Data Cleaning

Original Dataframe Shape

- Shape: (15,757 rows, 56 columns)

Column Renaming

- Consistency: Renamed columns for consistency:
    - 'SMOKING:' to 'SMOKING'
    - 'duration of intensive unit stay' to 'DURATION OF ICU STAY'

Duplicate Removal

- Duplicates: Removed 3,513 duplicate records.
- Latest Records: Kept the most recent record among duplicates to ensure each patient has only one record.

Exclusion of Specific Records

- Discharged Against Medical Advice: Removed 761 records for patients who were discharged against medical advice.

Redundant Variable Removal

- Dropped Columns:
    - 'SNO'
    - 'MRD No.'
    - 'D.O.A'
    - 'D.O.D'
    - 'month year'

Irrelevant Variable Removal

- Non-Predictive Variables: Dropped variables not relevant for predicting heart failure:
    - 'OUTCOME'
    - 'DURATION OF STAY'
    - 'DURATION OF ICU STAY'

Dropped Directly Predictive Variables

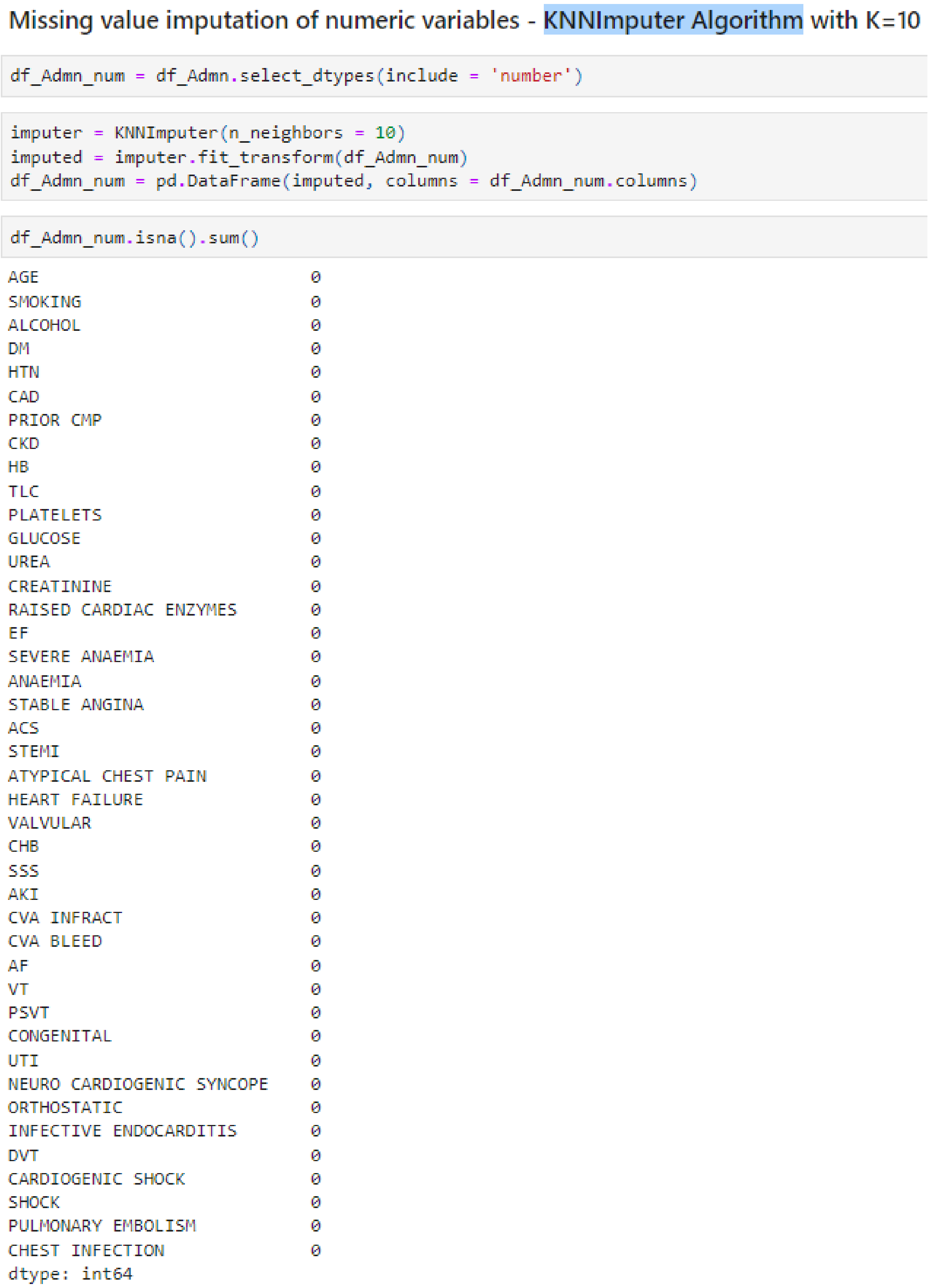
- Heart Failure Indicators: Removed variables that directly predict heart failure with 100% accuracy:
    - 'HFREF' (Heart Failure with Reduced Ejection Fraction)
    - 'HFNEF' (Heart Failure with Normal Ejection Fraction)

Datatype Conversion

- Conversion: Checked and converted datatypes from object to numeric where necessary to ensure proper data types for analysis.

Missing Values Handling

- BNP Column: Dropped the 'BNP' column due to more than 50% missing values.
- Imputation process: Applied imputation using KNNImputer Algorithm which fits the imputer to the data and transforms it by replacing missing values with KNN-imputed values.



Missing value imputation of numeric variables - KNNImputer Algorithm with K=10

```
df_Admn_num = df_Admn.select_dtypes(include = 'number')
```

```
imputer = KNNImputer(n_neighbors = 10)
imputed = imputer.fit_transform(df_Admn_num)
df_Admn_num = pd.DataFrame(imputed, columns = df_Admn_num.columns)
```

```
df_Admn_num.isna().sum()
```

```
AGE                          0
SMOKING                      0
ALCOHOL                      0
DM                           0
HTN                          0
CAD                          0
PRIOR CMP                    0
CKD                          0
HB                           0
TLC                          0
PLATELETS                    0
GLUCOSE                      0
UREA                         0
CREATININE                   0
RAISED CARDIAC ENZYMES       0
EF                           0
SEVERE ANAEMIA               0
ANAEMIA                      0
STABLE ANGINA                0
ACS                          0
STEMI                        0
ATYPICAL CHEST PAIN          0
HEART FAILURE                0
VALVULAR                     0
CHB                          0
SSS                          0
AKI                          0
CVA INFRACT                  0
CVA BLEED                    0
AF                           0
VT                           0
PSVT                         0
CONGENITAL                   0
UTI                          0
NEURO CARDIOGENIC SYNCOPE    0
ORTHOSTATIC                  0
INFECTIVE ENDOCARDITIS       0
DVT                          0
CARDIOGENIC SHOCK            0
SHOCK                        0
PULMONARY EMBOLISM           0
CHEST INFECTION              0
dtype: int64
```

Table. 3: Missing value imputation of numeric variables - KNNImputer Algorithm with K=10

Final Dataframe Shape

- Shape: (11,483 rows, 45 columns)

8.5. Data Transformation

K-Nearest Neighbors (KNN) Imputation

- Method: Used KNN Imputer from sklearn.impute to handle missing values in numeric columns.
- Algorithm: KNN Imputation fills missing values based on the values of the 10 nearest neighbors, using the mean (or weighted average) of these neighbors.

Imputation Process

- Application: Applied imputation using fit_transform(), which fits the imputer to the data and transforms it by replacing missing values with KNN-imputed values.

Resulting Dataframe

- Conversion: The imputed values were converted back into a pandas DataFrame, maintaining the original column names for consistency.

Outlier Analysis

Outlier Detection Method

- Method: Used the Interquartile Range (IQR) method to detect outliers. Identified outliers as data points falling outside the range defined by 1.5 times the IQR from the 25th (Q1) and 75th (Q3) percentiles.

