# Heart Disease Prediction using Machine Learning

Shubham Patel
*B.Tech. CSE*
*Ahmedabad University*
Ahmedabad, Gujarat, India
shubham.p2@ahduni.edu.in

Keyur Nagar
*B.Tech. CSE*
*Ahmedabad University*
Ahmedabad, Gujarat, India
keyur.n@ahduni.edu.in

Shrey Patel
*B.Tech. CSE*
*Ahmedabad University*
Ahmedabad, Gujarat, India
shrey.p2@ahduni.edu.in

***Abstract*—At present, the key factor responsible for mortality is Heart Disease. Heart Disease is among the most encountered disease in the Health Care Industry. It accounts for about one-third of the deaths occurring globally [1]. Not only its treatment is costly, but its also complicated. In most of the developing countries, its diagnosis isn't commonly available and is also spendthrifty. This paper compares different models for detecting heart disease using Machine Learning Algorithm. The dataset used is the Cardiovascular Disease Dataset which is published by Ryerson University on Kaggle. [2] The Heart Disease Dataset was used to train two different machine learning algorithms (K-Nearest Neighbours Classifier, and Logistic Regression). By introducing risk factor (feature extraction) - Body Mass Index (BMI), we have observed an increase in the accuracy. Feature selection was performed using Correlation-based Feature Selection (CFS) Algorithm to find out the important risk factors. Finally, Ensemble Classification Technique named Bagging is used to increase the model's accuracy by combining the results of several weak classifiers. This dataset takes into account 12 different attributes (11 attributes as input and 1 attribute as output) and 70,000 records for Heart Disease. We have used this dataset to train a model which will predict whether the patient has the disease or not. This paper summarizes some present research in this field and proposes the best performing algorithm (Bagged Logistic Regression) which has 78.01% accuracy in the prediction.**

***Index Terms*—Heart Disease, Machine Learning, Classification Techniques, K-Nearest Neighbors, Logistic Regression**

## I. Introduction

Heart Disease is a leading cause of death globally. Taking into account this alarming issue, strategic planning is needed to be enforced to measure the growing trend in the disease and to accordingly plan the preventive measures for both the individual and the population level. Ischemic Heart Disease and stroke are the major heart disease among all heart diseases. [1]

Researchers have found that Heart Disease normally occurs between middle age and older age. Factors accounting for heart disease are smoking habits, unhealthy diet, and physical inactivity. These factors trigger the risk factors such as diabetes, high cholesterolemia, and high blood pressure [1]. Although Heart Disease occurrence is common, it is preventable and hence becomes a priority for the Health Care Industry. Several studies suggests that the risk of cardiovascular diseases is reversible by lowering the level of the risk factors stated previously. Lowering the risk factors can delay the occurrence of the event or can reduce the severity and the occurrence of the event.

Heart Disease can reduce the quality of life and can lead to dependence on medications throughout the lifetime. This can result in negative outcomes in old age which include reduced mental ability, and physical ability. The disorders from the medications can also lead to premature death. Whereas the societal loss includes the impact on the family considering the monetary factor.

The use of Machine Learning in Heart Disease Detection has been discussed in several pieces of research paper. One such research piece includes the application of Artificial Intelligence for heart disease detection systems which improves the existing models performance which includes models provided by American College of Cardiology/American Heart Association for Cardiovascular Disease prediction and detection. [3]

## II. Literature Review

A paper named Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques [4] uses Naive Bayes, Decision Trees, and Neural Networks on the heart disease dataset. The paper presented the accuracy of Naive Bayes, Decision Trees, and Neural Networks algorithms as 90.74%, 99.62%, and 100% respectively for detecting heart disease. The authors added two more attributes namely smoking and obesity in the dataset. Adding the attributes resulted in more accuracy in the prediction because of their key role in cardiovascular diseases.

A paper named Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques [5] performs the research based on the technique of ensemble classification. This ensemble classification technique can be used to improve the accuracy of weak algorithms by the combination of multiple classifiers. The results showed an increase in the accuracy of prediction using ensemble classifier techniques such as boosting and bagging. The technique of feature selection was also implemented in the approach. The authors noted an increase of at most 7% in the prediction accuracy using this approach.

## III. Implementation

### A. Data Pre-processing

At first, the null values were removed from the dataset. Secondly some instances having outlier values of the ap_hi (Systolic Blood Pressure values $>$ 250 and $<$ 0) and ap_lo (Diastolic Blood Pressure values $>$ 200 and $<$ 0) were removed. The age feature was in days so it was also normalized to years. The Data Pre-processing step reduced the dataset to 68,975 instances.

