**A** **SUPERVISED MACHINE LEARNING PROJECT**

**ON HEART DISEASE (FINAL PROJECT).**

**BY**

**TAJUDEEN MOHAMMED AWAL**

**ABDUL RAHMAN ISSAH**

**ILYASS ABDUL RAHMAN**

**KAMIL HIDIR**

**AISHATU YUSIF**

**DATA SCIENCE BOOT CAMP**

**WITH**

**ZONGOVATION HUB**

**MAY, 2025.**

**Abstract**

Heart disease is inflammation or damage to the heart and blood vessels over time. The disease can affect anyone of any age, gender, or social status. After many studies trying to overcome and learn about heart disease, in the end, this disease can be detected using machine learning systems. It predicts the likelihood of developing heart disease. The results of this system give the probability of heart disease as a percentage. The datasets was handled in python programming using two main algorithms for machine learning; Logistic regression algorithm and the K-Nearest Neighbor’s Algorithm, which shows the best of both for heart disease prediction accuracy. The results we get from this study show that the K-Nearest Neighbor’s algorithm is the algorithm with the most excellent precision. And the highest accuracy with a score of 87% in predicting heart disease using machine learning algorithms.

**I. INTRODUCTION**

Heart disease is a range of conditions that affects the heart, while cardiovascular disease (CSD) encompasses both heart and blood vessel diseases. These two (heart disease and Cardiovascular disease) happen to be used interchangeably sometimes. As a leading cause of death globally, CVD accounted for 17.9 million deaths in 2019, representing 32% of all global deaths. With a high prevalence in low- and middle-income countries, including Ghana where CVD prevalence ranges from 0.5% to 65%, it is essential to implement evidenced-based strategies and policies for prevention and control.

**II. METHODS**

**2.1 Literature Review**

Machine learning is a powerful tool that enables systems to learn from experience and improve over time without explicit programing. This technology has numerous applications, including healthcare. Machine learning tackles various problems, primarily regression and classification. Regression methods handle numerical data, while classification addresses binary and multi-category issues. There are two main division: Supervised and Unsupervised learning. Supervised learning relies on labeled data, whereas unsupervised learning identifies patterns and structures in unlabeled dataset. Choosing the right machine learning algorithm is crucial for achieving accurate results. With our analysis on patient data, we will lean more towards supervised machine learning while using Logistic regression algorithm and the K-Nearest Neighbor’s Algorithm for our predictions of the label variable (Heart disease).

**2.1.1 Machine Learning Algorithm**

There are some techniques of machine learning algorithms like classification, clustering, regression, association analysis, and outlier surveys. One is in predicting the heart by technical learning methods, classification. We used classification because our dataset is a classification dataset as we will be predicting the likelihood of heart disease. Hence the classification model we will be using is the K-Nearest Neighbor’s Algorithm (KNN Algorithm) and Logistic Regression Algorithm**.**

**2.1.2 Logistic Regression Algorithm**

Logistics regression, a well-known regression model is a supervised machine learning algorithm used for classification problems (in this case predicting if a patient has heart disease or not). With regards to The heart disease dataset obtained from the UCI Machine Learning repository (dataset we are using), logistic regression can be used to predict the likelihood of heart disease based on the various features we have in our dataset. It contains 12 features and 918 entries. Using logistic regression model on our dataset we ended up with an accuracy score of 84%.

**2.1.3 K-Nearest Neighbor’s Algorithm (KNN Algorithm)**

KNN algorithm is one of the most popular machine learning algorithms, KNN is often used for the data classification process. Heart disease can be predicted also by analyzing the health parameters (Features in our dataset) of each patient. This study uses a UCI machine learning dataset with 12 definite parameters and 918 entries, in the end getting an accuracy of 87%.