Variable-Specific Outlier Analysis

- AGE:
  - Number of Outliers: 224 outliers detected.
  - Anomalies: Records with age exceeding 100 were removed.
- Hemoglobin (HB):
  - Number of Outliers: 108 outliers detected; no specific anomalies noted for removal.
- Total Leukocytes Count (TLC):
  - Number of Outliers: 582 outliers identified.
  - Anomalies: Records with TLC < 1 K per microliter were removed.
- Creatinine:
  - Number of Outliers: 1,230 outliers found; no specific anomalies highlighted for removal.
- Ejection Fraction (EF):
  - Number of Outliers: 0 outliers detected.
- Glucose:
  - Number of Outliers: 566 outliers detected; no specific anomalies removed.
- Urea:
  - Number of Outliers: 1,073 outliers detected; no specific anomalies removed.
- Platelets:
  - Number of Outliers: 200 outliers detected; no specific anomalies highlighted for
  - Anomalies: Records with Platelets < 10 K per microliter were removed.

Feature Scaling (section 9.5. for further details)

- Scaling: Applied RobustScaler to numerical features using median and interquartile range as the distributions were skewed. This was essential for the models used, including logistic regression and neural networks.

Categorical Variable Encoding

- One-Hot Encoding: To convert categorical variables into numerical form. This method was used for variables like gender, admission type, and comorbidities.

Variance Inflation Factor (VIF) Analysis (section 9.5. for further details)

- Method: Performed VIF analysis to identify and remove features with high multicollinearity. Variables with a VIF greater than 10 were considered for removal to mitigate multicollinearity.

Feature Importance (done after training the models)

- Permutation Importance: Applied permutation feature importance to assess the impact of each feature on model performance. Features with low importance scores were considered for exclusion to improve model efficiency and interpretability.

9. Design

9.1. Analytical Methods and Technology Used

To analyze the dataset, a comprehensive approach was adopted, utilizing various statistical tests and machine learning techniques. The following methods and technologies were employed:

- Statistical Tests: Shapiro-Wilk Test, Jarque-Bera Test, Two-Sample t-Test, and Chi-Square Test.
- Machine Learning: Utilized various models, including Logistic Regression, XGBoost, and Neural Networks (ANN) for predictive analysis.
- Scaling: Employed RobustScaler for feature scaling to handle outliers and ensure feature equality.
- Feature Selection: Used Variance Inflation Factor (VIF) to detect multicollinearity and Chi-Square Test for categorical variables.

9.2. Descriptive Statistical Analysis

Normality Test Summary for Numeric Variables: Two statistical tests were used to assess the normality of distribution for selected numeric variables: the Shapiro-Wilk Test and the Jarque-Bera Test. Both tests have a null hypothesis (H0) that assumes data are normally distributed and an alternative hypothesis (H1) that assumes data are not normally distributed.

- Variables Tested:
  - AGE
  - Hemoglobin (HB)
  - Total Leukocyte Count (TLC)
  - Creatinine
  - Ejection Fraction (EF)
  - Platelets
  - Glucose
  - Urea

Shapiro-Wilk Test Results: For all variables, the p-value was less than 0.05, leading to the rejection of the null hypothesis. This indicates that none of the variables are normally distributed.

- AGE: Not normally distributed (p-value = 2.03e-34)
- HB: Not normally distributed (p-value = 2.72e-23)
- TLC: Not normally distributed (p-value = 0.0)
- Creatinine: Not normally distributed (p-value = 0.0)
- EF: Not normally distributed (p-value = 0.0)

- Platelets: Not normally distributed (p-value = 0.0)
- Glucose: Not normally distributed (p-value = 0.0)
- Urea: Not normally distributed (p-value = 0.0)

Jarque-Bera Test Results: Similarly, all variables showed p-values less than 0.05, confirming that the data is not normally distributed.

- AGE: Not normally distributed (p-value = 1.57e-165)
- HB: Not normally distributed (p-value = 2.72e-43)
- TLC: Not normally distributed (p-value = 0.0)
- Creatinine: Not normally distributed (p-value = 0.0)
- EF: Not normally distributed (p-value = 4.44e-211)
- Platelets: Not normally distributed (p-value = 0.0)
- Glucose: Not normally distributed (p-value = 0.0)
- Urea: Not normally distributed (p-value = 0.0)

Conclusion: Both tests indicate that none of the numeric variables follow a normal distribution. This suggests that transformations or non-parametric methods might be more suitable for further analysis, particularly in modeling or statistical testing where normality is a key assumption.

9.3. Data Visualization

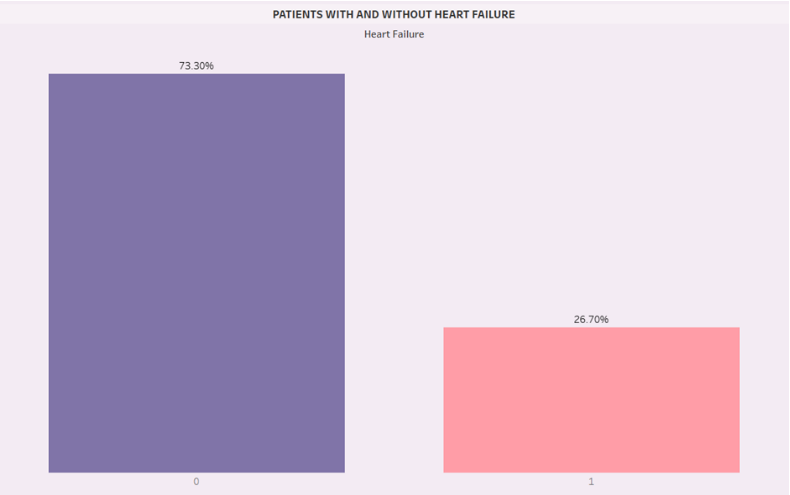


Fig. 5: Heart Failure Incidence: 26.70% of the cardiac patients have heart failure with 60.89% of these patients being male

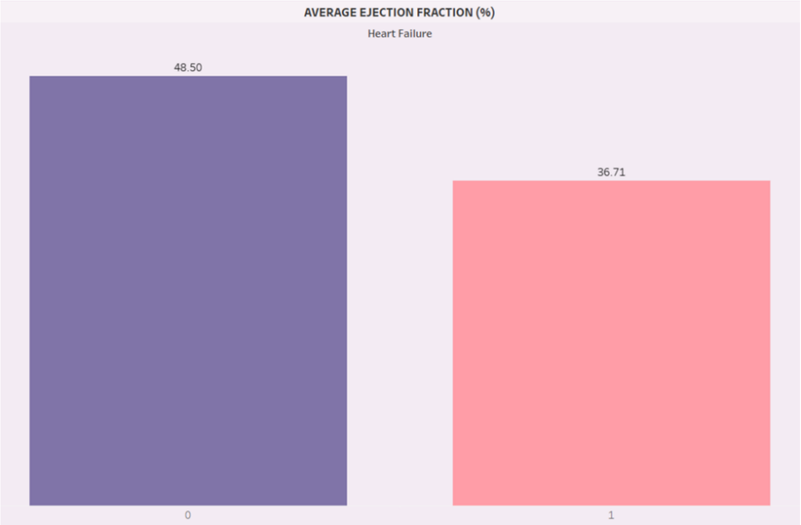


Fig. 6: Ejection Fraction: Patients with heart failure have a lower average ejection fraction(%) compared to those without heart failure

