AIT-664-DL3– Represent, Process & Visualize
Applied Information Technology

RESEARCH PROJECT PROPOSAL

UNVEILING CARDIOVASCULAR HEALTH PATTERNS USING STATISTICAL AND PREDICTIVE ANALYTICS

GROUP- 1

ABHISHEK ANUMALLA
AAKASH BOENAL
PAVAN TEJAVATH
SHASHANK YELAGANDULA

Prof. EBRIMA CEESAY

Introduction to the problem:

Heart disease is a significant health concern impacting individuals and communities worldwide. In fact, heart-related fatalities stand as the primary cause of mortality, with cancer-related deaths closely trailing behind. To effectively address this issue, it is crucial to understand the factors contributing to heart disease and its various manifestations. This research aims to analyze a dataset comprising demographic and clinical variables related to cardiovascular health, including age, sex, chest pain type (ATA, NAP, ASY), resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiogram (ECG), maximum heart rate, exercise-induced angina, oldpeak and ST slope. By unravelling patterns and relationships within this dataset, we seek to gain insights into the risk factors and predictors of heart disease. Specifically, we aim to identify associations between demographic, clinical, and ECG variables and the presence or absence of heart disease. Additionally, we aim to explore the predictive power of these variables in assessing heart disease risk. Ultimately, this research endeavors to contribute to the development of strategies for early detection, prevention, and management of heart disease, ultimately improving cardiovascular health outcomes for individuals and communities.

Why the problem is important:

Understanding the significance of the problem is crucial as heart disease stands as a pervasive threat to public health, claiming countless lives globally each year. It not only ranks as the leading cause of mortality but also imposes a substantial burden on individuals, families, and healthcare systems worldwide. It affects people's quality of life, productivity, and longevity. By understanding the factors contributing to heart disease and implementing targeted interventions, we can potentially reduce its prevalence and mitigate its impact on individuals and society. Moreover, promoting cardiovascular health aligns with broader public health goals of improving overall well-being and reducing healthcare disparities.

Literature Review:

Prediction of heart disease at early stage using data mining and big data analytics
The literature review in the paper focuses on the utilization of data mining (DM) models and techniques for forecasting heart disease (HD) based on patient datasets. It emphasizes how crucial data mining is for drawing insightful conclusions from massive volumes of medical data, assisting in the early diagnosis and prevention of heart disease. The review covers a range of DM approaches, including support vector machines, neural networks, naïve bayes, decision trees, genetic algorithms, K-NN, and clustering algorithms. These methods are used to create prediction models for heart disease with the goal of enhancing patient outcomes and diagnostic precision. The review also highlights the importance of big data analytics in managing extensive and intricate medical datasets, making it easier to extract useful information for the prediction of heart disease. Overall, the review of the literature highlights how DM and big data analytics can be combined to improve predictive modeling and develop strategies for managing heart disease.

Analysis of heart disease using statistical techniques
This paper centers on heart disease, emphasizing its diverse manifestations and related risk factors, including age, gender, obesity, smoking history, and particular symptoms like dyspnea and chest pain. Binary logistic regression is used in the research methodology to analyze data and determine associations between these risk factors and the chance of developing heart disease. The findings show that depression, obesity, smoking history, and chest pain are important indicators of heart disease. Understanding the connection between these variables and the presence of cardiovascular disease is possible thanks to the logistic regression model. The model's suitability for predicting heart disease risk is confirmed by statistical tests used to evaluate the model's goodness of fit. The significance of logistic regression in medical research is emphasized in the conclusion, especially when it comes to identifying and controlling cardiovascular risk factors to avoid unfavorable outcomes.

Cardiovascular disease analysis using supervised and unsupervised data mining techniques
The paper addresses the critical issue of cardiovascular diseases (CVDs), which are a leading cause of global mortality. Because CVDs have a major impact on public health, the study highlights the significance of early detection and treatment. A dataset with 14 characteristics linked to the diagnosis of heart disease is analyzed using a variety of data mining techniques, such as decision trees, support vector machines, Bayesian networks, and k-nearest neighbors. The methodology consists of preparing the dataset, applying various data mining algorithms, and segmenting the data using Simple K-Means clustering. Findings show that the support vector machines approach performed best in terms of precision (97.70%) and recall (97.70%), indicating that it is a useful diagnostic tool for cardiovascular disorders. The study highlights the potential of data mining techniques in healthcare decision-making and advances the development of diagnostic tools for CVDs.

Proposed Approach:

The proposed approach includes data collection, cleaning, modeling, exploratory data analysis (EDA), visualization, development of a prediction tool (tentative).

The proposed approach entails a systematic methodology for leveraging machine learning techniques to analyze a comprehensive dataset on cardiovascular health. Beginning with data collection from reliable sources such as Kaggle, rigorous data cleaning and preprocessing steps will be undertaken to ensure data quality and consistency. Exploratory data analysis (EDA) will follow, involving statistical analysis and visualization techniques to uncover patterns, correlations, and trends within the dataset. Feature selection and engineering will then be employed to identify relevant variables and enhance model performance. Subsequently, various machine learning algorithms will be applied to develop predictive models for heart disease risk assessment, with careful evaluation and validation to ensure robustness and generalization. Interpretation of model results will provide insights into the factors influencing heart disease risk. Finally, stakeholder engagement and communication will facilitate the dissemination of research findings and the development of evidence-based interventions and policies aimed at improving cardiovascular health outcomes. We are planning to develop a prediction tool with a potential web user interface to predict the heart disease of an individual based on their values.

Proposed Method:

In this project, we have acquired the dataset from Kaggle. This dataset consists of 1190 observations and 12 variables. We intend to implement the data cleaning process by first checking for the duplicates, null values, discrepancies in the data set and omitting them. For the exploratory data analysis, we intend to use either Python(Spyder) or R(RStudio) which would assist us in gaining valuable information with respect to the dataset. We intend to implement various machine learning models and find out which model is going to be the most accurate one. We will be implementing the models(Logistic Regression, SVM, Random Forest, Decision tree, Gradient Boosting Classifiers etc.) in Python. Furthermore, we intend to take our research to the next level by planning to develop a potential heart disease prediction tool which will provide the user an interface to predict if an individual has heart disease or not based on his vitals/data. The development of prediction tool is tentative and is contingent upon the workload throughout the semester.

Project Website Link:

mason.gmu.edu/~aanumall

Timeline:

We are still working out the exact details, but we intend to complete the project in a time period of 10-12 weeks. Below is the tentative schedule for the project:

TASK NAME	TENTATIVE TIME
Project Initiation Phase	Week 1-2
Data Collection, Cleaning & Preparation Phase	Week 3-4
Model Development Phase	Week 5-8
Model Evaluation and Interpretation Phase	Week 9-10
Reporting and Presentation Phase	Week 11-12
Heart Disease Prediction Tool Development	Week 12

REFERENCES

[1] Fabio Mendoza, A. P. (2016). *Cardiovascular Disease Analysis Using Supervised and.* JSW-Journal of Software. doi:10.17706/jsw.12.2.81-90

[2] Fedesoriano. (2022). *Heart Failure Prediction Dataset.* Kaggle. Retrieved from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

[3] Priyadarshini, R. G. (2021). *Analysis of heart disease using statistical techniques.* IOPScience. doi:10.1088/1742-6596/1770/1/012105

[4] Salma Banu, S. S. (2017). *Prediction of heart disease at early stage using data mining and big data analytics: A survey.* IEEE. doi:10.1109/ICEECCOT.2016.7955226