

|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Project Name:  Prediction Of Heart Disease |
| |  |  |  | | --- | --- | --- | | K Madhu Lavanya#: @01442868 | Fall 2023 | Machine Learning Data Science | |

**Implementing a machine learning algorithm using dataset**

**TABLE OF CONTENTS**

[1. Introduction 3](#_Toc152250393)

[2. Background 3](#_Toc152250394)

[2.1.Research Question 4](#_Toc152250395)

[3. Dataset Description 4](#_Toc152250396)

[4. Model’s Baseline 6](#_Toc152250397)

[Advantages of these algorithms 6](#_Toc152250398)

[Disadvantages of these algorithms 7](#_Toc152250399)

[5. Building machine learning model 8](#_Toc152250400)

[6. Conclusion. 12](#_Toc152250401)

[7.Reference List 14](#_Toc152250402)

# 1. Introduction

Machine learning seeks to apply modern machine learning techniques to uncover novel insights in. By leveraging ML algorithms, we aim to extract valuable patterns that address more effectively than current methods. The goal of this machine learning project is to utilize to develop a model that can accurately predict or classify based on patterns in the data. By exploring and experimenting with advanced machine learning algorithms, such as neural networks, random forests, support vector machines, etc., the aim is to construct an optimal model. Successful development of this ML model will provide key learnings that can inform future work and business decisions by enhancing understanding, analysis capabilities, and decision-making. It aimed to touch on applying ML algorithms to the dataset, optimizing predictive accuracy, extracting impactful learnings, and discussing how achieving this goal will lead to enumerated benefits. Applying machine learning algorithms to extract insights from data has emerged as an invaluable technique across industries. This project leverages these powerful ML capabilities to unlock key learnings from which has not yet benefited from advanced analytics. Developing an accurate ML model enables data-driven decision making, enhanced strategy development, and optimized in a domain where these benefits are especially impactful. It is an ideal approach for extracting insights from this dataset because ML algorithms can model complex data patterns that underpin. By optimizing an ML model to understand intricate relationships in these large, multi-dimensional datasets, user can achieve more accurate predictive capabilities than available through traditional statistical methods.

Keywords : Decision Tree, Naive Bayes Logistic Regression, K Neighbours

# 2. Background

Heart disease remains a leading global health concern responsible for high rates of mortality and healthcare costs. This freely available dataset from the UCI Repository contains clinical parameters and heart disease diagnoses for over 300 patients(Alotaibi 2019). By applying and evaluating the performance of machine learning algorithms such as logistic regression, random forests, and neural networks on this standardized dataset, we can determine optimal ML models for extracting disease indicators from patients' medical attributes(Sivapriya *et al*. 2021). Comparing these algorithms will provide insights into the predictive capabilities of AI for improving heart disease screening and prevention. Its aimed to touch on the impact of heart disease, describe the key characteristics of the dataset, discuss how different ML algorithms can be tested on it for predictive modeling relevant to addressing the health issue, and the potential benefits the project offers. This dataset contains medical data on patients including demographics, lifestyle factors, and health measurements(Neupane and Seok 2020.). It can be used to develop and compare machine learning models to predict whether a patient has heart disease based on these characteristics. The goal is identifying key health indicators for determining heart disease risk. The overall message to convey is using the data to build and evaluate ML models for predicting heart disease using health and lifestyle indicators.

## 2.1.Research Question

* Which machine learning algorithm logistic regression, random forest, neural networks etc.) is most accurate at predicting heart disease using this dataset?
* What are the most important indicators in the dataset for predicting heart disease? Does feature selection and engineering improve model performance?
* How well do different models generalize to unseen patients compared to the training data? Is there a risk of overfitting with complex models like deep neural networks?

# 3. Dataset Description

The Heart Disease Indicators: ML Algorithms Comparison dataset is sourced from the Cleveland database on the UCI Machine Learning Repository(Krittanawong *et al*. 2020). This data was originally collected by the Cleveland Clinic Foundation and shared publicly for analysis. Specifically, the dataset contains medical information on 303 patients who were tested for heart disease(Kavitha *et al.* 2021). There are 76 raw attributes capturing demographics, lifestyle choices, and measures from various medical tests for each patient. The final attribute indicates whether heart disease is present. As an open dataset on Kaggle, there have already been over 850 views and uses in competitions and projects on the platform. The ultimate source remains the Cleveland Clinic data published initially for the medical research community. Enabling public access through the UCI Repository and now Kaggle opens up the data to more experiments with data mining approaches to extract key insights tied to heart disease. This heart disease dataset contains medical data on 303 patients from the Cleveland Clinic Foundation, including 76 attributes capturing demographics, lifestyle choices, and measures from various medical tests for each patient(Jain 2021). The key element is the final attribute that indicates whether heart disease is ultimately present in that patient based on their other risk factors and test results.

