Plagiarism - Report

Originality Assessment

27%

Overall Similarity

Date: Mar 25, 2024
Matches: 904 / 3006 words
Sources: 23

Remarks: Moderate similarity detected, you better improve the document (if needed).

Verify Report:

17 HEART DISEASE PREDICTION USING RANDOM FOREST ALGORITHM

S.ANANDA KUMAR ,

Department Of Computer Science, Sri Kaliswari College, Sivakasi,

L.PRIYA M.Sc., M.Phil,

Assistant Professor, Department Of Computer Science, Sri Kaliswari College, Sivakasi.

ABSTRACT: Heart disease is a significant health concern worldwide, necessitating effective predictive models for early detection and intervention. 1 In this study, we employ five popular machine learning algorithms: Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), to develop predictive models for heart disease classification. The dataset used in this analysis undergoes extensive data cleaning, include handling missing values, standardizing features, and addressing class imbalance. Each algorithm is implemented and evaluated using appropriate performance metrics such as accuracy, precision, recall, and F1-score. Comparative analysis of these models provides insights into their effectiveness in predicting heart disease, aiding healthcare professionals in making informed decisions for patient care and management.

Keywords: Heart Disease, Random Forest, Logistic Regression, 1 Decision Tree, Support Vector Machine, K-Nearest Neighbor

1. INTRODUCTION:

Heart disease remains a prevalent health issue globally, posing significant challenges to healthcare systems and individuals alike. 4 Heart disease is a prevalent and potentially life threatening condition that affects millions of people worldwide. Understanding the factors that contribute to heart disease and developing effective predictive models are crucial steps in combating this health issue. 1 In this study, we aim to explore the effects of various factors on heart disease and develop predictive models using machine learning techniques such as Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Additionally, we will discuss three different types of data cleaning approaches to ensure the quality and reliability of our analysis.

Heart disease encompasses a range of conditions that affect the heart's functioning, including coronary artery disease, heart rhythm problems (arrhythmias), and heart defects present at birth (congenital heart defects). Risk factors for heart disease include high blood pressure, high cholesterol, smoking, diabetes, obesity, poor diet, lack of physical activity, and family history. The effects of heart disease can be devastating, leading to complications such as heart attack, heart failure, stroke, and even death. Early detection and management of risk factors are essential in preventing or mitigating the impact of heart disease. Heart disease, or cardiovascular disease (CVD), represents a broad category of conditions that affect the heart and blood vessels. It encompasses various disorders, including coronary artery disease, heart failure, arrhythmias, and congenital heart defects, among others. Heart disease is a leading cause of mortality globally, contributing to millions of deaths annually. It affects individuals of all ages, genders, and ethnicities, with certain risk factors predisposing individuals to its development.

Common risk factors for heart disease include high blood pressure (hypertension), elevated cholesterol levels, smoking, diabetes, obesity, physical inactivity, unhealthy diet, excessive alcohol consumption, and family history of cardiovascular conditions. These risk factors can lead to the build up of plaque in the arteries (atherosclerosis), narrowing the blood vessels and restricting blood flow to the heart muscle, which can result in various cardiac complications, such as heart attacks and strokes. Symptoms of heart disease can vary depending on the specific condition but may include chest pain or discomfort (angina), shortness of breath, fatigue, palpitations, dizziness, and swelling in the extremities. Diagnosis typically involves a combination of medical history assessment, physical examination, and diagnostic tests such as electrocardiography (ECG), echocardiography, stress tests, and cardiac catheterization. Treatment approaches for heart disease aim to alleviate symptoms, slow disease progression, and reduce the risk of complications. Treatment modalities may include lifestyle modifications (e.g., adopting a heart healthy diet, engaging in regular exercise, quitting smoking), medications (e.g., antihypertensives, statins, antiplatelet agents), interventional procedures (e.g., angioplasty, stent placement), and surgical interventions (e.g., coronary artery bypass grafting). Prevention strategies play a crucial role in mitigating the burden of heart disease. Efforts to prevent heart disease focus on addressing modifiable risk factors through health promotion and disease

prevention initiatives. These include promoting healthy lifestyles, implementing public health policies to encourage tobacco control, improving access to nutritious foods, facilitating regular physical activity, and enhancing screening and early detection programs. Overall, understanding the background information on heart disease is essential for healthcare professionals, policymakers, and the general public to implement effective prevention, early detection, and management strategies, ultimately reducing the prevalence and impact of cardiovascular disease on global health.