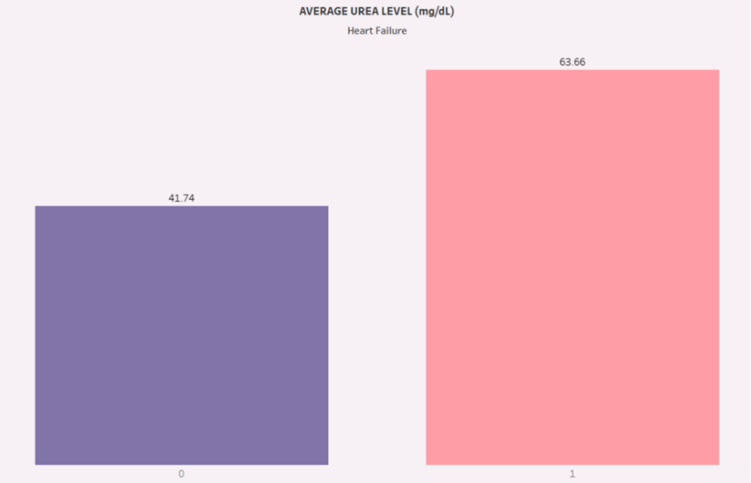
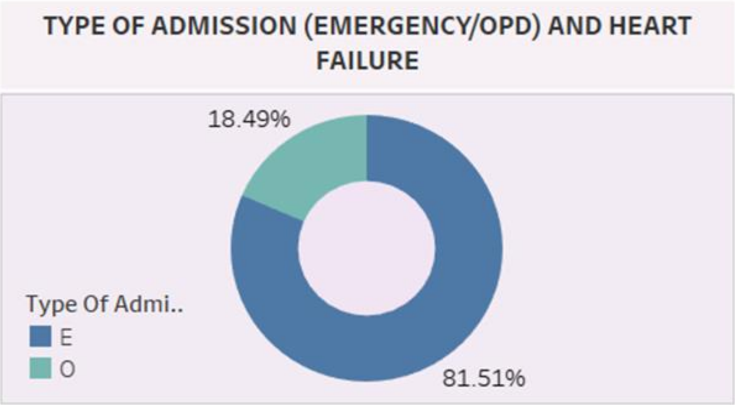


Fig. 7: Urea Level: The average urea level(mg/dL) is higher in patients with heart failure than in those without



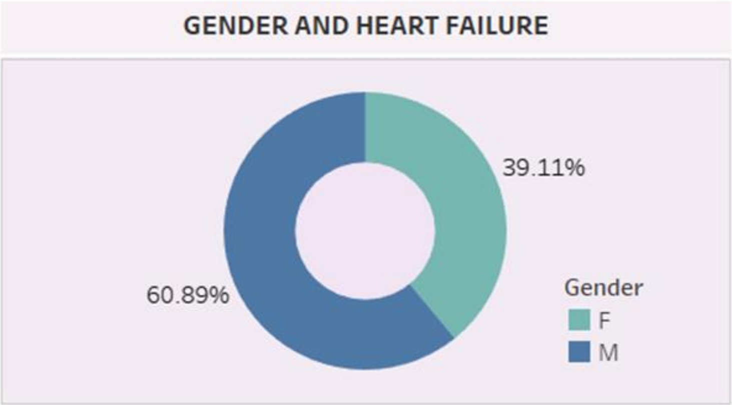


Fig. 8: Type of admission (Emergency vs. OPD) for individuals with heart failure

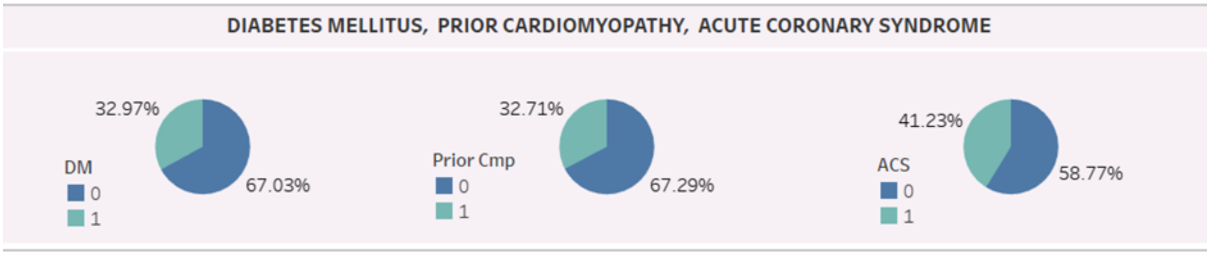


Fig. 9: Percentage of individuals with Diabetes Mellitus, Prior Cardiomyopathy and Acute Coronary Syndrome

9.4. Hypothesis Testing:

Using Two-Sample t-Test

The objective was to test whether there is a statistically significant relationship between HEART FAILURE (a binary categorical variable) and various numeric variables in the dataset.

Process:

1. Data Segmentation:
   - Divided the dataset into two groups:
     - Patients with HEART FAILURE (HEART FAILURE = 1)
     - Patients without HEART FAILURE (HEART FAILURE = 0)
2. Levene’s Test for Variance:

   - Checked whether the two groups have equal variance.
   - Null Hypothesis (H0): Variance between the two groups is equal.
   - Alternative Hypothesis (H1): Variance between the two groups is not equal.
   - If the p-value was greater than 0.05, the assumption of equal variance was accepted; otherwise, it was rejected.

3. Two-Sample t-Test:
   - Tested whether the means of the numeric variable differ significantly between the two groups.
   - Null Hypothesis (H0): The means of the two groups are equal (no relationship).
   - Alternative Hypothesis (H1): The means of the two groups are not equal (there is a relationship).
   - If the p-value was less than 0.05, the null hypothesis was rejected.

Results:

- Significant Relationships (p-value < 0.05):
  - AGE
  - Hemoglobin (HB)
  - Total Leukocyte Count (TLC)
  - Glucose
  - Urea
  - Creatinine
  - Ejection Fraction (EF)
- Insignificant Relationship (p-value > 0.05):
  - Platelets

Action Taken:

- The Platelets variable was dropped from the dataset as it was not significantly related to HEART FAILURE.

Chi Square Test

To test the association of categorical variables with HEART FAILURE

Results:

Removed insignificant Relationship (p-value > 0.05): RURAL, HTN, STEMI, VALVULAR, CHB, CVA INFRACT, CVA BLEED, UTI, ORTHOSTATIC, INFECTIVE ENDOCARDITIS

Conclusion: Insignificant features related to HEART FAILURE were removed from further analysis. This step helps in reducing dimensionality, improving model performance, and interpretability.

The final shape of the dataframe was (11459, 34)

9.5. Feature Engineering

Importance of Scaling:

1. Handling Different Ranges:
   - Numerical features often have different ranges. For instance, AGE ranges from 0 to 100, while CREATININE ranges from 0.1 to 5. Without scaling, features with larger ranges may dominate models like neural networks or distance-based algorithms.
2. Improving Model Performance:
   - Scaling ensures that all features contribute equally during training, helping the model converge faster and more stably.
3. RobustScaler:
   - Used for scaling numerical variables to handle outliers effectively. RobustScaler scales features using the interquartile range (IQR), making it less affected by extreme values compared to standard scaling methods.

Results of Scaling:

- Numerical Variables Scaled:
  - AGE
  - Hemoglobin (HB)
  - Total Leukocyte Count (TLC)
  - Creatinine
  - Ejection Fraction (EF)
  - Glucose
  - Urea
- Transformation:
  - Features were scaled around the median and scaled according to IQR. This transformation ensures that each feature contributes equally, preventing bias due to larger values.

Conclusion: RobustScaler was used to scale the numerical features effectively while handling outliers, leading to improved model performance and convergence.

Importance of VIF (Variance Inflation Factor) Calculation:

1. Detecting Multicollinearity:
   - High multicollinearity inflates the variance of regression coefficients, leading to unreliable estimates and making the model less interpretable.
2. Model Stability:
   - Multicollinearity affects coefficients and standard errors, making the model unstable and sensitive to changes in data.
3. Feature Selection:

   - Identifying variables with high VIF scores helps in removing or combining highly correlated features to improve model performance and interpretability.

4. Interpretability:
   - Lower multicollinearity improves interpretability, as the independent effects of variables are clearer.

VIF Calculation Process and Results:

- Findings:
  - None of the variables had a VIF higher than 5, a common threshold for concern.
  - Top features with the highest VIF values included:
    - CREATININE: 3.99
    - CAD (Coronary Artery Disease): 3.11
    - UREA: 2.91
    - ACS (Acute Coronary Syndrome): 2.80
    - AKI (Acute Kidney Injury): 2.77
  - Features like AGE, SMOKING, and ALCOHOL had very low VIF values (around 1), indicating no significant multicollinearity.

