

Assignment Stage 1: Report

Data Science for Innovation

Submitted By-

Kushal Ahuja (14191922)

## Table of Contents

[Section 1: Problem 3](#_bookmark0)

[Section 2: Literature review 4](#_bookmark1)

[Section 3: Approach 6](#_bookmark2)

[Section 4: Data 7](#_bookmark3)

[References 10](#_bookmark4)

# Section 1: Problem

Heart attack analysis and prediction is still a critical topic in healthcare since heart disease is one of the main reasons for mortality worldwide. Correctly identifying people with a risk of a heart attack is critical for avoiding heart disease and lowering healthcare expenditures. Age, gender, family background, smoking, diabetes,

hypertension, blood cholesterol, obesity, and lack of physical activity are all risk factors for heart disease. Algorithms based on machine learning can examine these risk variables and forecast the possibility of a patient having a heart attack. This enables healthcare practitioners to implement guidelines such as lifestyle

modifications, medicines, or surgery.

There are various advantages of using algorithms based on machine learning for heart attack investigation and prediction. Sophisticated algorithms can scan through huge quantities of patient data to uncover complicated correlations and trends between possible causes and the chance of a cardiac event. This may result in more accurate forecasts and better outcomes for patients. Moreover, machine learning algorithms may be programmed on massive datasets, permitting continuous quality improvement in predictive performance over a duration.

Cardiac arrest prediction algorithms can also help insurers understand the risks involved with covering certain clients and alter rates accordingly. This could lead to decreased expenses for both clients and insurance. Yet, there are some specific ethical problems with using patient information to predict health problems. To guarantee that patients are totally conscious of how personal data is utilized and to secure their rights, positive evaluations of data and informed authorization are required.

In addition, incorporating heart disease research and prediction into clinical practice has the possibility of

improving patient outcomes while also reducing healthcare expenditures. Early identification of those in high danger of cardiovascular disease, for example, can allow for early intervention and preventative measures that can prevent or postpone the development of heart disease and its effects. Furthermore, using algorithms for machine learning to aid with diagnostic and treatment decisions has the potential to increase the efficiency and efficacy of healthcare delivery.

The research questions that will be answered in stage 2 of the project.

 Are there any particularly important variables that contribute to the risk of heart disease?

 Which machine learning algorithms work best for assessing and predicting heart attacks?

 In regards to risk indicators and predictive variables, how do various forms of heart disease differ?

 How could heart disease research and prediction be utilized to establish individual preventative and treatment plans?

 How can heart disease prediction and analysis be incorporated into clinical practice to enhance patient outcomes and save healthcare costs?

 What are the present limits of cardiovascular disease research and prediction approaches, and how will they be addressed in future research

# Section 2: Literature review

**Paper 1**: Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, *64*(5), 304–310.

<https://doi.org/10.1016/0002-9149(89)90524-9>

The paper discusses a new discriminant function model that classifies people who have coronary artery disease. The reliability of the model is tested by examining 3 groups of patients undergoing angiography in Ohio,

Budapest, and Zurich. The probabilities from the Cleveland algorithm were compared with CADENZA (a

derived Bayesian algorithm published by medical studies called cadenza). Both algorithms overpredicted the probability of disease at American and Hungarian centers. Although the overestimations from the Bayesian models were very high compared to the discriminant function. It is observed that while using the Bayesian

model interdependencies of symptoms generates overconfidence in the model, in turn leading to overestimation.

**Paper 2:** Basheer, S., Mathew, R. M., & Devi, M. S. (2019). Ensembling Coalesce of Logistic Regression

Classifier for Heart Disease Prediction using Machine Learning. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12), 127–133. <https://doi.org/10.35940/ijitee.l3473.1081219>

The paper describes heart disease prediction using logistic regression classifier in 4 stages. Various ensembling methods like Extra Trees Regressor, Ada boost regressor, Gradient booster regress, Random Forest regressor, and Ada boost classifier extract essential features. Each important feature from ensembling methods is filtered and fitted to logistic regression classifier to analyse the performance. The Performance analysis is performed with the performance metrics such as Mean Squared error (MSE), Mean Absolute error (MAE), R2 Score,

Explained Variance Score (EVS) and Mean Squared Log Error (MSLE). It was observed that after feature

scaling, the feature importance extracted from Ada Boost was found to be the most effective compared to other ensembling methods.

