A Heart Disease Risk Prediction Model

Hangying Xie

Table of Content

1. Project Description

1.1 Project Background

Cardiovascular diseases (CVDs), which include coronary heart disease, cerebrovascular disease, and rheumatic heart disease, remain the leading cause of death worldwide. According to the World Health Organization (WHO), an estimated 17.9 million people die from CVDs each year, accounting for 31% of all global deaths.

However, early diagnosis of CVDs can significantly reduce the likelihood of premature death by timely lifestyle changes and proper treatments. Current challenges in cardiovascular health management include:

- Delayed detection: Late discovery of heart issuesmakes treatments less effective.
- Imprecise risk assessments: Current methods often fail to consider individual characteristics, leading to less effective prevention.
- High treatment costs: Expensive treatments result in incomplete or delayed care, especially in low-income areas.

These issues highlight the urgent need for innovative approaches to early prediction in cardiovascular health management.

1.2 Project Proposal

This project aims to develop a heart disease risk prediction model using wearable devices to continuously monitor and analyze real-time health data, such as heart rate and step frequency. By integrating this data with electronic health records and regular check-up results, the model will provide personalized risk assessments and early warnings of potential heart problems. This model is designed to benefit both insurance companies and healthcare providers.

1.3 Project Roles

- **Data engineers**
    - Set up and maintain secure, scalable data infrastructure for real-time data collection, storage, and processing.
    - Develop data pipelines for real-time processing and quality assurance, ensuring a steady flow of high-quality data for data analysts and data scientists.
    - Set up stream processing systems for real-time data analysis.
    - Deploy the finalized model from data scientists in production environment for real-time risk assessments.
    - Develop monitoring and alerting systems, and plan for scalability.
- **Data analysts**

    - Analyze heart health data from multiple sources (wearable device data, electronic health records, and check-up results).
    - Examine data provided by data engineers, identifying patterns and trends, which they communicate to both data scientists and stakeholders.
    - Use statistical tools and visualization techniques to identify key risk factors and trends.
    - Create intuitive dashboards based on the model's outputs, showing risk distributions and provide regular insight reports to insurance companies and healthcare providers.
    - Interpret model outputs for non-technical stakeholders to aid decision-making in patient care and risk management.
- **Data scientists**
    - Use insights provided by data analysts to develop an accurate heart disease risk prediction model.
    - Create a risk scoring system to translate the model outputs into recommendations, such as “Low Risk (0-20%): Continue current lifestyle, annual check-ups recommended.".
    - Design and implement machine learning algorithms by using multiple-source data to make prediction.
    - Work closely with data analysts to validate results and optimize the model.
    - Conduct feature engineering and ensure accuracy for diverse populations through continuous model optimization and validation.
    - Collaborate with medical professionals to ensure the model’s clinical relevance and real-world applicability.

2. Business Model

This innovative heart disease risk prediction model provides data-driven insights that deliver tangible business value for both insurance companies and healthcare providers. Unlike existing devices such as the Apple Watch that focus on basic health metrics, our model employs a sophisticated multi-source data integration approach. It combines real-time data from wearable devices with historical electronic health records and periodic check-up results, offering a more comprehensive and dynamic view of an individual's health status. This integrated approach, powered by advanced analytics, enables early identification of high-risk individuals and facilitates preventive care strategies. As a result, it can optimize resource allocation, potentially lowering healthcare costs while significantly improving patient outcomes.

2.1 Benefits

- **For health insurance companies**
    - Increased revenue through precise customer segmentation and targeted marketing of personalized insurance products.
    - Reduced long-term claim costs through preventive measures and early intervention.

    - Improved risk management strategies and optimized premium pricing.
    - Improved customer satisfaction by offering customized insurance products and services.
- **For medical professionals**
    - Improved efficiency by reducing time spent on data analysis.
    - Better patient stratification for more targeted care management.
    - Faster and more accurate identification of potential issues, enabling timely and appropriate intervention advice.
- **For individuals**
    - Reduced screening costs and earlier detection of potential issues, enabling timely treatment and lowering future expenses
    - Decreased death rates and risk of complications due to early intervention.
    - Increased health awareness and motivation to adopt positive lifestyle changes, supported by personalized health management recommendations based on individual risk profiles.