Conclusion: All VIF values are below 5, suggesting that multicollinearity is not an issue. No variables need to be removed based on this analysis, allowing the use of all features in the machine learning model.

Feature selection of categorical variables was done using Chi Square Test - test of association and with HEART FAILURE.

Feature selection of numerical variables was done using two sample t-Test.

10. Modeling

10.1. Selection of Model/Technique

Logistic Regression

- Method: Logistic Regression using Maximum Likelihood Estimation (MLE)
- Initial Model: Included all 33 features (excluding HEART FAILURE)
- Final Model: Refined to 20 significant predictors

XGBoost

- Initial Model:
  - Algorithm: XGBoost Classifier
  - Parameters: max_depth = 10, gamma = 1, random_state = 10
  - Train/Test Split: 70% train, 30% test
- Hyperparameter Tuning:
  - Parameters Tuned: learning_rate, max_depth, gamma, min_child_weight, n_estimators
  - Best Parameters: gamma = 4, learning_rate = 0.2, max_depth = 5, min_child_weight = 4, n_estimators = 20

ANN (Artificial Neural Network)

- Model Architecture:
  - Layers: Input layer followed by four hidden layers with LeakyReLU activation and Dropout layers
  - Output Layer: Sigmoid activation function for binary classification
- Training:
  - Epochs: 100
  - Batch Size: 64
  - Optimizer: Adam optimizer with a learning rate of 0.001
  - Loss Function: Binary cross-entropy

10.2. Challenges Faced

Logistic Regression

- Class Imbalance: The model showed higher precision and recall for Class 0 (No Heart Failure) compared to Class 1 (Heart Failure), indicating potential dataset imbalance.
- Recall for Class 1: A recall of 0.42 on the test set revealed that 58% of actual heart failure cases are missed, posing a critical issue in healthcare settings.

XGBoost

- Class Imbalance: The model exhibited a performance disparity between the two classes, with lower performance on the minority class (heart failure).
- Recall for Class 1: The recall for heart failure cases remains low, indicating that a significant number of heart failure cases are missed.

ANN

- Class Imbalance: Similar to Logistic Regression and XGBoost, the ANN model faced challenges in identifying heart failure cases due to class imbalance.
- Computational Resources: The ANN model's complexity requires significant computational resources, making training and optimization time-consuming.

10.3. Evaluation and Cross-Validation

Logistic Regression

- Test Set Performance:
  - Accuracy: 78%
  - Precision: Class 0 = 0.81, Class 1 = 0.62
  - Recall: Class 0 = 0.91, Class 1 = 0.42
  - F1-Score: Class 0 = 0.86, Class 1 = 0.50
  - AUC: 0.7919
- Train Set Performance:
  - Accuracy: 78%
  - Precision: Class 0 = 0.81, Class 1 = 0.65
  - Recall: Class 0 = 0.92, Class 1 = 0.41
  - F1-Score: Class 0 = 0.86, Class 1 = 0.50
  - AUC: 0.8068

XGBoost

- Model with Best Parameters Performance:
  - Train Set:
    - Accuracy: 81%
    - Precision: Class 0 = 0.83, Class 1 = 0.69
    - Recall: Class 0 = 0.92, Class 1 = 0.49
    - F1-Score: Class 0 = 0.87, Class 1 = 0.57
    - AUC: 0.7058

  - Test Set:
    - Accuracy: 79%
    - Precision: Class 0 = 0.82, Class 1 = 0.63
    - Recall: Class 0 = 0.90, Class 1 = 0.47
    - F1-Score: Class 0 = 0.86, Class 1 = 0.54
    - AUC: 0.6861

ANN

- Model with Best Parameters Performance:
  - Train Set:
    - Accuracy: 79%
    - Precision: Class 0 = 0.84, Class 1 = 0.62
    - Recall: Class 0 = 0.88, Class 1 = 0.55
    - F1-Score: Class 0 = 0.86, Class 1 = 0.58
    - AUC: 0.8153
  - Test Set:
    - Accuracy: 78%
    - Precision: Class 0 = 0.83, Class 1 = 0.59
    - Recall: Class 0 = 0.87, Class 1 = 0.53
    - F1-Score: Class 0 = 0.85, Class 1 = 0.56
    - AUC: 0.8019

10.4. Model Interpretation

Logistic Regression

- Significant Predictors:
  - Age and Ejection Fraction (EF) are strong predictors of heart failure.
  - Lower Hemoglobin levels and elevated glucose and urea levels are associated with increased heart failure risk.
  - Clinical history factors such as diabetes, coronary artery disease and prior cardiomyopathy are important.
- Negative Associations:
  - Variables like Pulmonary Embolism and Neurocardiogenic Syncope have negative coefficients, suggesting a complex relationship with heart failure.

XGBoost

- Feature Importance:

- Top Features:
    - EF (0.2436)
    - TYPE OF ADMISSION-EMERGENCY/OPD_O (0.0659)
    - UREA (0.0622)
    - PRIOR CMP_1.0 (0.0476)
    - STABLE ANGINA_1.0 (0.0465)
- Least Important Features:
    - DVT_1.0 (0.0000)
    - CKD_1.0 (0.0000)
    - CONGENITAL_1.0 (0.0000)

ANN

- Performance Metrics:
    - Before Feature Selection: The ANN model exhibited strong AUC scores, indicating good discriminative ability.
    - After Feature Selection: Feature selection improved the recall for heart failure cases slightly, demonstrating the importance of focusing on key features.

10.5. What Worked/What Did not Work

Logistic Regression

- What Worked:
    - The model demonstrated good discriminative ability with an AUC around 0.79-0.81, indicating effective differentiation between heart failure and non-heart failure cases.
    - Consistent performance between train and test sets suggests that the model is not overfitting.
- What Didn’t Work:
    - The model's performance in detecting heart failure cases (Class 1) is suboptimal, primarily due to class imbalance and low recall.
    - The imbalance between precision and recall highlights the need for strategies to improve performance on the minority class (heart failure).

XGBoost

- What Worked:
    - The model demonstrated good performance with high accuracy on the train set and reasonable performance on the test set.
    - Hyperparameter tuning improved model performance, although issues with class imbalance persisted.

- What Did not Work:
    - Despite hyperparameter tuning, the recall for heart failure cases remained low.
    - Class imbalance led to poorer performance for the minority class, indicating the need for additional techniques to handle imbalance.

ANN

- What Worked:
    - The ANN model showed robust performance before feature selection, with good AUC scores indicating effective prediction ability.
    - Feature selection improved the model's ability to identify heart failure cases slightly.
- What Did not Work:
    - The model's complexity made it computationally expensive and time-consuming.
    - Recall for heart failure cases remained lower than desired, suggesting further improvements are needed for effective prediction.

10.6. Model Deployment Considerations

- Class Imbalance Handling: Techniques such as SMOTE or class weighting should be considered to address the imbalance and improve model performance for the minority class.
- Feature Selection: Continuous refinement of features based on importance scores can enhance model performance and interpretability.

11. Key Results

11.1. Output of Intermediate Steps

Data Preprocessing and Feature Engineering

- Initial Dataset: The original dataset comprised 51 columns and 11,483 entries, with no missing values. The key variables included demographic data, clinical parameters, and biochemical readings.
- Feature Selection:
  - Logistic Regression: The initial model included all 33 features. After refining, 20 significant predictors were selected based on the VIF calculation and correlation analysis.
  - XGBoost: The model was trained with default parameters initially. Hyperparameter tuning was performed using GridSearchCV, optimizing for gamma, learning rate, max_depth, min_child_weight, and n_estimators.
  - ANN: The model architecture included an input layer, four hidden layers with LeakyReLU activation and Dropout layers, and an output layer with a sigmoid activation function. Training involved 100 epochs with a batch size of 64 and used the Adam optimizer with a learning rate of 0.001.