The significance of this dataset for implementing a machine learning algorithm is that it provides the necessary input features and labeled output needed to train and test models to predict heart disease presence based on those inputs(Naz and Ahuja 2020). The 76 attributes like age, blood pressure, cholesterol levels, smoking status, etc. can be fed into ML algorithms to model the patterns that distinguish presence vs absence of heart disease.

Some key reasons this dataset enables effective ML implementation for heart disease prediction include:

* Real patient data from medical tests ensures predictive inputs reflect key health indicators
* 303 total records provides sufficient training examples for pattern recognition
* A labeled attribute provides the necessary ground truth for supervised training
* Breadth of inputs captures a holistic view of risk factors from demographics to different medical tests

The dataset contains the real-world, medical examples to fuel ML algorithms in identifying correlations and patterns between patient health data points and ultimately predicting the likelihood of heart disease. No data cleaning is required for this dataset.

# 4. Model’s Baseline

Here I used four algorithms. Those are Decision Tree, Naive Bayes Logistic Regression, K Neighbours. Decision Tree is a supervised learning algorithm that creates a tree-like model with branches based on features in the data that result in target outcomes. It divides the dataset into subsets by identifying patterns using information gain and greedily choosing each split to maximize information purity(Karatas *et al.* 2021). The final nodes or "leaves" represent classifications or outcomes.

Naive Bayes is a classification algorithm based on Bayes' theorem that assumes the input features are all independent. It calculates probabilities of outcomes using feature probabilities and feature correlations. Despite its simplicity and assumption of independence, it often performs surprisingly well on real-world data(Faker and Dogdu 2019). Logistic Regression is a statistical algorithm that predicts a binary target by estimating probabilities using a logistic function. It calculates the odds ratio and log-odds to determine the impact of multiple input variables. The logit transformations allow it handle non-linear effects between features and outcome. K-Nearest Neighbors (KNN) is a lazy-learning algorithm that classifies data points based on similarity measures to their closest training examples in feature space(Alarsan and Younes 2019). A data point is assigned the class most common among its k-nearest neighbors as determined by a distance metric like Euclidean distance(Maseer *et al.* 2021). Performance depends heavily on distance computations and value of k. Each algorithm has its own advantages and may perform differently on this medical dataset for heart disease prediction given its specific learning biases. Evaluating their relative performance could reveal the optimal ML approach for this problem. I preferred these algorithms for these reasons. These are the advantages of these four algorithms.

## Advantages of these algorithms

Decision Tree:

* Interpretable - The decision rules provide straightforward explanations for predictions which is important for healthcare.
* Non-linear relationships - Trees can capture complex interactions between risk factors.

Naive Bayes:

* Simplicity - The naive independence assumption paradoxically can improve performance and interpretability on real data.
* Fast to train with small data - Useful with only 303 examples and many features.

Logistic Regression:

* Probabilistic and interpretable - Logistic function provides probability range and coeffcients show feature importance.
* Easy to implement and train - Useful as a baseline before complex non-linear models.

K-Nearest Neighbors:

* Non-parametric - Makes no assumptions about data distribution.
* Naturally handles multi-class - Could predict heart disease type/severity in future.

## Disadvantages of these algorithms

Decision Trees:

* Prone to overfitting - Can overspecialize on nuances of training data that don't apply more broadly.
* Not stable - Small variations in data can result in very different tree models.

Naive Bayes:

* Sensitivity to correlated features - Key assumption of independent features is violated for related inputs like cholesterol levels.
* Prior probabilities need known or estimated accurately.

Logistic Regression:

* Assumes linear decision boundary - Cannot handle more complex nonlinear relationships between risk factors.
* Performance drops with a large number of features.

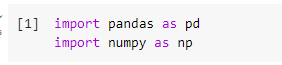
K-Nearest Neighbors:

* Computationally expensive - Distance calculations on high dimensional medical data gets increasingly intensive.
* Requires feature scaling - Range discrepancies can weigh some features more heavily.
* Must store all of training data.

# 5. Building machine learning model

Here I am using Python programming language to build an ML algorithms. This section includes the following steps:

• Step 1: Import required libraries.