### B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was carried out to get more insights on the dataset. The correlation analysis of 11 attributes with the output (12th attribute in our dataset) was performed. Removing the correlation is important because the correlation between the attributes can result in false predictions. All the attributes have a correlation value less than 0.41 with our output. Hence we can conclude that our attributes aren't highly correlated.

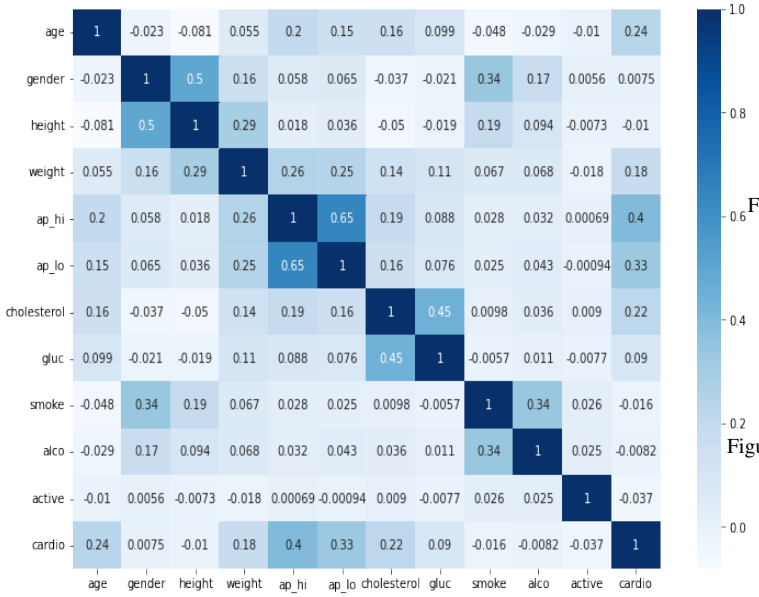


Figure 1. Correlation matrix of the Dataset

***Note**: From Fig. 3 to Fig. 9 the **orange color bar** indicates the **Heart Disease present** cases and the **blue color bar** indicates **Heart Disease absent** cases. **Also the plots are sized to fit the page limit so please zoom the pdf to see the plots clearly.***

- **Gender:** The gender attribute has two values: 1 (female) and 2 (male). It can be inferred from the Fig. 3 that there is not much difference between the cases for either gender.
- **Age:** From the Fig.4 is can be observed that the age attribute has significant effect on the heart disease cases. It can be inferred from that as the age increase the heart disease risk also increases.

1. **Age | Objective Feature |**
   **age** in int (days)
2. **Height | Objective Feature |**
   **height** in int (cm)
3. **Weight | Objective Feature |**
   **weight** in float (kg)
4. **Gender | Objective Feature |**
   **gender** in categorical code | 1: female, 2: male
5. **Systolic blood pressure | Examination Feature |**
   **ap_.hi** in int
6. **Diastolic blood pressure | Examination Feature |**
   **ap_lo** in int
7. **Cholesterol | Examination Feature |**
   **cholesterol** | 1: normal, 2: above normal, 3: well above normal |
8. **Glucose | Examination Feature |**
   **gluc** | 1: normal, 2: above normal, 3: well above normal
9. **Smoking | Subjective Feature |**
   **smoke** | binary | 0: non-smoker, 1: smoker
10. **Alcohol intake | Subjective Feature |**
    **alco** in binary | 0: non-alcoholic, 1:alcoholic
11. **Physical activity | Subjective Feature |**
    **active** | binary | 0: inactive, 1: active
12. **Presence or absence of cardiovascular disease | Target Variable |**
    **cardio** | binary | 0: absent, 1: present

Figure 2. Feature Information of Cardiovascular Disease Dataset

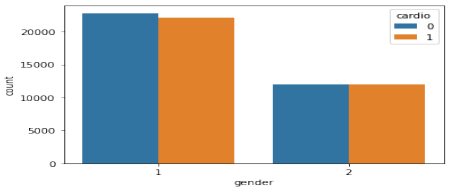


Figure 3. Plot of count for each gender with or without heart disease

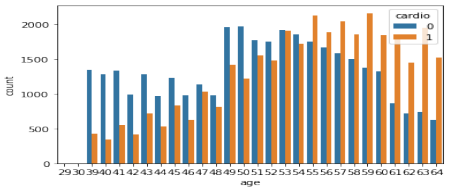


Figure 4. Plot of count for age with or without heart disease

- **Cholesterol:** The cholesterol attribute has 3 levels and the details about the each level can be found in Fig. 2. Although the level-1 cholesterol has more people without heart disease, however it can inferred from Fig. 5 that as the level of cholesterol increases the heart disease cases also increases.
- **Glucose:** The gluc attribute has 3 levels and the details about the each level can be found in Fig. 2. It can be inferred from the Fig. 6 that as the glucose level rises, the heart disease cases also increase.