2. LITERATURE SURVEY:

2.1. A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method - Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang and Qian Wang

Due to the widespread frequency of cardiac disease, efficient diagnostic techniques are required. In an effort to facilitate the detection of heart illness, this study presents ReliefF and Rough Set (RFRS), a hybrid categorization approach. This system, which consists of two subsystems, shows encouraging outcomes. The ReliefF method is used in three steps for the first subsystem, RFRS feature selection: (i) data discretization; (ii) feature extraction; and (iii) feature reduction using a heuristic Rough Set reduction approach created specifically for this purpose. An ensemble classifier based on the method is used by the second subsystem. A jackknife cross-validation strategy was used in the studies, which used the Statlog (Heart) dataset from the UCI database. An astounding 92.59% classification accuracy was attained at its highest point.

2.2 Heart Disease Prediction using Machine Learning Techniques - Devansh Shah1 · Samir Patel, Santosh Kumar Bharti

Heart disease is one of the world's leading causes of death; thus, prompt and precise diagnosis is essential for successful treatment. In order to anticipate cardiac disease, data mining techniques provide useful tools for examining large medical datasets. This study investigates several characteristics linked to heart disease using supervised learning algorithms such random forest, K-nearest neighbor, decision tree, and Naïve Bayes. Using the Cleveland dataset from the UCI repository, which has 76 attributes and 303 cases, just 14 attributes are tested in order to assess how well various algorithms perform. The purpose of the study is to forecast the patients' risk of heart

disease. The K-nearest neighbor algorithm had the greatest accuracy score, according to the results. Notably, using the Cleveland dataset, the random forest method attained an accuracy of 91.6%. with the Cleveland dataset and 97% using the dataset from People. These results highlight the promise of machine learning methods in heart disease prediction; the random forest and K-nearest neighbor algorithms show very encouraging outcomes.

2.3 Heart 11 Disease Prediction using Machine Learning and Data Mining - Keshav Srivastava, Dilip Kumar Choubey

Even though heart disease only weighs 300 grams, it has been a major factor in the decades-long decline in death rates. Significant contributions to its occurrence are made by factors including drinking, smoking, and eating an imbalanced diet. Analyzing clinical data is still difficult despite advances in technology. An exciting new direction in data analysis and interpretation is provided by machine learning approaches, especially 1 when it comes to the diagnosis of diseases like Dilated Cardiomyopathy, Arrhythmia, and Coronary Heart Disease. Python's Flask framework has been used to create a web application that lets users enter characteristics and forecast the course of heart disease. Utilizing technology for early illness diagnosis and intervention has advanced significantly with the inclusion of machine learning algorithms and user-friendly interfaces. This work uses 1 a variety of machine learning and data sources to present a unique method for determining the presence of illness. Data mining techniques. The research combines these algorithms' computational capacity with the Cleveland Heart Disease Dataset. Notably, K-Nearest Neighbors had the best accuracy (87%), out of all the algorithms examined.

2.4 Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques - Abid ishaq , Saima sadiq ,Muhammad umar, Saleem ullah , Seyedali mirjalili, Vaibhav rupapara, and Michele nappi

Cardiovascular disease continues 5 to be a major worldwide health problem, making patient survival prediction difficult. A solution is provided by data mining tools, which convert enormous amounts of healthcare data into useful insights. Several research stress how crucial it 2 is to find important characteristics in order to improve the performance of machine learning models in this field. The analysis of 299 hospitalized patients who survived cardiac failure is the main goal of this study.