**Paper 3:** Albahr, A., Albahar, M. A., Thanoon, M. I., & Binsawad, M. (2021). Computational Learning Model for Prediction of Heart Disease Using Machine Learning Based on a New Regularizer. *Computational Intelligence and Neuroscience*, *2021*, 1–10. <https://doi.org/10.1155/2021/8628335>

The paper discusses a computational learning model using regularization. As the model is trained, the learning model starts to memorize the data, which causes generalization errors. These models perform well in training sets but drastically poorly in test sets. L1 and L2 regularizers are used to help decide which feature to include in the model as they downsize the less critical one and remove the less important one. The lasso tends to select one variable from each group but ignores the rest, thereby reducing the model's accuracy. Cross-entropy loss

function is used over other loss functions to evaluate the classifier because the output value of cross-entropy is between 0 and 1, making it simple to convert probabilistic values to either one class or the other using

thresholds. Also, compared to other loss functions, it easily converges into corresponding class values.

**Paper 4:** Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013, April 18). *A Heart Disease Prediction*

*Model using SVM-Decision Trees-Logistic Regression (SDL)*. International Journal of Computer Applications; Foundation of Computer Science. <https://doi.org/10.5120/11662-7250>

The paper uses data from Cleveland Heart Disease database to predict heart disease using SDL (SVM, Decision Tress, Logistic Regression) and compares based on accuracy, sensitivity, and specificity. Various methods are used to split the data, like - node splitting, gini-coefficient, K- fold cross-validation, and statistical deviance.

Gini-coefficient is the most commonly used splitting criterion as it works on the population diversity of the

attribute before splitting it. The algorithm recommends that classification rule (C- rule ) as they are of the form of if-else ladders and provide the most straightforward and most understandable way of expressing knowledge. It is hypothesized that a result with higher sensitivity and specificity but lower accuracy will be attained from the results of this model, which is, in itself, a highly efficient model.

**Paper 5:** Teena Varma, Gaurav Kanojia, Hemant Gosavi, Atharva Jadhav, VIjayram Kanojiya, "HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING &AMP; DATA MINING TECHNIQUE",

*IJRAR - International Journal of Research and Analytical Reviews* (IJRAR), E-ISSN 2348-1269, P- ISSN 2349- 5138, Volume.7, Issue 2, Page No pp.415-419, June 2020, <http://www.ijrar.org/IJRAR19W1179.pdf>

The paper compares various heart disease predictions (hdp) using different machine learning algorithms. For a good prediction, hdp requires at least 12 features. We can find feature importance using Random Forest as it uses both classification and regression. Random forest obtains the best observations by creating decision trees on randomly selected data samples. It was observed that Random Forest Classifier has an accuracy of 99.29% followed by Decision Tree Classifier at 96.07%. Data scaling and Cross-validation are important to increase the performance of Support Vector Machine, K-means, and Logistic Regression.

# Section 3: Approach

The approach for solving the problem of predicting risk of having the heart disease will consists of the following machine learning algorithms –

* **Logistic Regression** is a statistical method that models the probability of an outcome, since the dataset has more than one independent variables, we feel it would be a good approach to start with. We would select the appropriate input variables and optimize model's parameter to generate the best predictive accuracy possible. We would also be mindful of using the input

variables that are not correlated as that could lead to biased estimates, for this we would also be performing feature selection.