Model Evaluation

- Logistic Regression:

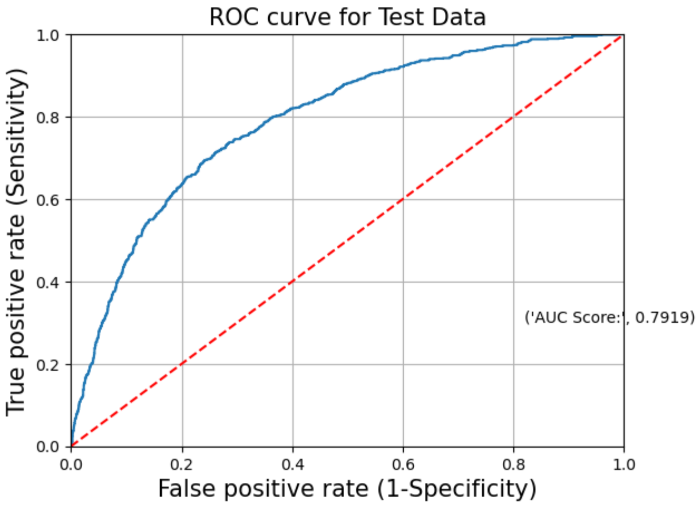


Fig 10: ROC for test data for the model with best parameters (Logistic regression)

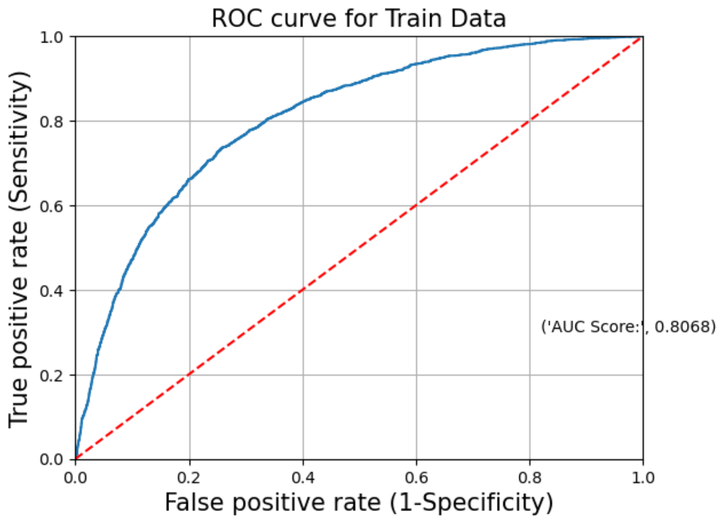
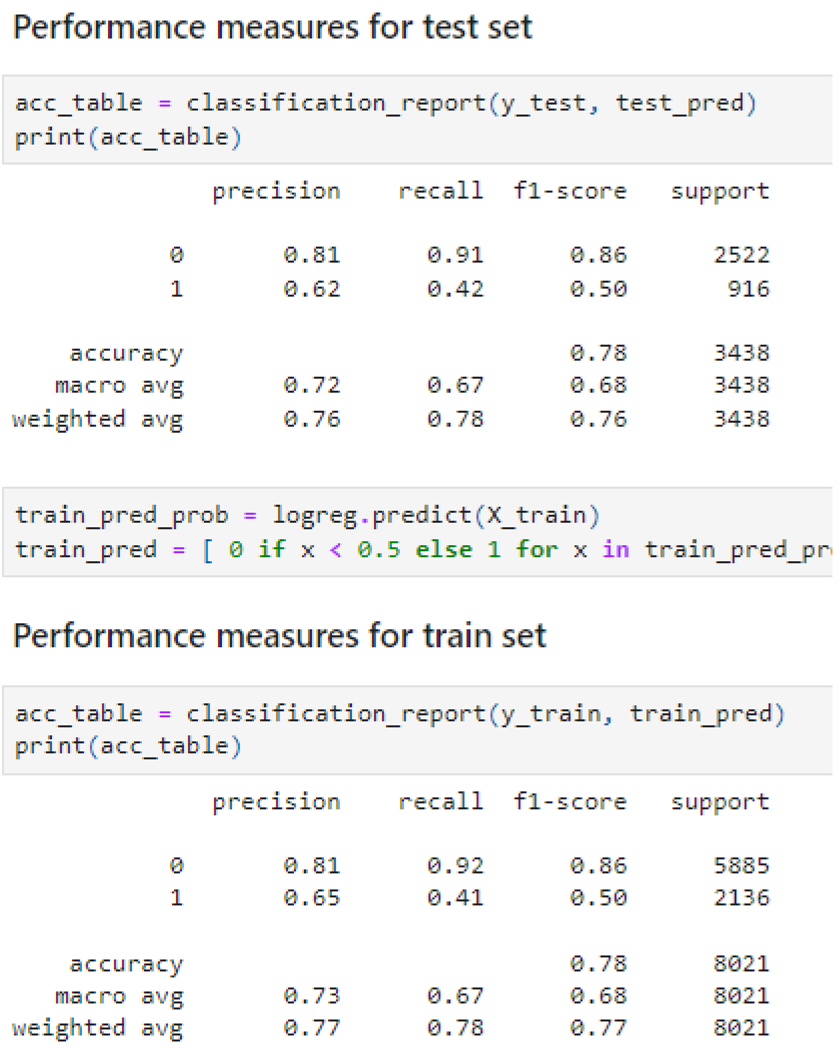


Fig 11: ROC for train data for the model with best parameters (Logistic regression)



Performance measures for test set

```
acc_table = classification_report(y_test, test_pred)
print(acc_table)
```

```
              precision    recall  f1-score   support

           0       0.81      0.91      0.86      2522
           1       0.62      0.42      0.50       916

    accuracy                           0.78      3438
   macro avg       0.72      0.67      0.68      3438
weighted avg       0.76      0.78      0.76      3438
```

```
train_pred_prob = logreg.predict(X_train)
train_pred = [ 0 if x < 0.5 else 1 for x in train_pred_pr
```

Performance measures for train set

```
acc_table = classification_report(y_train, train_pred)
print(acc_table)
```

```
              precision    recall  f1-score   support

           0       0.81      0.92      0.86      5885
           1       0.65      0.41      0.50      2136

    accuracy                           0.78      8021
   macro avg       0.73      0.67      0.68      8021
weighted avg       0.77      0.78      0.77      8021
```

Table 4 : Logistic regression classification report with best parameters

- XGBoost:

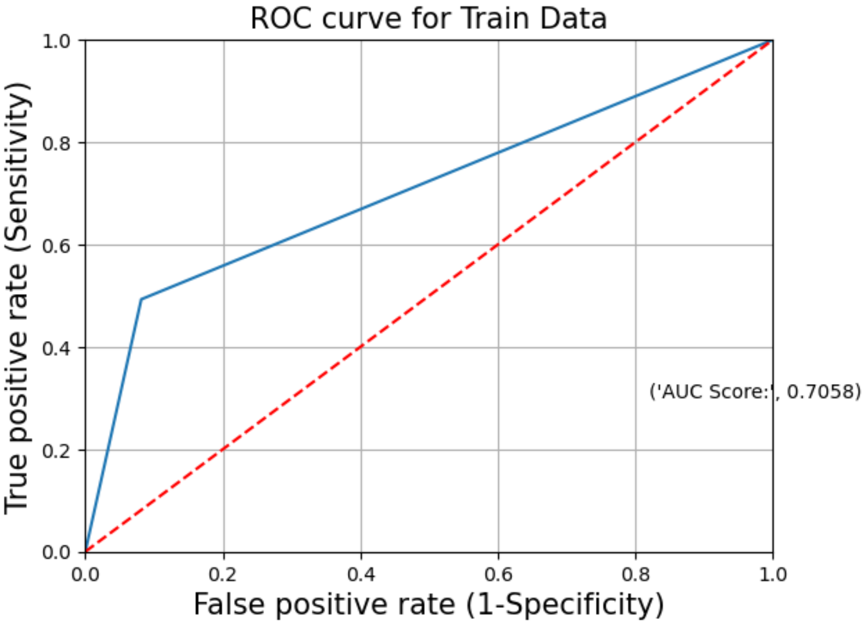


Fig 12: ROC for train data (XGBoost)