Finding important characteristics and useful data mining strategies will help to increase the precision of cardiovascular patient survival prediction. Decision Tree, AdaBoost, Logistic Regression, SGD, Random Forest, GBM, ETC, G-NB, and SVM are among the nine classification models used. SMOTE is used to solve the issue of class imbalance. These results highlight the significance of feature selection and the effectiveness of ETC in precisely forecasting the survival of cardiovascular patients. Therefore assisting in the improvement of clinical judgment in this crucial field of medicine. Furthermore, Random Forest selects the highest-ranked features to train 1 machine learning models on. Models with the complete feature set are compared to one another. The results show that ETC works better than other models, predicting the survival of cardiac patients with an accuracy value of 0.9262 using SMOTE.

2.5 3 Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System - Shadman Nashif, Md. 15 Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam

Cardiovascular illnesses are now the world's largest cause of death, impacting people in industrialized, developing, and undeveloped nations alike. Reducing death rates requires both ongoing clinical surveillance and early diagnosis of heart disorders. Unfortunately, due to budget limitations, access to round-the-clock medical professional care is frequently restricted. This study suggests a cloud-based machine learning-based cardiac 3 disease prediction system as a solution to these problems. Accurate illness identification depends on the choice of an effective machine learning algorithm, which is made possible by a comparative evaluation of several algorithms utilizing the WEKA platform. Additionally, an Arduino-based real-time 20 patient monitoring system is built to enable ongoing monitoring of patients with cardiac disease. Real-time data is recorded by this device, including body temperature, blood pressure, humidity, and heartbeat, sending information in batches of ten seconds to a central server. Medical practitioners can use an application to get 3 real-time sensor data and start live video streaming in case they need to act right away. Furthermore, the system uses GSM technology to automatically notify the designated physician when any patient parameter surpasses predetermined criteria. 1 The importance of using machine learning techniques for cardiac disease prediction and the creation of real-time patient monitoring systems to improve clinical treatment and

patient outcomes are generally highlighted by this research review. The suggested technique is validated on two popular open-access datasets using 10-fold cross-validation to evaluate the program's efficacy in detecting cardiac disease. Particularly intriguing is the 3 Support Vector Machine (SVM) method, which achieves a 97.53% accuracy level coupled with remarkable sensitivity and specificity rates.

3. METHODOLOGY:

3.1 Random Forest: To begin our investigation, we first load the necessary libraries: scikit-learn for machine learning features, NumPy for numerical operations, Matplotlib and Seaborn for data visualization, and Pandas for data processing. After loading the dataset from a CSV file, some preliminary research is done to understand its properties and organization. Next, preprocessing chores follow, including scaling numerical features with StandardScaler to guarantee scale consistency. To facilitate model assessment, 1 the dataset is then divided into training and testing sets using scikit-learn's train_test_split function. Using the training set of data, the Random Forest classifier is instantiated and trained. Test data is used to make predictions, and measures like accuracy score and classification report—along with the confusion matrix visualization—are used to evaluate the model's performance. Furthermore, a bar to begin this study, we import the necessary libraries, such as Seaborn, Matplotlib, NumPy, and Pandas for data management, numerical computations, and the creation of a chart to compare the average accuracy ratings of different classifiers. This all-inclusive technique outlines the phases of deployment and methods for 16 using Random Forest to forecast heart disease. The Random Forest algorithm has significant potential, achieving 93% accuracy.

3.2 3 Logistic Regression: To start, we load the necessary libraries: scikit-learn for machine learning features, Matplotlib and Seaborn for data visualization, NumPy for numerical calculations, and Pandas for data manipulation. After loading the dataset from a CSV file, some preliminary research is done to ascertain its properties and organization. Next, there are preprocessing chores 3 related to the data, such as scaling numeric characteristics with StandardScaler to guarantee scale consistency. To aid in the evaluation of the model, the dataset is then divided into training and testing sets using train_test_split from scikit-learn. Using training data, logistic regression is instantiated and trained as

the predictive modeling approach. Test data 2 is used to make predictions, and metrics like the confusion matrix, accuracy score, and classification report are used to assess the model's performance. This all-encompassing method outlines the technique and stages in execution for 3 using logistic regression to predict heart disease. The method for Logistic Regression has significant potential, with an accuracy rate of 82%.