* **Decision Tree** is another supervised learning algorithm that we would be using to predict the presence of heart disease based on the input variables. Our focus would be to optimize the model by tuning the hyperparameters like the max depth of the tree, the minimum number of samples required to split a node and to measure the quality of the split. We would perform the final evaluation of the model on the test dataset by computing accuracy, precision and recall metrics.
* **Random Forest** is a machine-learning method for handling classification and regression issues. It consists of multiple decision trees, thus resulting in more accuracy. It would build

various decision trees using randomly selected subsets of the data. Our focus would be to fine- tune the model by tuning the hyperparameters like the number of trees, max depth of each tree, number of samples required to split a node, etc. We will use this in our approach and

perform a final evaluation of the test data by measuring different metrics like accuracy, macro F1-score, micro F1-score and confusion matrix.

* **Naive Bayes**, a probabilistic machine learning algorithm, is often used to tackle classification problems. The approach would be beneficial with a dataset like this one with 14 input

variables and dealing with a high-dimensional structure. We would split the dataset into a training set and test set, apply the Naive Bayes model to the training data, and then, in turn, use it to make the prediction. We would then judge the model's accuracy by comparing the output with the true values in the testing set.

* **Support Vector Machines** are a supervised learning algorithm that finds the best hyperplane to separate the different classes in the dataset. Even though SVM is suited more for binary

classification, we would modify it to handle multiple classes for this dataset. The approach we would use would be the one-vs-one strategy, in which we would train a separate binary

classifier for each pair of target classes and then choose the class that wins the most comparison.

# Section 4: Data

The Heart attack analysis & prediction dataset is the dataset for the classification of heart attacks; This data set has been acquired from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Heart%2BDisease) which is sourced from creators –

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., PhD. (*UCI Machine Learning Repository: Heart Disease Data Set*, n.d.)

[https://archive.ics.uci.edu/ml/datasets/Heart+Disease](https://archive.ics.uci.edu/ml/datasets/Heart%2BDisease)

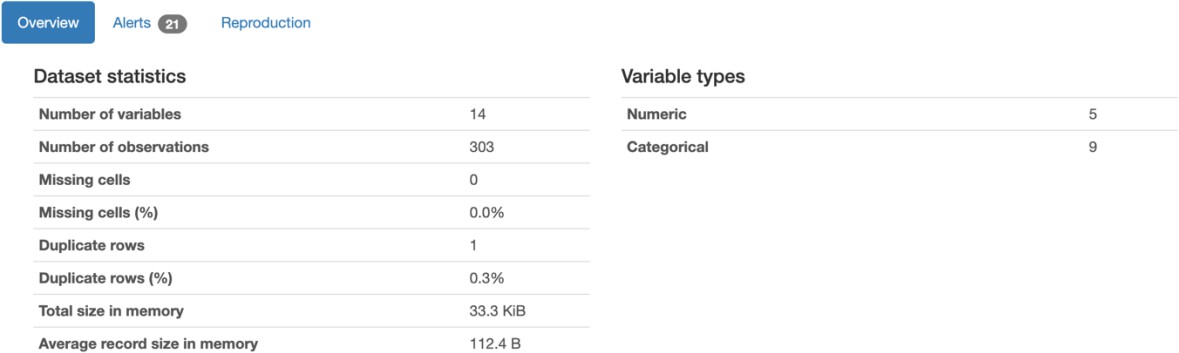
Dataset description

The dataset consists of 76 attributes, while all the research studies only describe employing a fraction of 14 attributes, i.e. age, sex, chest pain, blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate, Exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise, number of significant vessels and diagnosis of heart

disease status. The field of interest of this dataset is healthcare, as healthcare remains to be the topmost priority for any being; this data contributes awareness and crucial information about heart attack analysis.



(Source- *UCI Machine Learning Repository: Heart Disease Data Set*, n.d.)