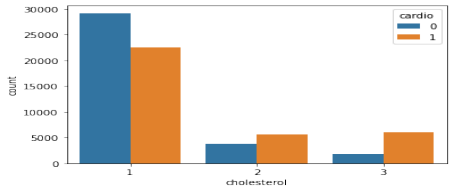


Figure 5. Plot of count for each level of cholesterol with or without heart disease

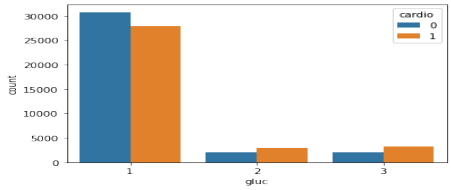


Figure 6. Plot of count for each level of glucose with or without heart disease

- **Smoking:** The smoke attribute has two values: 0 (non-smokers) and 2 (smokers). It can be inferred from the Fig. 7 that there is not much difference on the heart disease cases with regard to smoking.

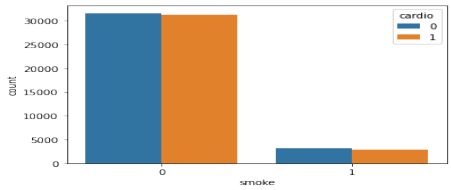


Figure 7. Plot of count for each smoker class with or without heart disease

- **Alcohol Consumption:** It can be inferred from Fig. 8 that the alco attribute has no effect on the heart disease cases.

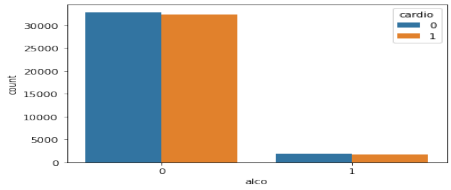


Figure 8. Plot of count for each alcohol consumer class with or without heart disease

- **Physical Activity:** The active attribute has two values: 0 (in-active) and 2 (active). It can be inferred from the Fig. 9 that there are more people with heart disease in physically in-active class compared to the physically active class.

### *C. Feature Extraction and Selection*

To increase the accuracy of our model, we introduced a new risk factor called Body Mass Index (BMI). The BMI

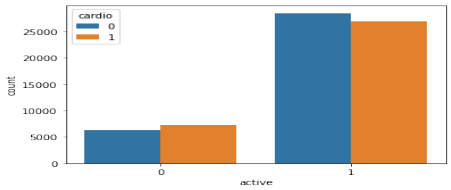


Figure 9. Plot of count for each physical activity class with or without heart disease

can be calculated from the information available in the dataset.

- **Body Mass Index (BMI):** BMI can be calculated from the height and weight of the individual. The weight should be in kgs. and the height should be in meters squared.

$$BMI = \frac{weight}{height * height}$$

Hence, the height and weight information is captured in our new BMI feature, so we can remove the height and weight features from the dataset and add the new BMI feature. We will now perform feature selection in which we will filter the dataset by considering only those attributes or that increase the risk of the disease.

We have used an open-source software named WEKA (Waikato Environment for Knowledge Analysis) for the Feature Selection developed by The University of Waikato, New Zealand. We have used a Correlation-based Feature Selection (CFS) algorithm to perform Feature Selection on our new data.

- **CFS Algorithm:** CFS is a simple filter algorithm that ranks feature subsets according to a correlation-based merit evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. The implementation of CFS allows the user

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Figure 10. Equation of Merit for CFS Algorithm

to choose from three heuristic search strategies: forward selection, backward elimination, and best first. Forward selection begins with no features and greedily adds one feature at a time until no possible single feature addition results in a higher evaluation. Backward elimination begins with the full feature set and greedily removes one feature at a time as long as the evaluation does not degrade. Best first can start with either no features or all features. [6]

Out of 10 features, we desire to select the 7 best performing features which account for maximum prediction accuracy. So in the CFS algorithm, we feed 7 as the number of output features. By running the CFS algorithm in WEKA, we got the Merit of Selection as 0.445. The CFS algorithm selected the following features: *age, ap_hi, ap_lo, cholesterol, gluc, active, and bmi*. These 7 features capture the maximum information of the dataset.