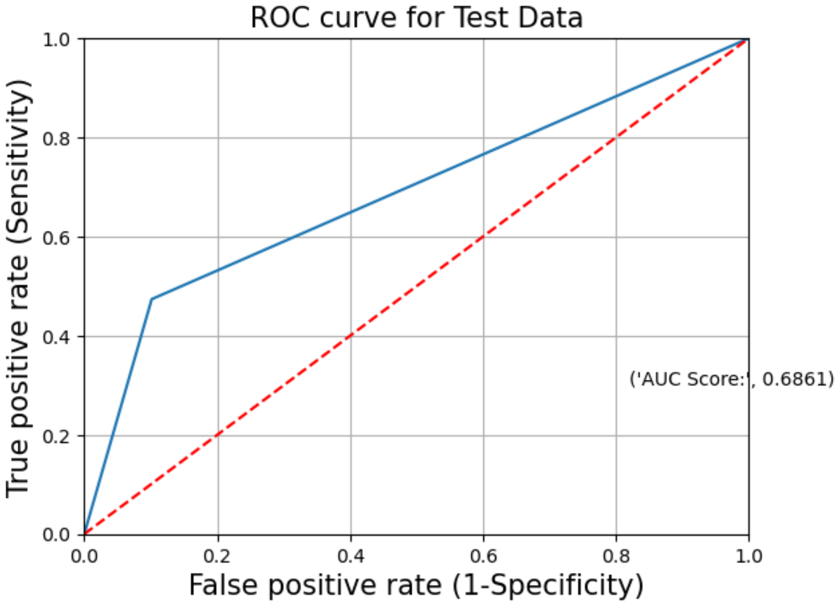
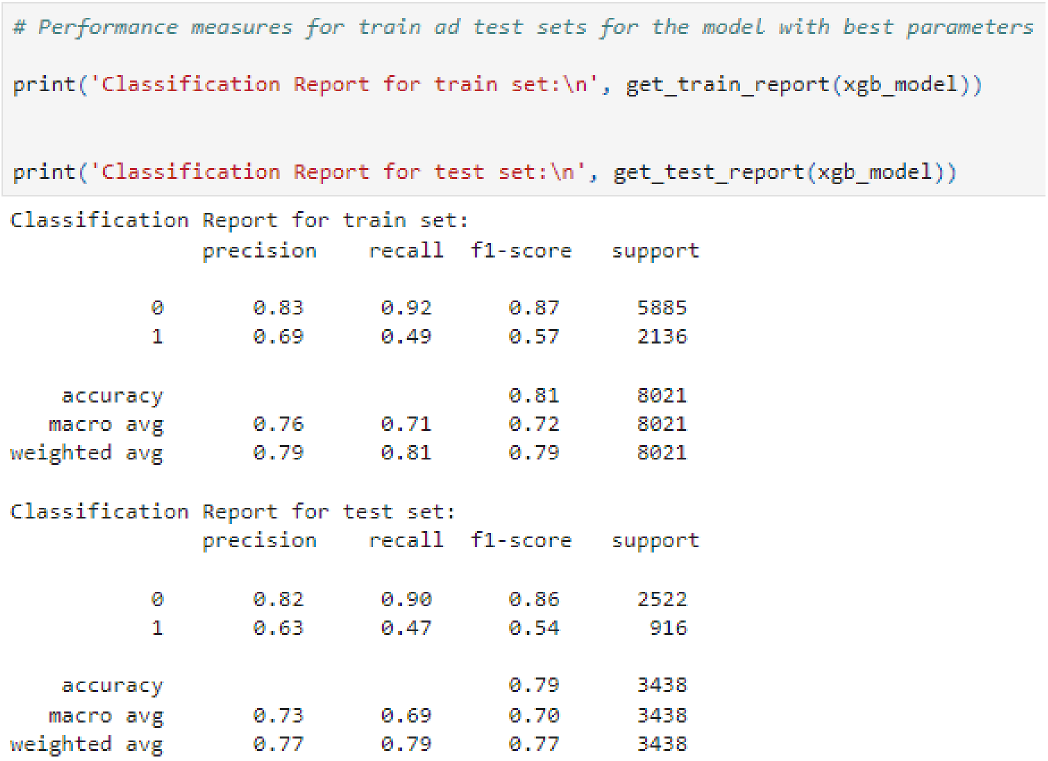


Fig 13: ROC for test data (XGBoost)

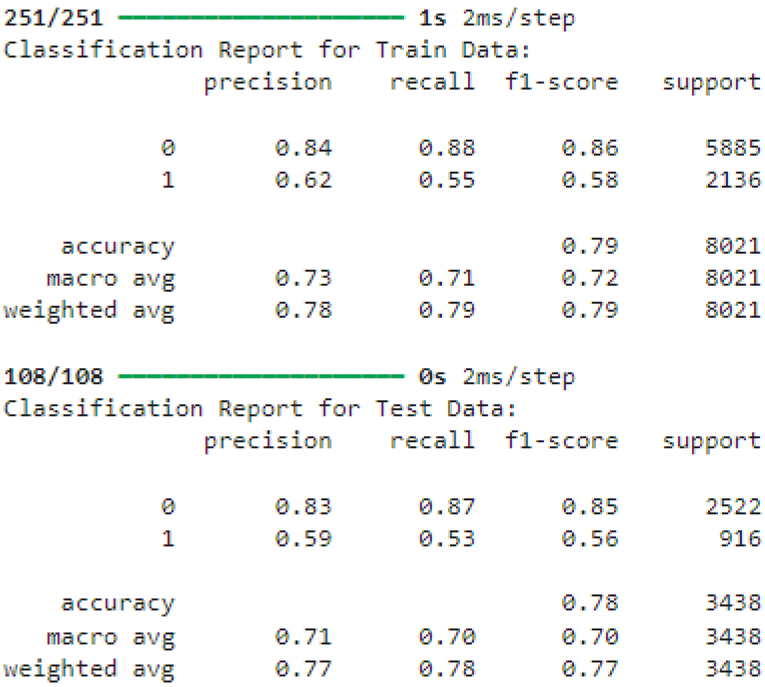


```
# Performance measures for train ad test sets for the model with best parameters

print('Classification Report for train set:\n', get_train_report(xgb_model))


print('Classification Report for test set:\n', get_test_report(xgb_model))
```

```
Classification Report for train set:
               precision    recall  f1-score   support

           0       0.83      0.92      0.87      5885
           1       0.69      0.49      0.57      2136

    accuracy                           0.81      8021
   macro avg       0.76      0.71      0.72      8021
weighted avg       0.79      0.81      0.79      8021

Classification Report for test set:
               precision    recall  f1-score   support

           0       0.82      0.90      0.86      2522
           1       0.63      0.47      0.54       916

    accuracy                           0.79      3438
   macro avg       0.73      0.69      0.70      3438
weighted avg       0.77      0.79      0.77      3438
```

Table 5: Performance measures for train ad test sets for the model with best parameters (XGBoost)

- ANN:



```
251/251 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step
Classification Report for Train Data:
               precision    recall  f1-score   support

           0       0.84      0.88      0.86      5885
           1       0.62      0.55      0.58      2136

    accuracy                           0.79      8021
   macro avg       0.73      0.71      0.72      8021
weighted avg       0.78      0.79      0.79      8021

108/108 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
Classification Report for Test Data:
               precision    recall  f1-score   support

           0       0.83      0.87      0.85      2522
           1       0.59      0.53      0.56       916

    accuracy                           0.78      3438
   macro avg       0.71      0.70      0.70      3438
weighted avg       0.77      0.78      0.77      3438
```

Table 6: Classification report for train and test data with top 9 selected features (ANN model)