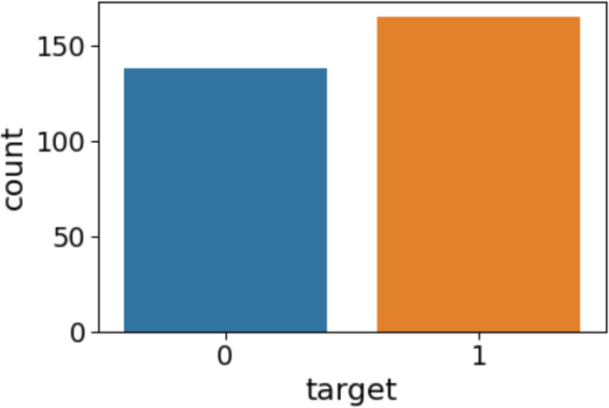
Heart Disease Dataset contains 303 instances of rows and 14 columns; rows in the dataset represent the patients, and columns demonstrate the demographics, medical data, and whether the person has heart disease or not.

Exploratory Data Analysis (EDA) is being performed on the dataset to understand the dataset and its attributes; it helps in achieving the solution to the problems by various steps. Visualisation in EDA helps achieve those goals and answers to the research question; EDA is also significant for conveying information to stakeholders.

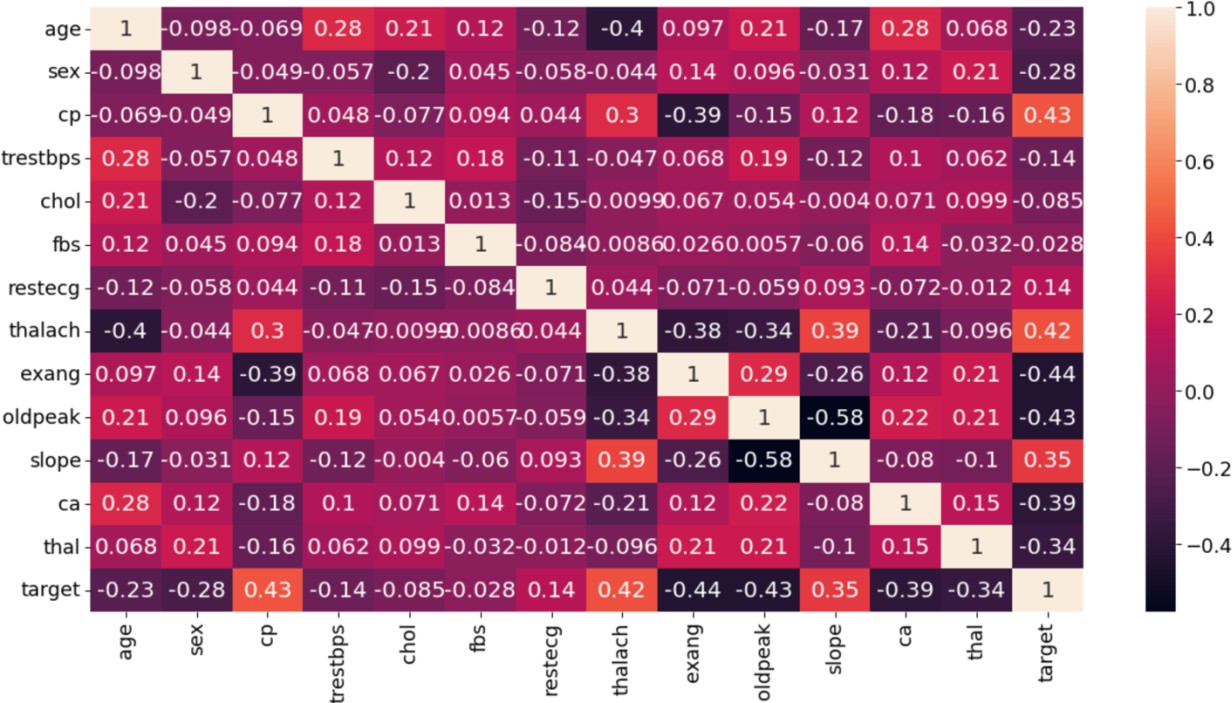
The quality of the dataset, including the outliers, missing values and duplicate values, was identified and treated through the process of EDA.