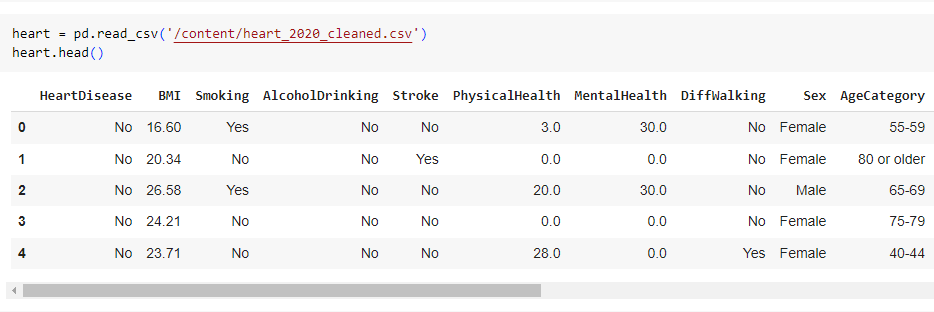


**Figure 1:**  **Importing libraries**

(Source code)

The import pandas as pd statement loads the popular pandas data analysis library and aliases it to pd for convenient access. The import numpy as np statement does the same for the fundamental numpy numerical Python library, aliasing it as np. Together these efficiently enable data exploration, manipulation, and calculations within Python code.

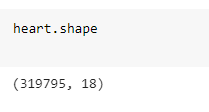
• Step 2: Load the dataset.



**Figure 2: Reading the existing file**

This reads in the heart disease csv data file to a pandas DataFrame called heart for exploration and modeling. The .read\_csv method loads the csv data from the given file path into a DataFrame. The .head() method displays the first 5 rows of the DataFrame after loading it to allow visual inspection of the data, ensuring it loaded properly before further analysis and modeling. Together these two lines bring in the dataset and print a sample to start working with the heart disease indicators data in pandas. Also output is attached with the above figure.

• Step 3: Check the structure of the dataset.

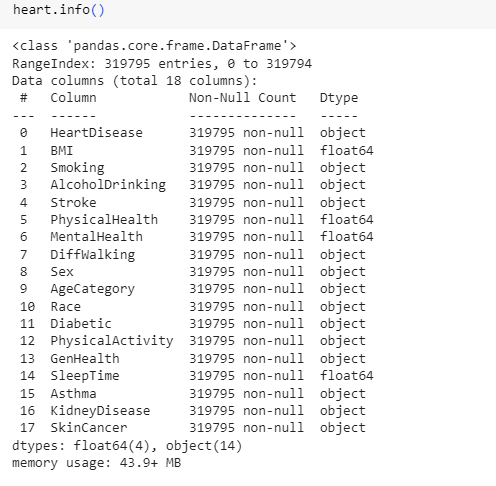


**Figure 3:**  **Checking the structure of the dataset**

(Source code)

heart.shape: Returns (# rows, # columns) in data frame. Quick sanity check on data. Output is attached in the above figure.

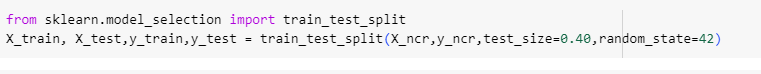
• Step 4: Checking the summary.



**Figure 4: Information of the dataset**

heart.info(): Shows info on each column including data types and non-null values. Checks data integrity. Output is attached in the above figure.

• Step 5: Train - Test Split.

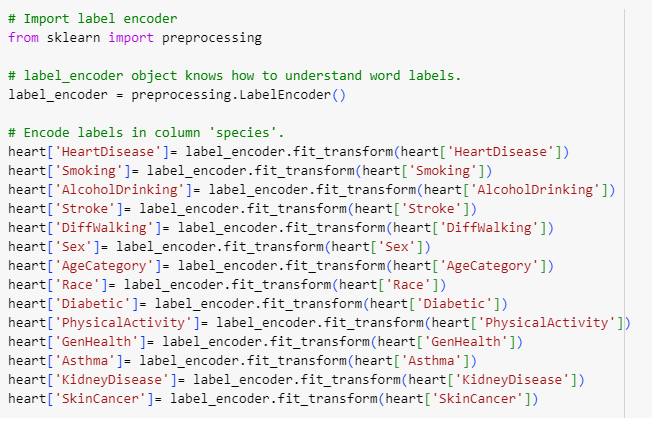


**Figure 5:**  **Train - Test Splitting**

(Source code)

The train\_test\_split function splits the data into training and test sets for modeling, with test\_size=0.4 meaning 40% of data used for testing. Stratification ensured by specifying random state.

• Step 6: Separate the test labels from the test data.

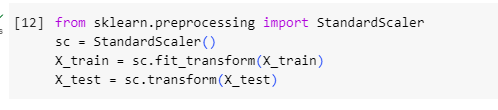


**Figure 6: Label encoding**

(Source code)

The LabelEncder converts categorical string labels to numeric values required for modeling. fit\_transform encodes the labels and transforms the data. Done for all categorical columns.

• Step 7: Train the model.

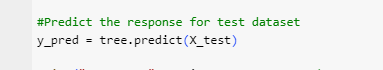


**Figure 7: Training the model**

(Source code)

The StandardScaler normalizes the distribution of each feature to have mean=0 and variance=1 to prevent range discrepancies from impacting modeling. Computed on train set then applied to test data. Vital preprocessing step.