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 7 node expansions
        Total number of subsets evaluated: 58
        Merit of best subset found:    0.445

Attribute Subset Evaluator (supervised, Class (numeric): 10 cardio):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,3,4,5,6,9,11 : 7
                     age
                     ap_hi
                     ap_lo
                     cholesterol
                     gluc
                     active
                     bmi
```

Figure 11. Output window of WEKA for CFS Algorithm

### D. Disease Prediction

The data was split into two parts namely training data (80%) and testing data (20%). The K-Nearest Neighbour algorithm was implemented on the training data and testing data. The plot of misclassification suggests that the value of k (neighbors) around 80 would result in the minimum gap between the training and testing error. The accuracy of the KNN model calculated at k = 80 is 72.65%.

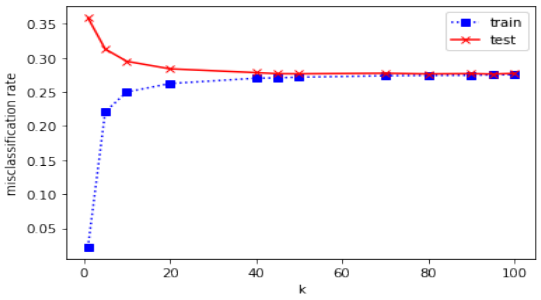


Figure 12. Plot of misclassification vs. k

The Logistic Regression algorithm was implemented on the training data and testing data. The accuracy of the Logistic Regression model calculated on the testing data comes to be 72.85%.

### E. Ensemble Classification

Ensemble Classification is a Machine Learning technique in which several models are trained to solve the same problem, and their results are combined to produce better accuracy. The Ensemble technique we used was Bagging.

- **Bagging:** Bagging, commonly known as Bootstrap aggregating, is an ensemble classification technique that improves the accuracy and performance of machine learning algorithms. It is widely used to deal with the problem of bias-variance trade-off and reduce the model's variance. Bagging avoids overfitting and can be used with classification and regression models.

We implemented Bagging with Logistic Regression as a base model, which increased the model accuracy by 5%. Hence the final accuracy of our model increased to 78.01%.

## IV. Results

$$Acc. = \frac{TruePos. + TrueNeg.}{TruePos. + TrueNeg. + FalsePos. + FalseNeg.}$$

*Acc. indicates Accuracy*

### A. K-Nearest Neighbours Algorithm

The testing data (20% of 68,975) contained 13,795 records and our KNN Algorithm correctly predicted 9978 records (True Pos. + True Neg.). The accuracy of our model can be defined as:

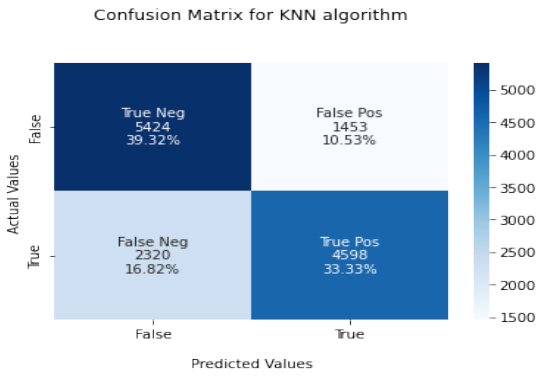


Figure 13. Confusion Matrix for KNN Algorithm

$$Acc. = \frac{5424 + 4598}{5424 + 4598 + 2320 + 1453} = 72.65\%$$

### B. Logistic Regression Algorithm

The testing data (20% of 68,975) contained 13,795 records and our Logistic Regression Algorithm correctly predicted 9888 records (True Pos. + True Neg.). The accuracy of our model can be defined as:

$$Acc. = \frac{5396 + 4654}{5396 + 4654 + 2264 + 1481} = 72.85\%$$

### C. Bagged Logistic Regression Algorithm

The testing data (20% of 68,975) contained 13,795 records and our Logistic Regression Algorithm correctly predicted 9888 records (True Pos. + True Neg.). The accuracy of our model can be defined as:

