Sri Lanka Institute of Information Technology

Statement of Work document

Fundamentals of Data Mining – IT3051

Mini Project

Group-16

Predicting Risk Factors for Cardiovascular Heart Disease

IT21924750	Senevirathna W.P.U.
IT21468360	Rizvi F. A
IT21276200	Samarakoon S.M.R.S.B
IT21329760	Nanayakkara N.W.B.S

Table of Contents

Background 3

Scope of work 4

1. User Interface Layer 4

2. Data wrangling and data cleansing layer 4

3. Data Mining Layer 4

4. Model building and Analysis layer 5

5. Data Visualizing Layer 5

Activities 5

1. User Interface Layer 5

2. Data wrangling and data cleansing layer 5

3. Data Mining Layer 5

4. Model building and Analysis layer 6

5. Data Visualizing Layer 6

Approach 6

1. User Interface Layer 6

2. Data wrangling and data cleansing layer 6

3. Data Mining Layer 7

4. Model building and Analysis layer 7

5. Data Visualizing Layer 7

Deliverables 7

1. User Interface Layer 7

2. Data wrangling and data cleansing layer 7

3. Data Mining Layer 8

4. Model building and Analysis layer 8

5. Data Visualizing Layer 8

Project Plan & Timeline 8

Assumptions 9

Project team, roles, and responsibilities 9

Background

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions.

Individuals need to be aware of their risk factors and take steps to prevent and manage cardiovascular diseases to improve their overall health and longevity. Regular check-ups with healthcare providers can also help in the early detection and management of CVD risk factors.

To address this vital problem, our data mining project intends to use advanced analytics and machine learning to predict the risk of cardiovascular diseases by uncovering the hidden patterns within a diverse dataset.

The dataset provides extensive information regarding the multitude of elements that influence the likelihood of cardiovascular disease. It incorporates in-depth profiles of more than 70,000 individuals, encompassing details such as their age, gender, height, weight, blood pressure measurements, cholesterol levels, blood glucose levels, smoking habits, alcohol consumption, physical activity levels, and whether they have been diagnosed with cardiovascular diseases.

We intend to develop an accurate and reliable approach for predicting the risk of cardiovascular disease that can be used both by healthcare professionals and individuals concerned about their heart condition. It will not only help with predicting risks but will also provide significant insight into the root causes that cause CVDs.

Scope of work

Our project consists of 5 layers.

1. User Interface Layer
2. Data wrangling and data cleansing layer
3. Data Mining Layer
4. Model building and Analysis layer
5. Data Visualizing layer

1. User Interface Layer

Objective: To create a user-friendly environment, enhancing user interaction with the backend analytics.

The User Interface Layer constitutes the system's front end, serving as the interface through which users interact. It facilitates data selection and input and provides users with the analytics they need.

2. Data wrangling and data cleansing layer

Objective: Transform the raw data into a more suitable and valuable format for downstream analytical purposes.

In the Data Wrangling and Data Cleansing Layer, the focus is on cleaning and preprocessing data. This phase involves identifying and rectifying corrupt or inaccurate records while also pinpointing incomplete, inaccurate, or irrelevant data segments.

3. Data Mining Layer

Objective: Apply data mining techniques to the prepared data to uncover useful insights and patterns.

The Data Mining Layer involves the analysis of datasets using algorithms to extract valuable numeric information. Its main function is to unearth insights from the data and convert this information into a structured format that is easily comprehensible, paving the way for further analysis.

4. Model building and Analysis layer

Objective: Develop predictive models and conduct in-depth analysis to aid decision-making.

This is the process of modeling data to predict whether or not a person has been diagnosed with cardiovascular disease. This layer creates predictive models that use the chosen dataset to predict desired outcomes from new data**.**

5. Data Visualizing Layer

Objective: Present the analyzed data and model outputs in graphical form for user comprehension.

The Data Visualizing Layer focuses on graphically representing the final analyzed results of characteristic data in an appealing and easily understandable manner for users. It utilizes appropriate graphs and a user-friendly interface to accomplish this task.

Activities

1. User Interface Layer

- Design and develop the user interfaces (UI) for data entry and selection.
- Implement interactive features to enhance user-friendliness.

2. Data wrangling and data cleansing layer

- Analyze and handle missing data, outliers, and inaccuracies.
- Standardize data formats and address inconsistencies in data.
- Perform data transformations as needed.

3. Data Mining Layer

- Choose relevant data mining algorithms (clustering, classification, and association).
- Analyse the dataset using data mining techniques.
- Extract and prepare useful data for further study.

4. Model building and Analysis layer

- Implement predictive models with the dataset and the selected methods.
- Use appropriate evaluation metrics to evaluate model performance.
- Assist in decision-making based on model results.

5. Data Visualizing Layer

- Create data visualizations (charts, graphs, dashboards) to display the results of the analysis.
- Integrate the visualizations into the frontend user interface for easy access.

Approach

The Dataset which we selected to build a model for our problem,

https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

1. User Interface Layer

- Design and Create wireframes and prototypes to visualize the UI, using tools such as Figma.
- Implement the UIs using HTML, CSS, and JS.