**2.2 Collection and Features Selection**

For our Project, the heart disease data was sourced from UCI Cleveland data source from the UCI machine learning repository website, this dataset is a publicly available dataset source. The data set consists of 918 agencies (entries/rows) with 12 attributes. The basis of the data is that there are 76 attributes or features, but in published research, only 14 attributes have been processed. The data is contained in the dataset repository, namely the UCI Cleveland dataset. In the dataset, there are 12 attributes and 1 attribute for prediction or known as the dependent variable name (label variable), the ’Diagnosis’ attribute, and the rest will be entered as input or known as the independent variable. The attribute descriptions are shown in Table below

**DESCRIPTION OF VARIOUS FEATURES**

|  |  |
| --- | --- |
| Features | Description |
| Age | age of the patient (years) |
| Sex | sex of the patient (M: Male, F: Female) |
| ChestPainType | TA: Typical Angina  ATA: Atypical Angina  NAP: Non Angina Pain  ASY: Asymptomatic |
| RestingBP | resting blood pressure (mm Hg) |
| Cholesterol | serum cholesterol (mm/dl) |
| FastingBS | fasting blood sugar (1: if FastingBS > 120 mg/dl, 0: otherwise |
| RestingECG: resting electrocardiogram results (Normal: Normal, ST: having ST | T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria) |
| MaxHR | maximum heart rate achieved (Numeric value between 60 and 202) |
| ExerciseAngina | exercise induced angina (Y: Yes, N: No) |
| Oldpeak | ST (Numeric value measured in depression) |
| ST\_Slope: | the slope of the peak exercise ST segment (Up: up sloping, Flat: flat, Down: down sloping) |
| HeartDisease | output class (1:has heart disease, 0:Patient do not have) |

**2.3 Data Preparation**

This the first stage of our process as we have already outsourced dataset (disease data) for our analysis. In preparing our dataset we will extract all the needed features and get our dataset ready for the next stage.

**2.4 Data Preprocessing**

This is where we begin with our EDA, we first have look at our dataset using the pandas library, then do a .describe(), .info(), .head(), and .shape on the data, check for missing values(isna), check for outliers, duplicates and encode categorical variables among others using some libraries like numpy, seaborn, matplotlib, LabelEncoder from scikit learn, LogisticRegression from sklearn.linear\_model, TRUE from pickle, StandardScaler from sklearn.preprocessing and KNeighborsClassifier from sklearn.neighbors.

**2.4 Feature Selection**

In this stage we will feature scale some features using the StandardScaler, this is done to prevent our models from being bias during prediction. After scaling we then split our dataset into two ie; 80% of our dataset as our train data while the remaining 20% is our test data.

**2.5. Model Building and Prediction**

Using logistics regression algorithm and K-Nearest Neighbor’s Algorithm to build a model that will learn from our train data in other to predict the outcome of our test data.

**2.6. Model Evaluation**

After training and prediction our logistics regression model had an accuracy score of 84% whereas our K-Nearest Neighbor had an accuracy score of 87. The KNN gave a higher accuracy compared to the logistic regression model because of its simplicity, non-parametric nature and the ability to capture complex relationships with the data. We had the following after we used the confusion matrix.

i. 68 Patient were correctly predicted of not having heart disease.

ii. 9 Patient were incorrectly predicted to have heart disease (they actually did not have heart disease).

iv. 20 Patient were incorrectly predicted to not have heart disease (but they did have heart disease).

v. 87 Patient were correctly predicted to have heart disease.

This model can be said to be effective in predicting heart disease by using a different rating algorithm. In this model, we use multiple compounds like chest pain type, rest blood pressure, Chol and others to predict.

**3. CONCLUSION**

Cardiovascular disease (CVD) is a major global health issue, particularly affecting men over 40 years old. Despite its severity, public awareness remains low. Key risk factors including smoking, high blood pressure, high cholesterol, poor diet, physical inactivity, and obesity increase the likelihood of heart disease.

Fortunately, many heart conditions are preventable through lifestyle changes, routine health screenings, and increased education. Strengthening public sensitization and improving access to preventive healthcare are essential for reducing heart disease prevalence and saving lives.

In this study, we introduce the heart disease prediction system with different two type of classifier techniques which are Logistic regression and K-Nearest Neighbor for the prediction of heart disease. The history of the patient that contains a collection of data that leads to heart disease, the data collection of the medical history includes CP (Chest Pain), MaxHR (MaxHearthRate), cholesterol, fasting blood sugar, resting blood pressure, exerciseangina, age, gender amongst others.

Comparing both models we used to predict heart disease; the results we got shows that the K-Nearest Neighbor algorithm is the algorithm with the highest accuracy with a score of 87% in predicting heart disease. In the future we will develop the algorithm to be more efficient and have a higher accuracy rate for predicting heart disease from previous research or studies.