$$Acc. = \frac{6284 + 6391}{6284 + 6391 + 1776 + 1798} = 78.01\%$$

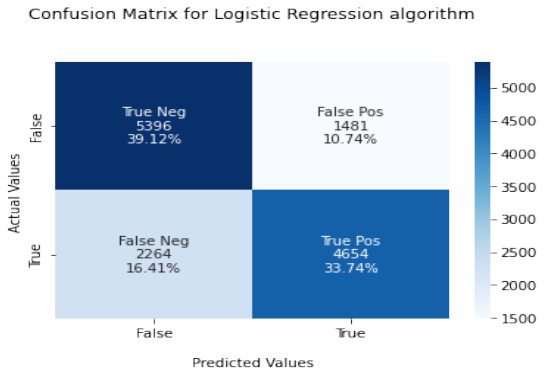


Figure 14. Confusion Matrix for Logistic Regression Algorithm

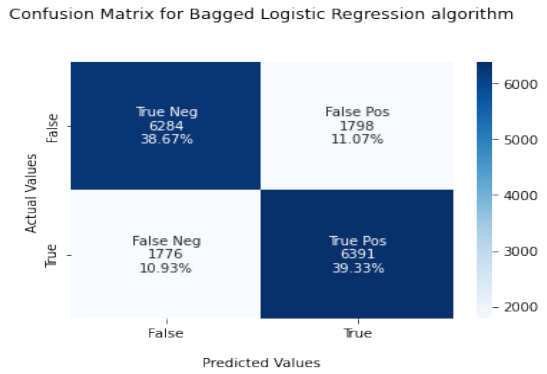


Figure 15. Confusion Matrix for Bagged Logistic Regression Algorithm

| Classifier | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| KNN | 72.65 | 75.99 | 66.46 | 70.91 |
| Logistic Regression | 72.85 | 75.86 | 67.27 | 71.31 |
| Bagged Logistic Regression | 78.01 | 78.04 | 78.25 | 78.15 |

Table I
PERFORMANCE MATRIX FOR THE IMPLEMENTED ALGORITHMS

## V. Deployment

The Ensembled Machine Learning model was deployed on Heroku (a cloud platform for hosting web applications). The model was saved in a pickle file with the help of the Pickle() module of Python. The deployed application uses this pickle file to predict the output when new data is provided. Our web application contains 7 features, namely: *age, ap_hi, ap_lo, cholesterol, gluc, active, and bmi*. Here the user has to simply enter the requested inputs, and the model provides its prediction of whether the person is Healthy or Unhealthy.
Link to: Web Application

## VI. Conclusion

This paper compared the two algorithms namely K-Nearest Neighbours and Logistic Regression to predict Heart Disease using the Cardiovascular Disease Dataset. After testing the algorithms, the KNN algorithm resulted in 72.65% accuracy whereas the Logistic Regression algorithm resulted in 72.85% accuracy. Adding one risk feature - BMI and

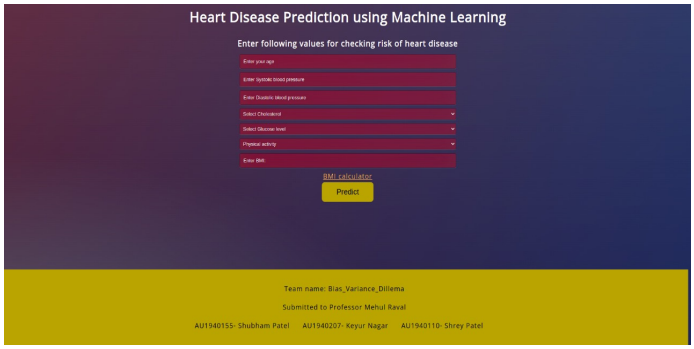


Figure 16. Image of the Web Application deployed on Heroku

performing Feature Selection using the Correlation-based Feature Selection (CFS) algorithm increased the model accuracy. Finally, ensemble classification technique such as Bagging was used with Logistic Regression as a base model. The Bagged Logistic Regression model increased the accuracy by 5%, increasing the model accuracy to 78.01%. Hence we can conclude that the Bagged Logistic Regression algorithm best fits our dataset.

## GitHub Repository

## References

[1] Who.int. 2022. Cardiovascular diseases (CVDs). [online] Available at:https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
[2] Ryerson University, "Cardiovascular Disease dataset," 2019. [Online] Available: www.kaggle.com.
[3] K, Vanisree Jyothi, Singaraju. (2011). Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. International Journal of Computer Applications. 19. 10.5120/2368-3115.
[4] Chaitrali S Dangare and Sulabha S Apte. Article: Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications 47(10): pp.44-48, June 2012.
[5] C.B.C. Latha, S.C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine, Unlocked 16, pp.100203, 2019.
[6] M. A. Hall, Correlation-based Feature Selection for Machine Learning- University of Waikato. [Online]. Available: https://www.cs.waikato.ac.nz/ mhall/thesis.pdf. [Accessed: 22-Apr-2022].