DATA ANALYSIS ON “HEART FAILURE PREDICTION”

**BY: SHANTANU SHAW**

**SCHOOL OF COMPUTER APPLICATION , KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**

**[shantanushaw0102@gmail.com](mailto:shantanushaw0102@gmail.com)**

**Abstract :**

Cardiovascular diseases (CVDs) are a major global health concern, causing the highest number of deaths worldwide. With an estimated 17.9 million lives lost each year, they account for 31% of all global deaths. Heart failure, a common consequence of CVDs, can be predicted by analyzing a dataset containing 12 features. **I aim to use logistic regression for this analysis.** Preventing CVDs involves addressing behavioral risk factors like tobacco use, unhealthy diet, obesity, physical inactivity, and harmful alcohol consumption through widespread strategies. Early detection and management for individuals with cardiovascular disease or high risk are crucial, and a machine learning model can play a key role in achieving this goal.

**References :**

1. <https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data>
2. <https://www.kaggle.com/code/abhaysen07/heart-failure-prediction-using-logistic-regression/notebook>
3. <https://scholar.google.com/>

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.NO.** | **PAPER REFERENCES** | **PURPOSE OF PAPER** | **METHODOLOGY USED WITH ACCURACY** | **IS AI/ML USED? NAME IT** | **DETAILS OF DATA SET** | **OUR ANALYSIS WITH ACCURACY** |
| 1. | * Usha. K Dr, “Analysis of Heart Disease Dataset using neural network approach”, IJDKP, Vol 1(5), Sep 2011. * Shadab Adam Pattekari and Asma Parveen, ”PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES”, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624, Vol 3, Issue 3, 2012, pp 290-294. | The purpose of this paper is to conduct a survey on various data mining techniques for the prediction and diagnosis of heart disease. The authors aim to identify the algorithm that best suits heart disease prediction with a high level of accuracy. The paper explores several single and hybrid data mining algorithms and discusses their advantages and limitations in the context of heart disease prediction. | The paper employs a survey methodology to analyze various data mining techniques. It reviews several algorithms, including Support Vector Machine (SVM), Decision Tree, K-Mean, Naïve Bayes, and Neural Network.  **Accuracy : 86.70%** | * Support Vector Machine (SVM) * Decision Tree * K-Mean * Naïve Bayes * Neural Network | Various datasets are used .No specific data set is mentioned | Data set used : Heart failure prediction  Source : Kaggle.com  **Analysis Details :**  **Dataset Visualization :**   * Column : 13 columns age, anaemia, creatinine\_phosphokinase,diabetes,ejection\_fraction, high\_blood\_pressure, platelets,serum\_creatinine, serum\_sodium, sex, smoking, time, deaths. * No.of entries : 299 * No null values * Mean age : 60 * Mean deaths : 0.32 * Analysis is done through different graphs such as count-plot , scatter-plot , dist-plot.   **Data transformation :**  **Used Logistic regression to find the accuracy , precision, recall and F1-score.**  Performance of the dataset with outliers :  **Accuracy Score: 80.0 %**  **Recall Score: 60.0 %**  **Precision Score: 68.57142857142857 %**  **F1 Score: 64.0 %**  **Confusion matrix :**   | **col\_0** | **0** | **1** | | --- | --- | --- | | **row\_0** |  |  | | **0** | 84 | 16 | | **1** | 11 | 24 |   **Checking outliers :**  To check for outliers in a dataset, we used the **IQR (Interquartile Range) Method.**  There are 75 rows x 13 columns outliers.  Performance of the dataset without the outliers:  **Accuracy Score : 86.13861386138613 %**  **Recall Score : 73.33333333333333 %**  **Precision Score : 78.57142857142857 %**  **F1 Score : 75.86206896551724 %**  **Confusion matrix :**   | **col\_0** | **0** | **1** | | --- | --- | --- | | **row\_0** |  |  | | **0** | 65 | 8 | | **1** | 6 | 22 |   Performance analysis done through pie chart  And AUC-ROC curve graph.  **AUC-ROC curve:**  Screenshot 2023-11-14 131507  **PIE CHART:**  **Screenshot 2023-11-14 131657** |
| 2. | * B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin and X. Wei, "Predicting the Risk of Heart Failure with EHR Sequential Data Modeling," in IEEE Access, vol. 6, pp. 9256-9261,2018. * S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019. | The paper aims to address the critical issue of heart disease, which is a leading cause of death worldwide. It emphasizes the importance of early detection and proposes a tool using data mining and machine learning to analyze patient data for early prediction of heart disease. | The research utilizes machine learning techniques, specifically the Support Vector Machine (SVM) algorithm, for predicting heart disease. The paper outlines the steps of data pre-processing, division into training and testing sets, model construction using SVM, and evaluation of the model's accuracy.  **Accuracy : 85.20%** | The main algorithm used is Support Vector Machine (SVM). | * Size: The dataset consists of 13 attributes related to patients. * Attributes: Age, Sex, Chest Pain type, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting Maximum Heart Rate achieved |
