TDT4259 - Applied Data Science
Individual Report

Martin Hegnum Johannessen

2024/07/18

Contents

Chapter 1

Overview

This document outlines a heart disease prediction system using machine learning. Cardiovascular diseases (CVDs) are the leading cause of death globally. By leveraging a dataset containing 11 clinical features, the goal is to build a system capable of accurately predicting heart disease and enabling timely interventions.

1.1 Intent

The intent of this project is to create a data-driven heart disease prediction system using machine learning models. The system will be integrated into clinical workflows to assist healthcare professionals in making early and informed decisions about at-risk patients.

1.2 Why?

Heart disease is a major health challenge worldwide, causing 17.9 million deaths annually. Early diagnosis significantly reduces mortality, and this system aims to assist in that process. By developing a model that can accurately predict the likelihood of heart disease, the system helps identify high-risk individuals for further medical evaluation, improving patient outcomes and reducing healthcare costs.

1.3 How?

The system will take clinical data, preprocess it, and apply machine learning models to predict the likelihood of heart disease. The system will be cloud-hosted for scalability, with horizontal scaling to meet performance requirements. Security measures such as OAuth 2.0 authentication and GDPR compliance will be implemented. Monitoring tools will also be employed.

Chapter 2

Motivation

Cardiovascular diseases (CVDs) are the leading cause of death globally. The diseases are responsible for 17.9 million deaths annually [1]. Heart failure is a common outcome of CVDs. This outcome pose a significant health risk, especially when left undiagnosed. Early detection is crucial in reducing mortality and improving patient outcomes.

The *Heart Failure Prediction Dataset* [1] provides 11 clinical features to predict heart disease. To apply machine learning for early diagnosis presents an opportunity to create predective models that can identify high-risk individuals. Ultimately, enabling timely interventions. Given the global impact of CVDs, this project aims to create data-driven solutions for more efficient and accurate heart disease prediction.

Chapter 3

Success metrics

The success of the heart failure prediction project can be assessed through a combination of model performance and healthcare-related impact metrics. These metrics evaluate both technical accuracy of the model and its applicability in improving healthcare outcomes.

The metrics used to evaluate the performance of the machine learning model are:

- **Accuracy:** measures the percentage of correctly predicted cases out of all predictions made. High accuracy is desired.
- **Precision:** measures the ratio of correctly predicted positive cases to the total predicted positives. High precision reduces false positives.
- **Recall (Sensitivity):** captures the model's ability to correctly identify true positive cases.
- **F1 Score:** balances precision and recall, providing a more holistic view of model performance.[1]
- **AUC-ROC:** measures the trade-off between true positive and false positive rates. A higher AUC indicates a model that can better distinguish between heart disease and non-heart disease cases [2].

Beyond statistical performance, the success of the model must also be measured in terms of its healthcare impact:

- **Early Detection Rate:** The effectiveness of the model in identifying heart disease cases earlier than traditional diagnostic methods.
- **Reduction in Hospital Readmissions:** Whether the model can help reduce heart disease-related hospital readmissions by enabling timely interventions.
- **Improved Patient Outcomes:** Ultimately, the model's success is measured by improved patient outcomes.

[1] Especially for imbalanced datasets.

Chapter 4

Requirements & Constraints

This chapter outlines the functional requirements and constraints necessary to implement a heart failure prediction system. In addition, the scope of the system is defined.

4.1 Functional Requirements

- **Accurate Prediction of Heart Disease:** The system must be capable of predicting the likelihood of heart disease using the 11 clinical features provided in the dataset (e.g., age, cholesterol, blood pressure). The goal is to develop a model that can correctly identify patients at risk of heart failure.
- **Minimization of False Negatives:** The model should prioritize high recall (sensitivity) to ensure that as many true heart disease cases as possible are detected.
- **User-Friendly Interface for Clinicians:** The system should present predictions in a simple and interpretable way for healthcare professionals.
- **Integration with Existing Medical Systems:** The heart failure prediction system must be compatible with standard Electronic Health Record (EHR) systems [3].

4.2 Constraints

- **Accuracy and Performance:** The model must achieve a minimum accuracy of 85% and an F1 score of at least 0.80 to be considered reliable for clinical use.
- **Data Privacy and Security:** All data used must comply with data protection regulations such as GDPR.
- **Latency:** The system should provide predictions within 1 second of input data being processed.

- **Scalability:** The system must be able to handle a throughput of up to 10,000 predictions per day.[1]
- **Model Maintenance:** The predictive model must be retrained every six months to account for any changes in patient demographics, medical practices, and updated clinical guidelines.
- **Cost Limitations:** The implementation cost should not exceed 15 million NOK.