3.3 Decision Tree: The loading 1 of the dataset from a CSV file marks the beginning of the data preparation step, which is then followed by an initial investigation to ascertain the structure and properties of the dataset. Then, in order to guarantee scale consistency throughout the dataset, numerical characteristics are scaled using StandardScaler. Next, using scikit-learn's train_test_split method, the dataset is split into training and testing sets. Predictive modeling is done using decision tree classification, which is instantiated and trained using training data. 2 On the basis of the test data, predictions are produced, and metrics like the confusion matrix, accuracy score, and classification report are used to assess the model's performance. This thorough approach demonstrates the stages needed in 1 using Decision Tree classification to the prediction of heart disease, including the technique and execution. 2 The Decision Tree algorithm has significant potential, producing an accuracy level of 86%.

3.4 Support Vector Machine (SVM): The implementation of a Support Vector Machine (SVM) classifier for heart disease prediction. 4 The first step in the data preprocessing phase is to import the dataset from a CSV file and perform a preliminary investigation to comprehend its features and organization. To guarantee consistency in scale throughout the dataset, numerical features are then normalized using StandardScaler. Next, using the train_test_split function from scikit-learn, 1 the dataset is divided into training and testing sets. Using 23 support vector machines for classification, which are instantiated and trained using training data, is a predictive modeling technique. 2 The test data is used to generate predictions, and metrics like the confusion matrix, accuracy score, and classification report are used to assess the performance of the model. This thorough method shows how to use 14 Support Vector Machine classification for heart disease prediction, including the implementation procedures and methodology. 83% accuracy is achieved with the support Vector machine approaches.

3.5 K-Nearest Neighbors (KNN): The dataset for this research is first loaded from a CSV file, and then it is explored to learn more about its features and structure. StandardScaler is then used to standardize numerical characteristics in order to guarantee scale consistency throughout the dataset. The train_test_split function from scikit-learn is then used to divide the dataset into training and testing sets. The predictive modeling method used is KNN classification, which is instantiated and trained using the training set. On the basis of the test data, predictions are produced, and metrics like the confusion matrix, accuracy score, and classification report are used to assess the model's performance. This thorough approach demonstrates the stages needed in utilizing KNN classification for heart disease prediction, including methodology and implementation. The K-Nearest Neighbors(KNN) algorithm has significant potential, achieving 87% accuracy.

RESULT ANALYSIS:

We evaluated three data cleaning techniques and discovered that scaling features, eliminating outliers, and imputed missing data greatly enhanced model performance for heart disease prediction. Outlier removal improved model robustness, feature scaling guaranteed consistent model convergence, and imputation maintained data integrity. Together, these methods produced predictions that were more accurate, demonstrating how important it is to preprocess data for machine learning tasks.

Figure.1:Accuracy graph

TABLE:

Data Cleaning-1: Data Loading, Exploration, Data Visualization, Data Preprocessing, Model Training.

Data Cleaning-2: Data Loading, Exploratory Data Analysis (EDA), Handling Missing Values, Handling Duplicates, Data Splitting.

Data Cleaning-3: Data Loading, Data Preparation, Handling Missing Values, Encoding Categorical Variables, Data Splitting, Feature Scaling.