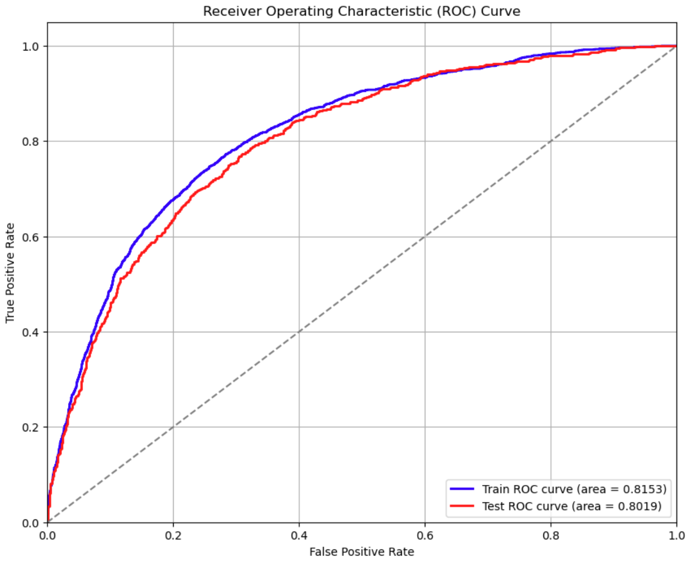


Fig 14: ROC for train and test data with top 9 features (ANN model)

Feature Importance

- Logistic Regression:
  - Significant predictors included Age, Ejection Fraction (EF), Hemoglobin, Glucose, and Urea.
  - Clinical history factors such as diabetes, coronary artery disease, and prior cardiac procedures were important.

- XGBoost:

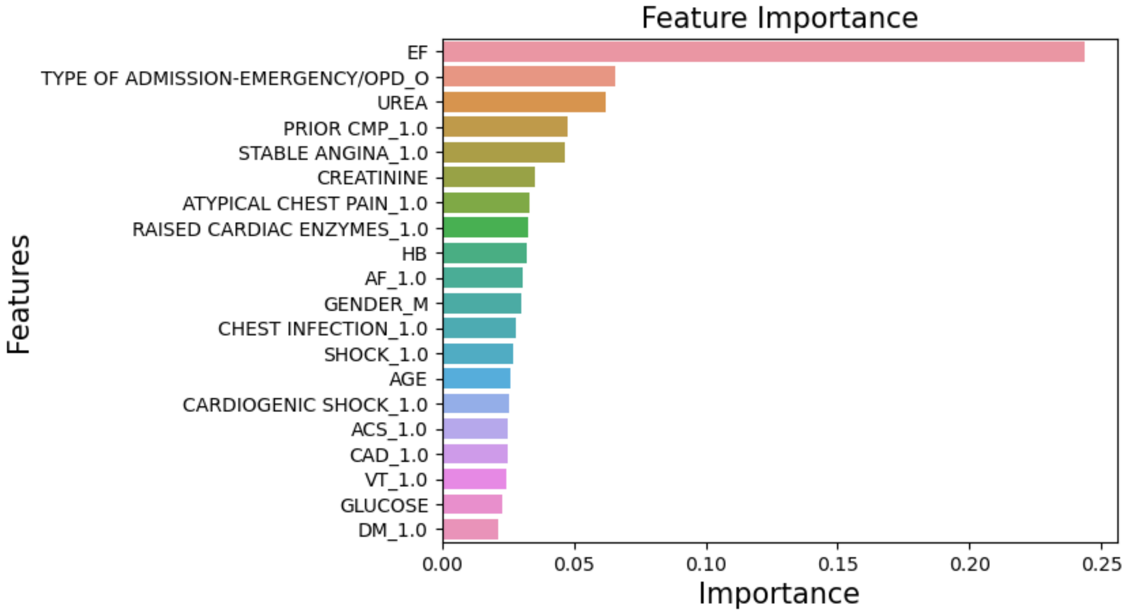


Fig. 15: Feature importance from XGBoost model

- ANN:

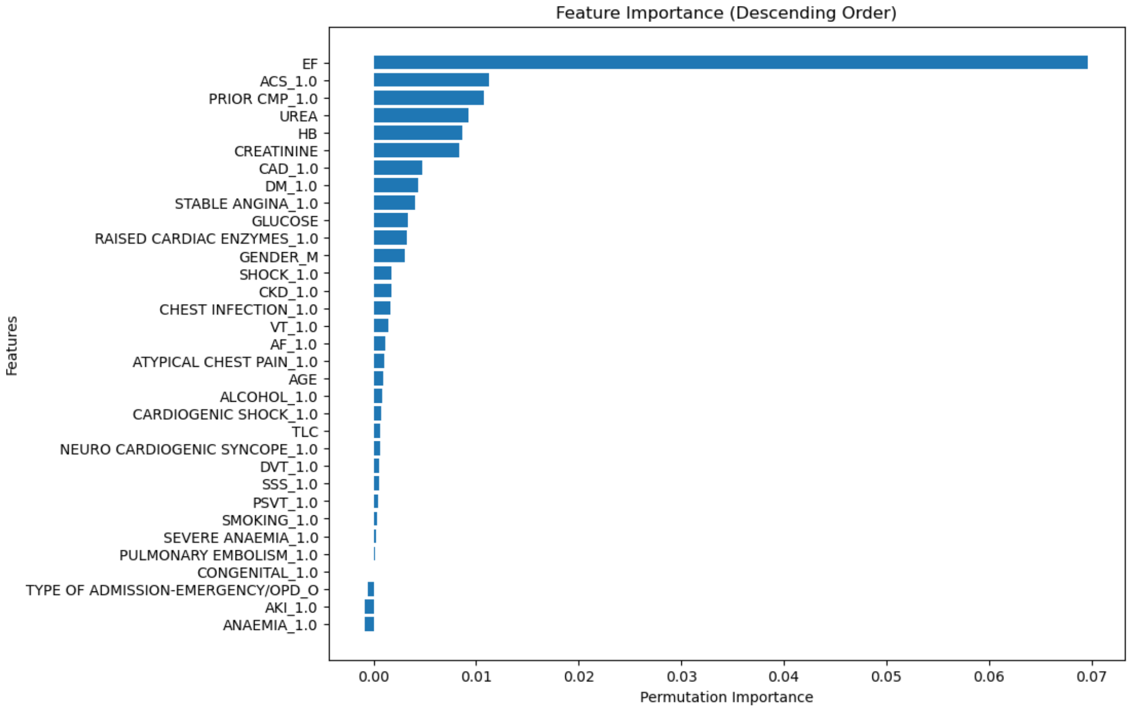


Fig. 16: Feature importance from ANN model (top 9 were selected for further analysis)

11.2. Analysis of the Results

Logistic Regression

- Strengths:
  - Provides a baseline model with good interpretability.
  - Demonstrates reasonable performance in distinguishing between classes with an AUC of 0.79.
- Weaknesses:
  - Struggles with class imbalance, particularly in predicting heart failure cases, as evidenced by low recall for Class 1.
  - Feature selection revealed that several predictors have a significant impact on the outcome, but the model's performance in identifying heart failure cases needs improvement.

XGBoost

- Strengths:
  - High accuracy and good performance on the train set; hyperparameter tuning led to improved model performance.
  - Effective in handling non-linearity and complex interactions between features.
- Weaknesses:
  - Despite tuning, the recall for heart failure cases remains low, indicating that the model may still be struggling with class imbalance.

  - The model’s performance on the test set suggests a need for further optimization.

ANN

- Strengths:
  - Robust performance with high AUC scores, indicating effective prediction ability.
  - Feature selection improved performance slightly, demonstrating the importance of focusing on relevant features.
- Weaknesses:
  - High computational cost and complexity.
  - Recall for heart failure cases still lower than desired, highlighting the need for further refinement.

12. Conclusion

12.1. Summary of the Project Outcome

The project successfully developed and evaluated multiple machine learning models for predicting heart failure and other in-hospital outcomes using a comprehensive dataset. The key models assessed were Logistic Regression, XGBoost, and an Artificial Neural Network (ANN). Each model was evaluated based on accuracy, precision, recall, F1-score, and AUC.

- Logistic Regression: Provided a good baseline model with an AUC of 0.79. However, it faced challenges with class imbalance, particularly in detecting heart failure cases.
- XGBoost: Demonstrated high accuracy and robust performance after hyperparameter tuning. Despite improvements, the model struggled with class imbalance, particularly in recall for heart failure cases.
- ANN: Showed strong performance with high AUC scores and improved recall for heart failure cases after feature selection. The model's complexity required significant computational resources.