4.3 What's In-Scope & Out-of-Scope

The scope of this system focuses on predicting the risk of heart disease using the given dataset's 11 clinical features. This includes building and deploying a machine learning model to assist in early detection of heart disease. The project will not cover other cardiovascular conditions not included in the dataset or broader healthcare issues.

4.4 Assumptions

It is assumed that the healthcare providers have sufficient data quality and availability to train the machine learning model. The dataset used is assumed to be representative of the patient population for which the system will be used. It is also assumed that necessary IT infrastructure, such as EHR integration, is already in place in the healthcare institutions where the system will be deployed. Additionally, the system will rely on clinical staff to interpret and act on the model's predictions rather than making autonomous treatment decisions.

[1] A rough estimate of usage in medium- to large-scale hospitals.

Chapter 5

Methodology

5.1 Problem Statement

The problem at hand is predicting the likelihood of heart disease in patients based on various clinical features. This problem can be framed as a supervised classification task, where the objective is to categorize patients into two classes: those with heart disease (labeled as 1) and those without heart disease (labeled as 0). The dataset contains clinical data that can be used to train machine learning models to make these predictions. Given that heart disease is a binary outcome, this is a binary classification problem. The model aims to minimize false negatives as early detection is important within this setting.

5.2 Data

The dataset contains 918 observations with 12 attributes. Table **??** presents the features that are used as input for the machine learning model.

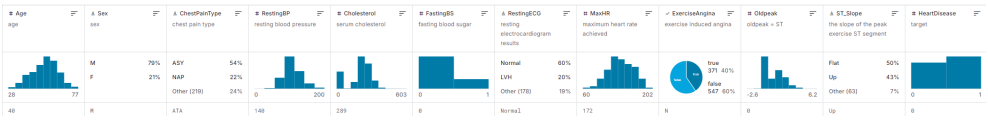

Figure 5.1: Data features for heart disease prediction

5.3 Techniques

Several machine learning models will be used, including Logistic Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks. Logistic Regression provides easy interpretation, while Random Forest handles non-linear relationships. SVM captures complex decision boundaries. Neural Networks will help prevent overfitting, given the dataset size.

Data preparation is essential for optimal model performance. Missing or incorrect entries will be addressed using imputation. Continuous variables, like *Cholesterol* and *RestingBP*, will be filled with mean or median values. Categorical variables, such as *RestingECG*, will be imputed using mode. Invalid data points (e.g., extreme values in *MaxHR* or *Oldpeak*) will be corrected or removed. Outliers will be managed using box plots and statistical techniques to prevent model bias.

Once cleaned, continuous features will be normalized using z-score scaling. This ensures that models like Logistic Regression, SVM, and Neural Networks perform effectively. Tree-based models, like Random Forest, will remain unaffected by scaling. Feature engineering will introduce new variables, such as interactions between *Oldpeak* and *ST_Slope*, to improve accuracy.

Categorical variables like *ChestPainType*, *RestingECG*, and *ST_Slope* will be one-hot encoded for compatibility with most models. For Random Forest, they will be used as-is. Undersampling will address class imbalance in the target variable.

Hyperparameter tuning will use grid or random search. For Logistic Regression, regularization will be adjusted. Random Forest will be optimized for the number of trees and depth. SVM will be fine-tuned for kernel and regularization. Neural Networks will be tuned for hidden layers, learning rate, and dropout. Cross-validation will ensure model generalization.

5.4 Experimentation & Validation

To validate our approach, we will split the dataset into training and testing sets using an 80-20 split. The cross-validation procedure and evaluation metrics are mentioned in the previous sections.

In a real-world deployment, the heart failure prediction model could undergo A/B testing in clinical settings. Patients could be randomly assigned to either the treatment group or the control group.

5.5 Human-in-the-loop

Since the model will be used in healthcare, a human-in-the-loop approach is essential. The approach ensures that predictions are both accurate and actionable. Medical professionals will review the predictions before taking any action. Their domain expertise will be integrated into the decision-making process. Alerts and thresholds can be adjusted based on clinical guidelines. Any edge cases or uncertain predictions can be flagged for manual review. This way, the model supports human decision-making rather than replacing it in critical heart disease cases [4].

Chapter 6

Implementation

6.1 High-level design

The system follows a straightforward flow for heart disease prediction. Data is ingested from clinical databases, preprocessed (e.g., normalization, imputation), and passed to the machine learning model. The model outputs predictions, which are stored in a database. Clinicians access the predictions through a user interface, allowing for manual review and intervention. Figure 6.1 visualizes this process.