Various patterns are identified and analysed due to EDA, thus answering further research questions. EDA assists in evaluating the result of statistical methods or machine learning algorithms.

### Count plot of the variable target, which demonstrates the heart disease(0 as No and 1 as Yes)



Correlation matrix – The correlation matrix visualises the correlation coefficients between the variables; each cell in the table demonstrates the correlation between the two particular variables. (Bock, 2022). The information related to correlation can be interpreted by evaluating the colour of the cells; the darker colour denotes the higher correlation, while the light colour represents the lower correlation.

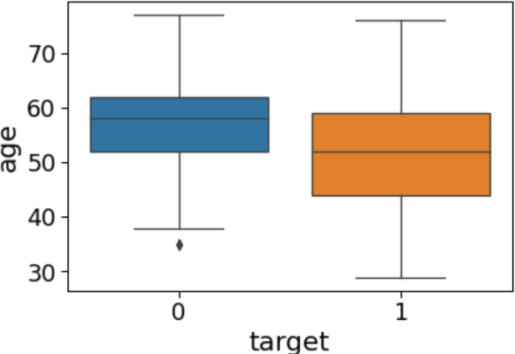


The above table shows the relation between all the variables and thus assists in identifying the pattern.

It is evident from the above illustration that cp (chest pain), thalach (maximum heart rate) and slope (the slope of the peak exercise) are the variables highly correlated to the target variable, i.e. heart disease, which states that

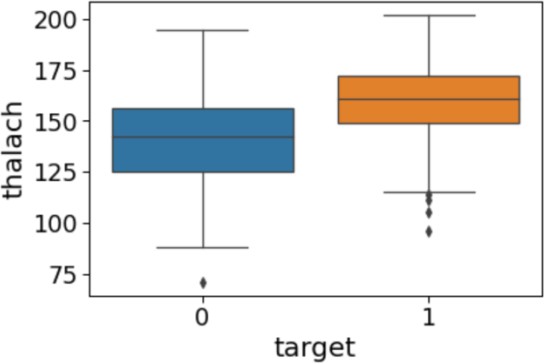
chest pain, maximum heart rate and slope of peak exercise contribute the most towards determining the existence of heart disease. In contrast, exang(exercise-induced angina), oldpeak(ST depression induced by exercise relative to rest) and ca(number of significant vessels) contribute the least towards determining the existence of heart disease.

### The relationship between age and target variable



The above box plot shows whether there is a relationship between the age of the person and the risk of having heart disease; it is evident that people aged between 45- 60 tend to have the risk of having heart disease.

### The relationship between thalach (maximum heart rate achieved) and target variable



The above box plot shows the relation of thalach with having the risk of heart disease, It is evident from the figure that individuals having higher heart rate i.e. 150 bpm-175 bpm, contributes more towards the risk of the presence of any heart disease.

 The tools used to clean and export the dataset are libraries named pandas and to make the visualisation, the seaborn library is being used.

# References

*UCI Machine Learning Repository: Heart Disease Data Set*. (n.d.). [https://archive.ics.uci.edu/ml/datasets/Heart+Disease](https://archive.ics.uci.edu/ml/datasets/Heart%2BDisease)

Bock, T. (2022, August 23). *What is a Correlation Matrix? - Displayr*. Displayr. <https://www.displayr.com/what-is-a-correlation-matrix/>

Gupta, P. (2018, June 20). *Decision Trees in Machine Learning - Towards Data Science*. Medium. <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>

*What is Logistic regression? | IBM*. (n.d.). <https://www.ibm.com/au-en/topics/logistic-regression>

*Introduction to Random Forest in Machine Learning*. (n.d.-b). Engineering Education (EngEd) Program | Section. [https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-](https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/) [learning/](https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/)