The models provided valuable insights into the predictors of heart failure and highlighted areas for further refinement, particularly in handling class imbalance and improving recall for the minority class.

12.2. Future Work

1. Address Class Imbalance:

- Implement techniques such as SMOTE or class weighting to enhance the models' ability to identify heart failure cases. Evaluate the impact of these techniques on model performance.

2. Feature Engineering and Selection:

- Explore additional feature engineering techniques to enhance model performance. Investigate interactions between features and their impact on model predictions.

3. Model Optimization:

- Further tune hyperparameters for XGBoost and ANN models to improve performance. Experiment with advanced algorithms and architectures to enhance predictive accuracy.

4. External Validation:

- Conduct evaluations using external validation datasets to assess the generalizability of the models. This will ensure that the models perform well in diverse real-world scenarios.

5. Model Integration:

- Develop and deploy the models in a real-world clinical setting, integrating them with existing healthcare systems. Ensure that the models provide actionable insights and support clinical decision-making.

6. Continuous Monitoring and Refinement:

- Implement a system for continuous monitoring of model performance and update the models based on real-world feedback and data. This will ensure that the models remain relevant and effective over time.

7. Expansion of Scope:

- Consider expanding the scope of the project to include additional outcomes and predictors. Explore the potential of integrating other data sources to provide a more comprehensive decision support system.

By addressing these areas, the project can continue to advance towards its goal of improving patient outcomes, optimizing healthcare resources, and supporting clinical decision-making through effective predictive modeling.

13. References

Amarasingham, R., Patzer, R. E., Huesch, M., Nguyen, N. Q., & Xie, B. (2014). Implementing electronic health care predictive analytics: Considerations and challenges. *Health Affairs*, *33*(7). https://doi.org/10.1377/hlthaff.2014.0352

Beniczky, S., Karoly, P., Nurse, E., Ryvlin, P., & Cook, M. (2021). Machine learning and wearable devices of the future. *Epilepsia*, *62*(S2). https://doi.org/10.1111/epi.16555

Chishtie, J., Sapiro, N., Wiebe, N., Rabatach, L., Lorenzetti, D., Leung, A. A., … Eastwood, C. A. (2023). Use of Epic Electronic Health Record System for Health Care Research: Scoping Review. *Journal of Medical Internet Research*. https://doi.org/10.2196/51003

Ehwerhemuepha, L., Carlson, K., Moog, R., Bondurant, B., Akridge, C., Moreno, T., … Feaster, W. (2022). Cerner real-world data (CRWD) - A de-identified multicenter electronic health records database. *Data in Brief*, *42*. https://doi.org/10.1016/j.dib.2022.108120

Jahangiry, L., Farhangi, M. A., & Rezaei, F. (2017). Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome. *Journal of Health, Population and Nutrition*, *36*(1). https://doi.org/10.1186/s41043-017-0114-0

Kikuchi, S., Kadama, K., & Sengoku, S. (2021). Characteristics and classification of technology sector companies in digital health for diabetes. *Sustainability (Switzerland)*, *13*(9). https://doi.org/10.3390/su13094839

Mizan, T., & Taghipour, S. (2022). Medical resource allocation planning by integrating machine learning and optimization models. *Artificial Intelligence in Medicine*, *134*. https://doi.org/10.1016/j.artmed.2022.102430

Morgenstern, J. D., Buajitti, E., O’Neill, M., Piggott, T., Goel, V., Fridman, D., … Rosella, L. C. (2020). Predicting population health with machine learning: A scoping review. *BMJ Open*, *10*(10). https://doi.org/10.1136/bmjopen-2020-037860

Ravaut, M., Sadeghi, H., Leung, K. K., Volkovs, M., Kornas, K., Harish, V., … Rosella, L. (2021). Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *Npj Digital Medicine*, *4*(1). https://doi.org/10.1038/s41746-021-00394-8

Ren, Y., Loftus, T. J., Datta, S., Ruppert, M. M., Guan, Z., Miao, S., … Bihorac, A. (2022). Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Predict Postoperative Complications and Report on a Mobile Platform. *JAMA Network Open*. https://doi.org/10.1001/jamanetworkopen.2022.11973

14. Appendices

Appendix: Patient Variables and Medical Conditions

1. Demographic Variables

- **Rural (R) / Urban (U)**: Patient locality categorized as either rural or urban.

2. Admission Type

- **Emergency (E)**: Patients admitted through emergency services.
- **OPD (O)**: Patients admitted through outpatient departments.

3. Lab Parameters

- **HB (Hemoglobin)**: A measure of the amount of hemoglobin in the blood.
- **TLC (Total Leukocyte Count)**: The number of white blood cells (WBC) in the blood.
- **Platelets**: A measure of platelet count in the blood.
- **Glucose**: Blood glucose level, indicative of blood sugar control.
- **Urea**: Blood urea concentration, a marker of kidney function.
- **Creatinine**: Serum creatinine level, a marker of kidney function.
- **BNP (B-Type Natriuretic Peptide)**: A hormone produced by the heart, elevated in heart failure.
- **EF (Ejection Fraction)**: The percentage of blood leaving the heart each time it contracts.

4. Patient History (Comorbidities)

- **DM (Diabetes Mellitus)**: Presence of diabetes.
- **HTN (Hypertension)**: Presence of high blood pressure.
- **CAD (Coronary Artery Disease)**: History of coronary artery disease.
- **PRIOR CMP (Cardiomyopathy)**: History of cardiomyopathy.
- **CKD (Chronic Kidney Disease)**: History of chronic kidney disease.

5. Other Comorbidities

- **Raised Cardiac Enzymes**: Elevation in cardiac enzymes, indicating heart muscle damage.
- **Severe Anemia**: Critically low hemoglobin levels.
- **Anemia**: Lower than normal hemoglobin levels.
- **Stable Angina**: Chronic chest pain or discomfort due to reduced blood flow to the heart.
- **ACS (Acute Coronary Syndrome)**: A range of conditions associated with sudden, reduced blood flow to the heart.
- **STEMI (ST-Elevation Myocardial Infarction)**: A severe heart attack caused by complete blockage of a heart artery.
- **Valvular Heart Disease**: Disorders involving one or more of the heart's valves.
- **Atypical Chest Pain**: Chest pain not typical of heart disease.

- **CHB (Complete Heart Block)**: A condition where the heart’s electrical signals are blocked.
- **SSS (Sick Sinus Syndrome)**: A group of heart rhythm disorders.
- **AKI (Acute Kidney Injury)**: A sudden episode of kidney failure or damage.
- **CVA Infarct (Cerebrovascular Accident)**: A stroke caused by a blockage in the brain.
- **CVA Bleed**: A stroke caused by bleeding in the brain.
- **AF (Atrial Fibrillation)**: An irregular, often rapid heart rate that can cause poor blood flow.
- **VT (Ventricular Tachycardia)**: A fast heart rhythm originating from the heart's ventricles.
- **PSVT (Paroxysmal Supra Ventricular Tachycardia)**: A rapid heart rate starting in the upper part of the heart.
- **Congenital Heart Disease**: Heart defects present at birth.
- **UTI (Urinary Tract Infection)**: Infection in any part of the urinary system.
- **Neurocardiogenic Syncope**: Fainting caused by a sudden drop-in heart rate and blood pressure.
- **Orthostatic Hypotension**: A form of low blood pressure that happens when standing up.
- **Infective Endocarditis**: Infection of the inner lining of the heart chambers and valves.
- **DVT (Deep Venous Thrombosis)**: A blood clot in a deep vein, usually in the legs.
- **Cardiogenic Shock**: A life-threatening condition where the heart cannot pump enough blood to meet the body's needs.
- **Shock**: A critical condition brought on by the sudden drop in blood flow.
- **Pulmonary Embolism**: A blockage in one of the pulmonary arteries in the lungs.
- **Chest Infection**: Infection in the lungs or airways, such as pneumonia.