| 3. | * World Health Organization, Cardiovascular Diseases. * American Heart Association, Classes of Heart Failure. * American Heart Association, Heart Failure. | The paper aims to explore the use of machine learning (ML) and deep learning (DL) models for diagnosing heart disease. It emphasizes the significance of timely and accurate diagnosis due to the alarming global statistics on cardiovascular diseases. The focus is on utilizing various datasets, particularly the Public Health Dataset from 1988, to train models and predict heart disease, considering factors such as age, sex, chest pain type, blood pressure, cholesterol levels, and other relevant attributes. | * Machine Learning Classifiers: Logistic Regression, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, Support Vector Machine (SVM), and XGBoost. * Deep Learning: Sequential model with fully connected dense layers, dropout layers, ReLU activation functions, and binary cross-entropy loss. * Evaluation Process: Confusion matrix, accuracy score, precision, recall, sensitivity, and F1 score.   Accuracy :   * In the first approach (without feature selection and outliers detection), SVM achieved the highest accuracy among ML models, and DL **accuracy was 76.7%.** * In the second approach (with feature selection but no outliers detection), Random Forest outperformed other ML models, and DL **accuracy was 86.8%.** * In the third approach (with feature selection and outliers detection), KNeighbors achieved the highest accuracy among ML models (83.29%), while DL accuracy reached **94.2%.** | Various machine learning models, including Logistic Regression, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, SVM, and XGBoost, are applied to predict heart disease based on the dataset. Additionally, deep learning is employed using a sequential model for comparison. | * Size: The dataset comprises four databases (Cleveland, Hungary, Switzerland, and Long Beach V). * Attributes: Initially, 76 attributes are present, but experiments focus on a subset of 14 attributes, including age, sex, chest pain type, blood pressure, cholesterol, fasting blood sugar, electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression, slope, number of major vessels, thalassemia, and target. * Preprocessing: Outlier handling, feature selection, and normalization are applied to enhance data quality. |
| 4. | * World Health Organization (WHO) report * Australian Institute of Health and Welfare (AIHW) report and various studies like Amin et al. | The main purpose of the paper is to address the limitations in existing studies related to heart disease prediction, specifically focusing on the shortcomings of datasets, dataset standardization, and algorithm tuning. The goal is to develop an improved and accurate heart disease prediction model using machine learning classifiers. | * Data Preprocessing: Refining and standardizing the collected datasets to address inconsistencies and missing values. * Machine Learning Classifiers: Applying classifiers such as Random Forest, XGBoost, Decision Trees, SVM, Naïve Bayes, Logistic Regression. * Performance Evaluation: Assessing classifier performance using metrics like accuracy, precision, sensitivity, recall, and F-measures. * SVM shows comparatively good performance for negative classes, with a recall, precision, and F-measure of **94%, and an accuracy of 96.72%.** * LR presents a small **precision of 78%**. | Machine learning (ML) classifiers are used for heart disease prediction. Classifiers mentioned include Random Forest, XGBoost, Decision Trees, SVM, Naïve Bayes, Logistic Regression, Linear Discriminant Analysis, AdaBoost, and Extra Trees. | The primary dataset used is the Cleveland heart disease dataset, obtained from the University of California, Irvine (UCI) online repository. It contains 303 records, with 76 features in its original form. |
| 5. | * D. Shah, S. Patel, and S. K. Bharti, “Heart disease prediction using machine learning techniques,” SN Computer Science, vol. 1, no. 6, pp.1–6, 2020. * R. Saravanan and P. Sujatha, “A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches data classification,” in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 945–949. | The primary purpose of the paper is to address the challenges in early detection and diagnosis of cardiovascular diseases (CVD) using machine learning (ML) classification algorithms. It emphasizes the significance of ML in analyzing data related to heart health and explores the use of various ML techniques for cardiac ailment diagnosis. | The paper evaluates the performance of ML algorithms using the following metrics:  Accuracy:   * Logistic Regression **(85.3%)** * KNN (90.2%) * SVM (94.56%) * Random Forest (87.5%) * Gradient Boosting (87.5%). | The paper uses the following ML classification algorithms for cardiac ailment diagnosis:   * Logistic Regression * K-Nearest Neighbors (KNN) * Support Vector Machines (SVM) * Random Forest (RF) * Gradient Boosting (GB) | * The dataset used is a combination of five heart datasets, resulting in a total of 918 insights. * After pre-processing, the final dataset contains 918 insights with 14 attributes, including age, sex, cholesterol, etc. |

NAME : SHANTANU SHAW

CLASS: SEC ‘C’

ROLL NO: 2270325

EMAIL: [shantanushaw0102@gmail.com](mailto:shantanushaw0102@gmail.com)

MOB.NO : 8825380997