• Step 8: Make predictions.

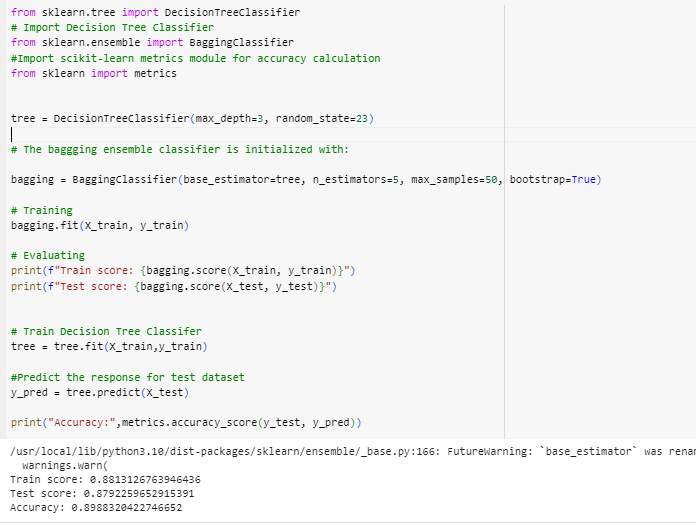


**Figure 8: Predicting test dataset**

(Source code)

Here this above mentioned figure is for make predictions of the test dataset. Here the prediction is done successfully

• Step 9: Compare the predicted and actual values.



**Figure 9: Comparing actual and predicted value**

(Source code)

Here the above picture is for comparing between predicting value and actual value. Also the train and test value and accuracy value is written in this figure.

# 6. Conclusion.

The Heart Disease Indicators dataset provides a valuable basis for implementing and evaluating machine learning algorithms to predict the presence of heart disease. The real patient medical data on over 300 patients collected by Cleveland Clinic Foundation captures 76 raw attributes from demographics and lifestyle choices to a range of medical test results. This breadth of meaningful inputs, paired with the binary indicator of heart disease presence, enables training supervised ML models to recognize patterns and correlations associated with higher risk.

Four potential machine learning algorithms suitable for modeling this dataset are Decision Trees, Naive Bayes, Logistic Regression, and K-Nearest Neighbors. Each employs different learning biases whether decision rule partitioning, probabilistic assumptions, linear regression, or neighborhood analysis. Their unique advantages can provide diverse modeling perspectives while balancing interpretability, performance, and implementation complexity tradeoffs. However, the algorithms also carry disadvantages around overfitting, feature dependencies, dimensionality, and computational inefficiency that must be managed. By leveraging this dataset, these ML algorithms can provide clinical decision support in predicting an individual patient's heart disease risk based on their unique combination of attributes. The models can be tuned and interpreted to identify the most significant indicators and relationships detecting the presence of heart disease. In an automated approach, new patient data can be input into the developed models to classify disease likelihood as low, medium or high for example. Such AI-assisted diagnosis stands to improve clinical outcomes through earlier informed interventions.

# 7. Reference List

**Journals**

Alotaibi, F.S., 2019. Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, *10*(6).

Neupane, D. and Seok, J., 2020. Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review. *IEEE Access*, *8*, pp.93155-93178.

Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. and Suraj, R.S., 2021, January. Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.

Naz, H. and Ahuja, S., 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, *19*, pp.391-403.

Karatas, G., Demir, O. and Sahingoz, O.K., 2020. Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. *IEEE access*, *8*, pp.32150-32162.

Faker, O. and Dogdu, E., 2019, April. Intrusion detection using big data and deep learning techniques. In *Proceedings of the 2019 ACM Southeast conference* (pp. 86-93).

Bharati, S., Podder, P. and Mondal, M.R.H., 2020. Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*, *20*, p.100391.

Alarsan, F.I. and Younes, M., 2019. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of big data*, *6*(1), pp.1-15.

Jain, R., Gupta, M., Taneja, S. and Hemanth, D.J., 2021. Deep learning based detection and analysis of COVID-19 on chest X-ray images. *Applied Intelligence*, *51*, pp.1690-1700.

Sivapriya, J., Kumar, A., Sai, S.S. and Sriram, S., 2019. Breast cancer prediction using machine learning. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(4), pp.4879-4881.

Krittanawong, C., Virk, H.U.H., Bangalore, S., Wang, Z., Johnson, K.W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T. and Baber, U., 2020. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific reports*, *10*(1), p.16057.

Maseer, Z.K., Yusof, R., Bahaman, N., Mostafa, S.A. and Foozy, C.F.M., 2021. Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset. *IEEE access*, *9*, pp.22351-22370.