6.2 Infrastructure

The system will be hosted in the cloud for scalability, reliability, and ease of management. Cloud platforms such as AWS or Azure will provide necessary services like compute instances, storage, and security management. A cloud-based approach enables the system to easily scale up or down depending on demand and ensures high availability.

6.3 Performance

To meet the system's performance and throughput requirements, horizontal scaling will be used. This ensures that the system can handle increased loads by adding more computing resources. Autoscaling features in cloud platforms like AWS EC2 will help automatically adjust the number of servers based on real-time traffic. Latency will be minimized by geographically distributing resources to reduce response time.

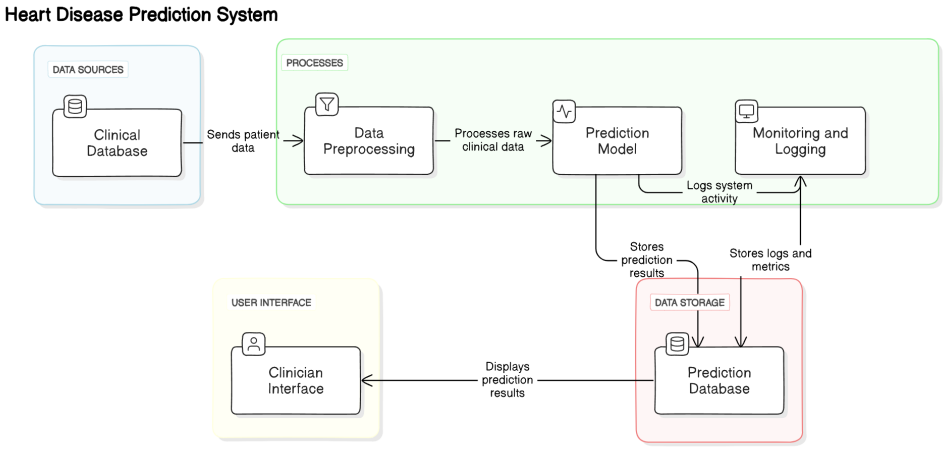

Figure 6.1: Data flow diagram

6.4 Security

The system will implement user authentication through OAuth 2.0. Additionally, role-based access control (RBAC) will be enforced to limit access based on user roles (e.g., clinicians, admins). If the system is publicly accessible, it will be secured behind a firewall, with inbound traffic monitored and controlled. Data in transit will be encrypted using HTTPS, and encryption at rest will be applied for sensitive patient data.

6.5 Data Privacy

The system will be designed to comply with GDPR and other data privacy regulations. Sensitive patient data will be anonymized where possible. All personally identifiable information (PII) will be encrypted both in transit and at rest. The system will also follow data retention policies.

6.6 Monitoring & Alarms

The system will employ logging and monitoring solutions such as AWS CloudWatch or Azure Monitor to track metrics like response times, system uptime, and error rates. Alerts will be set up for any metric breaches, such as high latency, high error rates, or resource utilization beyond thresholds.

6.7 Cost

The estimated monthly cost for operating the system is projected based on cloud resources. For instance, compute resources (EC2 or VMs) may cost between $500 and $1000 monthly, depending on usage. Additional costs will include storage, monitoring tools, and data transfer fees.

6.8 Integration Points

The system will integrate with existing clinical databases for data ingestion. The output will be consumed by downstream systems. APIs will be provided for integration with external applications.

6.9 Risks & Uncertainties

There are several risk and uncertainties. Variability in patient data quality could impact model accuracy. Uncertainties around scaling the system under high-load scenarios or during peak usage times are also present. Data privacy is another concern, especially with GDPR compliance, and any data breach could have legal consequences. All risks and uncertainties are presented in Table 6.1.

Risk	**Comment**	**Severity**
Data quality	Could lead to inaccurate predictions.	High
Scaling under load	System may struggle during peak use.	Moderate
GDPR breach	Non-compliance could result in fines.	High
Integration failure	May disrupt hospital workflows.	Moderate
Security breach	Patient data could be exposed.	High
Model drift	Accuracy may decrease without retraining.	Moderate

Table 6.1: Risk assessment

Bibliography

[1] F. S. Palacios, *Heart failure prediction dataset*, https://www.kaggle.com/fedesoriano/heart-failure-prediction, Accessed: [13.09.2024], Sep. 2021. [Online]. Available: https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[2] A. Rajkomar, J. Dean and I. Kohane, 'Machine learning in medicine,' *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[3] A. K. Jha, D. Doolan, D. Grandt, T. Scott and D. W. Bates, *The use of health information technology in seven nations*. Elsevier, 2010, vol. 79, pp. 848–854.

[4] E. J. Topol, 'High-performance medicine: The convergence of human and artificial intelligence,' *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.