2.2 Challenges

- **Model Reliability and Interpretability**
  The model differs from traditional diagnostic methods that rely on various medical examinations. Ensuring the model's reliability and interpretability requires close collaboration with medical experts and rigorous testing and validation to build trust among healthcare providers and patients.
- **Acceptance by the Medical Community**
  Reducing concerns and resistance from medical professionals is critical. The model is designed to aid medical decision-making, not replace professionals. Open communication and demonstrations of the model's effectiveness can help gain acceptance.
- **Data Accuracy and Completeness**
  The model's success depends on accurate and complete data. Ensuring high-quality data inputs from wearable devices, electronic health records, and regular check-ups is essential.
- **Data Privacy and Compliance**
  Protecting patient data privacy is paramount. Strategies to ensure compliance with relevant data protection regulations (such as HIPAA and GDPR) must be in place to safeguard sensitive information.

3. Characterising and Analysing Data

3.1 Data characteristics, processing and storage

Properties	Justification
Data Source	● Continuous health data (e.g. heart rate, blood pressure, step frequency, stress levels etc.) from wearable devices ● Electronic Health Records (EHRs) from hospitals and clinics ● Regular health check-up results and lab tests ● Patient self-reported data (e.g. lifestyle, diet)
Data type	● Primarily include numerical, textual, and time-series. Formats vary (.csv, .json, .xml, .pdf, etc.), covering: ○ Structured data: vital signs, lab results, patient demographics. ○ Semi-structured data: wearable device outputs, ECG readings. ○ Unstructured data: clinical notes, patient self-reports.
Data Characteristics	**<u>Volume</u>** ● Wearable device data: Approximately 250MB per user per day ● EHR data: Average of 1.4GB per patient for a comprehensive record ● Assuming 1 million users, potential daily data volume could reach 250TB. ● Annual data volume could exceed 90PB, including all data sources.
	<u>Velocity</u> ● Wearable devices: Continuous stream, up to 250 data points per second per device ● EHR updates: Sporadic, with frequency varying by healthcare provider and patient visit patterns ● Lab results and check-ups: Batch updates, frequency depends on individual health management plans
	<u>Variety</u> ● Structured data: Vital signs (e.g. heart rate, blood pressure), lab results (e.g. cholesterol, glucose levels) ● Semi-structured data: JSON/XML outputs from wearable devices, ECG readings ● Unstructured data: Physician notes, patient-reported symptoms, lifestyle information ● Various formats including numeric data, text, and potentially medical imaging data (e.g. ECG graphs)
	<u>Variability</u> ● Data quality and frequency may vary between different wearable device brands ● Inconsistent use of medical terms across different healthcare providers ● Seasonal variations in health metrics and behaviors ● Changes in data collection methods or wearable technology over time

Properties	Justification
	<u>Veracity</u> Possible issues include: ● Varying accuracy of consumer-grade wearable devices (5-10% error rate in heart rate measurements) ● Potential errors in EHR data entry (error rates ranging from 5.9% to 17.3%) ● Subjective nature of patient-reported data ● Inconsistencies in medical coding practices across different healthcare providers
	<u>Visualisation</u> It will be extensively used in data exploration, pattern recognition, and for presenting risk assessments to healthcare providers and patients. It's crucial for interpreting complex health data.
Data Storage	● Time-series data (wearables): Distributed NoSQL database (e.g., Apache Cassandra) ● Structured medical data (EHR): Relational database (e.g., PostgreSQL) ● Unstructured data (clinical notes): Document-oriented database (e.g., MongoDB) ● Large-scale storage: Hadoop Distributed File System (HDFS)
Data Processing	● Real-time processing: Apache Spark Streaming for wearable data ● Batch processing: Apache Hadoop MapReduce or Spark for EHR and lab results ● Complex event processing: Apache Flink or Spark for risk pattern detection ● Machine learning: Apache Spark MLlib for model training and updates ● Interactive analysis: Apache Zeppelin or Jupyter Notebooks for data exploration and visualization

3.2 Data Analysis and Statistical Methods

This heart disease risk prediction model utilizes supervised machine learning to classify individuals into risk categories (low, medium, high) based on their likelihood of developing heart disease.

- **Multinomial Logistic Regression**
  Multinomial logistic regression is introduced as a simple yet powerful method for predicting outcome probabilities across multiple categories. Its specific characteristics are as follows:
    - Mathematical logic: Utilizes the logistic function to predict the probabilities of each outcome class.
    - Trainable parameters: Feature weights and intercepts for each class.
    - Key hyperparameters: Regularization strength and type.

This method provides interpretable weights for risk factors, making it suitable for multi-class classification problems (e.g., categorizing heart disease risk as low, medium, or high) while remaining computationally efficient.

