Data Mining in Healthcare for Heart Diseases

Published BY: Umair Shafique, Fiaz Majeed, Haseeb Qaiser, and Irfan Ul Mustafa
ISSN: 2028-9324 Vol. 10 No. 4, Mar. 2015, 1313

1 INTRODUCTION:

Data Mining, or Knowledge Discovery in Databases (KDD) [1], is one of the most prominent areas of research. It involves identifying interesting patterns and extracting meaningful insights from data. Various data mining tools and techniques are employed to predict behaviors and trends, enabling experts to make proactive and accurate decisions based on the derived knowledge. Globally, data mining has been successfully applied in diverse fields such as marketing, banking, and business. However, its potential in the healthcare sector remains underutilized.

Currently, heart or cardiovascular diseases are a critical concern in the healthcare industry worldwide. In April 2011, the World Health Organization (WHO) reported [2] that coronary heart disease accounted for 15.36% of total deaths in Pakistan. Furthermore, the WHO projected that by 2030, over 23 million people globally will die annually from heart diseases [3].

Despite the massive amount of data generated by the healthcare industry on heart diseases, much of it remains underutilized, failing to contribute effectively to decision-making. Doctors and healthcare experts typically rely on their personal experience to predict the presence of specific heart conditions in patients, which can sometimes lead to inaccurate outcomes. This highlights the need for leveraging data mining techniques to analyze patient data, uncover hidden patterns, and support decision-making in multiple ways. Each data mining technique serves a specific purpose depending on the requirements and objectives. While data mining in healthcare is both essential and complex, it holds great promise for addressing real-world challenges in diagnosing and treating diseases.

The application of data mining in healthcare offers numerous benefits, such as grouping patients with similar health conditions to enable tailored treatments, ensuring the availability of medical solutions at lower costs, facilitating safe and timely healthcare, identifying the causes of diseases, and optimizing treatment methods. Additionally, it enhances resource utilization and aids healthcare organizations in formulating efficient policies.

In our study, we will utilize an online dataset of heart patients and analyze it using Weka, a data mining software. We will implement three algorithms—Decision Tree, Neural Network, and Naïve Bayes—with and without attribute selection, evaluating the performance of each.

Methodology:

A substantial amount of research has been conducted on heart disease prediction using data mining techniques. Some notable studies are outlined below:

Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm (2010)
This study proposed a heart disease prediction system utilizing a Genetic Algorithm, an

optimization technique inspired by natural selection and genetics [6]. The system aimed to enhance prediction accuracy by reducing the number of attributes from 13 to 6. Three classifiers—Decision Tree, Classification via Clustering, and Naïve Bayes—were employed, and the Weka data mining tool was used for experimentation. Results showed that the Decision Tree classifier achieved the highest accuracy and fastest construction time compared to Naïve Bayes and Classification via Clustering.

HDPS: Heart Disease Prediction System (2011)
This system used a single data mining algorithm, Artificial Neural Networks (ANN), to classify heart disease based on 13 attributes [7]. The dataset, containing 303 instances, was sourced from the UCI Machine Learning Repository. ANN was used for classification, showcasing its effectiveness in heart disease prediction.

Data Mining Neural Network Approach for Heart Disease Prediction (2012)
This study employed a Multilayer Perceptron Neural Network (MLPNN) with the Back Propagation (BP) algorithm [8]. MLPNN, a prominent neural network model, consists of multiple interconnected layers, with BP calculating the error between predicted and actual values. Using the Weka tool, experiments were conducted on a dataset of 573 records, divided into training and testing sets. A total of 15 attributes were used, achieving nearly 100% prediction accuracy.

Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers (WAC) (2013)
This system assigned weights to attributes based on their predictive capability [9]. Implemented in Java, it used a benchmark dataset with 303 records and 14 attributes from the UCI repository. The Weighted Associative Classifiers achieved approximately 80% prediction accuracy.

Comparison of Data Mining Approaches for Heart Disease Prediction (2013)
This study compared three techniques—Naïve Bayes, J48 Decision Tree, and Bagging (Bootstrap Aggregation)—to predict heart disease [10]. The experiments, conducted using the Weka tool, employed a dataset from the Hungarian Institute of Cardiology with 76 raw attributes, of which 11 were selected. Results showed Bagging achieved the highest accuracy (85.03%), followed by J48 Decision Tree (84.35%) and Naïve Bayes (82.31%). The use of Bagging significantly improved classification accuracy.

Frequent Feature Selection Method for Heart Disease Prediction (March 2014)
This study utilized a Frequent Feature Selection method combined with fuzzy measures and nonlinear integrals to predict heart disease [11]. Experiments conducted with the Weka tool involved 1,000 records and 8 attributes. Weights were assigned to attributes to predict heart attacks. The mining process involved two stages:

Weighted Support: Assigning relative weights to transactions in the dataset.

Maximal Frequent Itemset Algorithm (MAFIA): Combining existing and new algorithmic ideas to mine frequent itemsets.

At the final stage, the significance weight of each pattern was calculated, resulting in improved prediction accuracy.

Conclusion:

Our study focused on the application of data mining techniques in the healthcare domain, specifically targeting heart diseases. Heart disease is a life-threatening condition that can lead to severe complications, including death. We utilized publicly available heart patient data from the UCI repository, comprising 597 unique instances. Classification, a key data mining technique, was implemented using three algorithms: Decision Tree, Neural Networks, and Naïve Bayes.

To select the most suitable tool for our experiments, several important factors were considered, leading to the choice of the Weka machine learning software. The performance of the algorithms was evaluated using multiple metrics, including accuracy, precision, F-measure, ROC curve values, true positive rate (TP rate), and false positive rate (FP rate).

Four experiments were conducted under two different scenarios. In the first scenario, all attributes were used, while in the second scenario, only selected attributes were utilized. The dataset was prepared in ARFF format, compatible with Weka. Results from the experiments revealed that the Naïve Bayes classification algorithm achieved the highest accuracy of 82.914%.

This study demonstrates that data mining techniques can be effectively and efficiently applied to predict heart diseases. The outcomes of this research can serve as an assistive tool to support more consistent and accurate diagnoses of heart disease, ultimately improving patient care and decision-making in the healthcare industry.