Accuracy	
Data cleaning-1	

Data cleaning-2

Data cleaning-3

Random Forest

92

92

93

Logistic Regression

85

82

82

Decision Tree

89

91

86

SVM

72

76

83

K-Nearest Neighbors

82

74

87

CONCLUSION:

In summary, this study underscores the significance of utilizing 3 machine learning algorithms for early heart disease detection. Leveraging Random Forest, Logistic Regression, Decision Tree, SVM, and KNN models, we've developed effective predictive tools. Rigorous data cleaning, including 1 handling missing values and standardizing features, ensured the reliability of our analyses. Through thorough evaluation, we've gained valuable insights into each model's performance,

aiding healthcare decision-making. Moving ahead, refining these models could revolutionize heart disease diagnosis and treatment globally.

REFERENCES:

[1] Heart Disease Prediction using Machine Learning Techniques, Devansh Shah1. · Samir Patel1 · Santosh Kumar Bharti(2020)

[2] Machine learning algorithms using binary classification and multi model ensemble techniques for skin diseases prediction, Vikas Chaurasia and Saurabh Pal - Int. J. Biomedical Engineering and Technology, Vol. 34, No. 1 (2020)

[3] Heart Disease Prediction using Machine Learning and Data Mining, Keshav Srivastava, Dilip Kumar Choubey - International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1 (2020)

[4] Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques - Abid ishaq , Saima sadiq ,Muhammad umar, Saleem ullah , Seyedali mirjalili, Vaibhav rupapara, and Michele nappi(2021)

[5] Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System - Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam (2018)

[6]"Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", C. B. C. Latha, S.C.Jeeva, Informatics in Medicine Unlocked 16,2019

[7] An improved ensemble learning approach for the prediction of heart disease risk, Ibomiye Domer Mienve, Yanxia sun, Zenghui Wang(2020)

[8] Classification models for heart disease prediction using feature selection and PCA, Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès(2020)

[9] Survival prediction of heart failure patients using machine learning techniques,

Asif Newaz, Nadim Ahmed, Farhan Shahriyar Haq(2021)

Sources

1	https://www.nature.com/articles/s41598-023-40717-1 INTERNET 9%
2	https://www.academia.edu/80992122/Improving_the_Prediction... INTERNET 5%
3	www.ncbi.nlm.nih.gov/pmc/articles/PMC8330430/ INTERNET 3%
4	bing.com/videos INTERNET 2%
5	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7123129 INTERNET 1%
6	https://www.inderscienceonline.com/doi/abs/10.1504/IJBET.2020.110361 INTERNET 1%
7	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10460604 INTERNET 1%
8	https://www.researchgate.net/publication/342210798_Heart_Disease... INTERNET <1%
9	https://doaj.org/article/79e54a2a745e49658d089301c4cc35fa INTERNET <1%
10	https://www.mayoclinic.org/diseases-conditions/heart-disease INTERNET <1%
11	https://www.researchgate.net/profile/Dilip-Choubey-2/pub... INTERNET <1%
12	www.verywellhealth.com/key-symptoms-of-heart-disease ... INTERNET <1%
13	www.ncbi.nlm.nih.gov/pmc/articles/PMC5740998/ INTERNET <1%
14	www.mdpi.com/2411-9660/6/5/87 INTERNET <1%

15	https://www.jetir.org/download1.php?file=JETIR2305031.pdf INTERNET <1%	
16	www.sciencedirect.com/science/article/pii/S235291482300... INTERNET <1%	
17	iopscience.iop.org/article/10.1088/1742-6596/1817/1/012... INTERNET <1%	
18	https://www.pythonprog.com/sklearn-preprocessing-robustscaler INTERNET <1%	
19	https://www.datacamp.com/tutorial/data-preparation-with-pandas INTERNET <1%	
20	www.ncbi.nlm.nih.gov/pmc/articles/PMC10422369/ INTERNET <1%	
21	machinelearningmodels.org/python-tutorial-data-cleanin... INTERNET <1%	
22	https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20... INTERNET <1%	
23	www.imsl.com/blog/using-support-vector-machines-jmsl INTERNET <1%	

EXCLUDE CUSTOM MATCHES	OFF
EXCLUDE QUOTES	OFF
EXCLUDE BIBLIOGRAPHY	OFF