- **Random Forest**
  To address potential non-linear relationships and feature interactions, the project introduces the random forest method. The characteristics of random forest include:
  - Mathematical logic: Ensemble of decision trees, uses majority voting
  - Trainable parameters: Tree structure and node splitting rules
  - Key hyperparameters: Number of trees, tree depth, feature sampling ratio

  Random forest can handle non-linear relationships and feature interactions, provide feature importance rankings, and reduce the risk of overfitting. The output of random forest includes heart disease risk prediction probabilities, feature importance rankings, and insights into complex feature relationships.

- **XGBoost**
  XGBoost (eXtreme Gradient Boosting) is adopted as an advanced method. The features of XGBoost are as follows:
  - Mathematical logic: Gradient boosting decision trees, iteratively optimizes loss function
  - Trainable parameters: Tree structure and leaf weights
  - Key hyperparameters: Learning rate, number and depth of trees, regularization parameters

  XGBoost improves prediction accuracy by iteratively optimizing the loss function. It can handle various types of features and missing data, and has built-in regularization to reduce overfitting. The output of XGBoost includes high-precision heart disease risk predictions, capture of complex feature interactions, and model interpretation through SHAP values.

  Through statistical analysis and machine learning techniques, the model aims to deliver:
  - Risk Classification
    - Categorize individuals into risk levels (low, medium, high)
    - Provide probability estimates for each category
  - Comprehensive Evaluation:
    - Combine multiple model outputs for robust risk assessment
    - Integrate diverse data sources: wearables, health records, and check-ups

3.3 Demonstration

To validate our heart disease risk prediction model, I analyzed the UCI Heart Disease Dataset, containing data from 303 patients with 14 attributes, and a target variable indicating heart disease risk level (num). This dataset aligns with our model's approach of integrating multiple data sources, including wearable devices and electronic health records. Key indicators like heart rate and blood pressure, which can be continuously monitored by wearables, are included,

supporting our model's real-time monitoring and predictive capabilities. This analysis provides insights to guide and validate our innovative approach to heart disease risk assessment.

- **Data Preparation and Exploration**
  - The dataset was cleaned and preprocessed, handling missing values and converting categorical variables to appropriate formats.
  - Exploratory data analysis revealed key insights into the relationships between various health indicators and heart disease risk.
- **Model Development**
  - A multinomial logistic regression model was developed to predict heart disease risk levels.
  - The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance.
- **Model Performance**
  - The model achieved an overall accuracy of 60.7% on the test set.
  - Performance varied across different heart disease severity levels, with better prediction for the absence of heart disease.
- **Interpretation of Results**
  The model's moderate accuracy (60.7%) can be attributed to several factors:
  - Limited dataset size (303 patients) may not fully represent the diverse spectrum of heart disease cases.
  - Complex nature of heart disease risk, influenced by numerous factors not all captured in this dataset.
- **Despite these challenges, this demonstration provides valuable insights:**
  - It confirms the feasibility of developing a heart disease risk prediction model using multiple health indicators.
  - The results highlight the complexity of heart disease risk prediction and the need for comprehensive data.
  - It establishes a baseline for further model refinement and improvement.

For detailed analysis and visualizations, please refer to the accompanying R Markdown (Rmd.) file.

4. Data Governance and Management

Our heart disease risk prediction model adheres to the CRISP-DM standard, ensuring a structured approach throughout the data lifecycle, from collection to archival. For data governance and management, we implement robust practices addressing accessibility, security, and confidentiality. This includes role-based access control, data encryption, and anonymization techniques, aligned with the Capability Maturity Model for scientific data management.

We strictly comply with HIPAA and GDPR regulations, and address ethical concerns through informed consent, fairness in predictions, and transparent data usage policies. Regular audits, data quality assessments, and comprehensive data management plans (DMPs) are in place, following best practices from organizations like the Australian National Data Service (ANDS).

Our approach engages key stakeholders, including data scientists, IT developers, and end-users, in decision-making processes. We also implement robust data backup strategies and version control systems to ensure data integrity and traceability.

These measures collectively ensure the integrity of our data processes, protect patient privacy, and maintain ethical standards while delivering valuable health insights through our innovative risk prediction model. By adhering to these principles, we aim to avoid common pitfalls in data management, such as those seen in high-profile cases like the Robodebt scheme or the British COVID-19 data mishandling incident.

References

Sony, M. R. K. (2020). *Heart disease data* [Data set]. Kaggle.

https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

World Health Organization. (n.d.). *Cardiovascular diseases.* World Health Organization.

https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1