2. Data wrangling and data cleansing layer

- Identify missing values, outliers, and data quality issues by profiling the raw data.
- Execute data cleaning operations such as imputing missing data, dealing with outliers, and addressing data discrepancies.
- Use data transformation methods such as Normalization, Smoothing, and Discretization.
- Validating that the data that has been cleaned and preprocessed fulfills quality criteria.
- The preprocessed data is divided into training data and testing data.

3. Data Mining Layer

- Techniques willing to be used – Classification.
- Algorithms willing to be used – Logistic Regression (One of the best algorithms for binary classification)
- Use selected techniques and algorithms to extract insights and patterns from prepared data.

4. Model building and Analysis layer

- Languages willing to be used: Python.
- Model Development: Using the training dataset, create prediction models.
- Model Evaluation: Evaluate the model using the testing data set. Use applicable measures to assess model performance (e.g., accuracy, precision, recall).

5. Data Visualizing Layer

- Data Visualization: Design and build data visualisations (e.g., charts, graphs, dashboards) to display analysed results.
- User Integration: Integrate visualizations into the frontend UI for user access.
- Test the usability of visualizations.

Deliverables

1. User Interface Layer

- User-Friendly User Interface
- Usability Testing Reports

2. Data wrangling and data cleansing layer

- Cleaned and Preprocessed dataset

3. Data Mining Layer

- Data mining results and patterns.

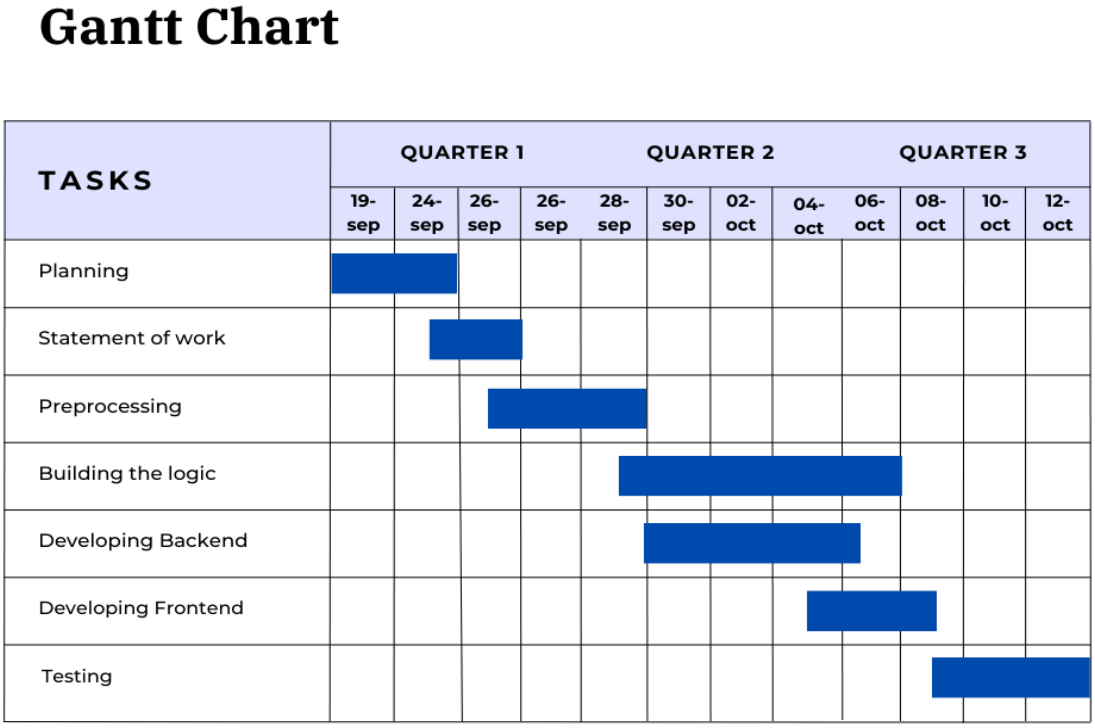
4. Model building and Analysis layer

- Predictive Models
- Model Evaluation Reports
- Analysis Findings

5. Data Visualizing Layer

- Data visualization
- Completed Project
- Final Documentation

Project Plan & Timeline



Assumptions

- All the source data is accurate, accessible, and available for analysis.
- Other than the features contained in the dataset, none of the data has been altered by any external factors.
- Outliers may have an impact on normal data. As a result, they are removed during the data preprocessing stage.

Project team, roles, and responsibilities

	IT Number	Name	Roles	Responsibilities
1	IT21924750	Senevirathna W.P.U.	Solution developer / Solution tester	• Implement the model. • Documentation • Integrate • Data analysing • Testing Data
2	IT21468360	Rizvi F. A	Solution developer / Solution tester	• Implement the model. • Documentation • Data analysing • Integrate • Testing Data
3	IT21276200	Samarakoon S.M.R.S.B	Solution developer / Solution tester	• Implement the model. • Develop Front End • Data visualization • Integrate • Testing Data
4	IT21329760	Nanayakkara N.W.B.S	Solution developer / Solution tester	• Implement the model. • Data visualization • Integrate • Testing Data