Department of Computer Science & Information Systems College of Liberal Arts and Sciences  
Bradley University



**CS 562 Machine Learning**

(Department of Computer Science)

**PROFESSOR: Babu k Baniya**

**PROJECT REPORT**

**Data Set : Heart Disease Prediction**

**Submitted By: Submitted to:**

Bhargavi Peta (338443) Babu k Baniya

**TABLE OF CONTENTS**

1. Abstract…………………………………………………………………….3
2. **Introduction………………………………………………………………** 3
3. **Methodology………………………………………………………………4**
4. **Experiments……………………………………………………………….** 5
5. **Conclusions………………………………………………………………….7**
6. **References …………………………………………………………………..8**

# Abstract:

In this research paper, we will analyze the use of machine learning, specifically binary logistic regression, for identifying predictors of heart diseases and predicting overall risks is a promising approach to address the urgent need for early detection and treatment of asymptomatic diseases, such as cardiovascular diseases, which are overwhelming national and corporate budgets in healthcare expenditures. Python, along with JupyterLab, is a popular and powerful tool for data analysis and machine learning tasks. It provides a wide range of libraries for data manipulation, visualization, and model building, making it suitable for conducting the data analysis required for this study.

# Introduction:

Machine learning is employed extensively throughout a wide range of industries, including the healthcare sector. Artificial intelligence (AI) applications such as machine learning enable systems to automatically learn from their experiences and get better over time without having to be explicitly programmed Additionally, machine learning is the process of utilizing algorithms to analyze data, learn from it, and then predict or determine anything about the outside world . Regression and classification are the two main kinds of issues that machine learning is frequently used to address.Regression algorithms are typically employed for numerical data, while binary and multicategory issues are examples of classification problems .Algorithms for machine learning are Unsupervised learning and supervised learning are two further categories [5]. Unsupervised learning does not use predefined labels and instead aims to infer the natural structures within the dataset In essence, supervised learning is carried out by employing prior information in output values. As a result, choosing a machine learning algorithm requires careful consideration.

Chart

Description automatically generated

Binary logistic regression is a type of supervised machine learning algorithm used for classification tasks, where the goal is to predict the probability of an outcome belonging to one of two classes. In this study, the classes would be "heart disease" and "no heart disease", and the logistic regression model would be used to identify the most significant predictors or risk factors associated with heart diseases. Once the model is built, it can be validated using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess its performance in predicting heart diseases. The results of the analysis can then be interpreted to identify the most significant predictors of heart diseases

# Logistic Regression

Logistic regression is a one of the machine learning classification algorithm for analyzing a dataset in which there are one or more independent variables that determine an outcome and also categorical dependent variable . Linear regression uses output in continuous numeric whereas logistic regression transforms its output using the logistic sigmoid function to return a probability Values

The logistic function in linear regression is a type of sigmoid, a class of functions with the same specific properties. Sigmoid is a mathematical function that takes any real number and maps it to a probability between 1 and 0.

The formula of the sigmoid function is:

Chart

Description automatically generated

According to the given data set, 1 indicates the high risk of 10-year future coronary heart disease and 0 indicates non or no heart risks. The independent variables n in the logistic model as x1, x2, x3……., xn ( ) Logisticregressionachievesthisbytakingthelogoddsofthe event ln(P/1−P), where, P is the probability of event which is risk of CHD. Therefore, P always lies between 0 and1

# 3.Methodology

# Input Variables:

# Research in the field has identified the following as ****important risk factors**** for heart disease and other chronic illnesses like diabetes (not in strict order of importance):

# Table:Input Variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Data type |
| Heart DiseaseorAttack | Risk of Heart attack | Binary |
| High BP | Adults who have been told they have high blood pressure by a doctor | Continuous |
| High Chol | Adults who have been told they have high cholestrol by a doctor | Continuous |
| CholCheck | Cholesterol check within past five years | Continuous |
| BMI | Body Mass Index (BMI) | Continuous |
| Smoker | Smoke or not? | Nominal |
| Stroke | Whether previously had a stroke? | Nominal |
| Diabetes | Whether had Diabeties? | Nominal |
| PhysActivity | Adults who reported doing physical activity or exercise during the past 30 day | Nominal |
| Fruits | Consume Fruit 1 or more times per day | Continuous |
| Veggies | Consume Vegetables 1 or more times per day | Continuous |
| HvyAlcoholConsump | Heavy drinkers (adult men having more than 14 drinks per week | Nominal |
| AnyHealthcare | Medicare, or Indian Health Service | Nominal |
| NoDocbcCost | when you needed to see a doctor but could not because of cost? | Nominal |
| GenHlth | general your health | Nominal |
| MenHlth | mental health, which includes stress, depression, and problems with emotions, | Nominal |
| PhysHlth | physical health, which includes physical illness and injury | Nominal |
| DiffWalk | difficulty walking or climbing stairs? | Nominal |
| Sex | indicate sex of respondent. | Nominal |
| Age | Fourteen-level age category | Continuous |
| Education | highest grade or year of school you completed | Continuous |
| Income | annual household income from all sources: | Continuous |

**Loading Data and Other Required Libraries**

It has loaded the heart prediction data using Framingham CSV file into Jupiter Lab in Order to build the logistic regression model. In addition to that, required libraries which used as supportive applications are loaded. It has removed the education field from the database.

**Dependent Variable** : Heart Attack

**Independent Variables**: HighBP,HighChol,CholCheck,BMI,Smoker,Stroke,Diabetes,PhyActivity,Fruits,Veggies,HighAlcoholConsump,AnyHealthCare,NODoctorCost,GenHealth,MentalHlth,PhyHlth,DiffWalk,Sex,Age,Education,Income

Graphical user interface, text, application, email

Description automatically generated

No. of instances and attributes :

A picture containing diagram

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

According to the above data, there are 229787 patents with no heart disease and 23893 patients with risk of heart disease.

**Implementing Logistic Regression**

The following outcomes are used to indicate the logistic regression. Logistic regression is mainly used to for prediction and also calculating the probability of success.

Table

Description automatically generated

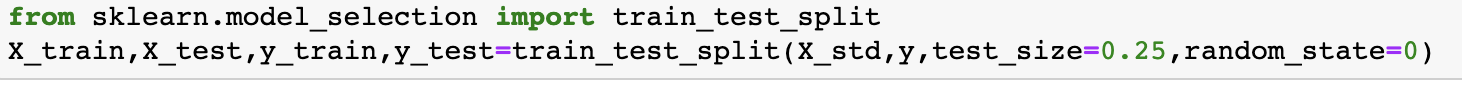
**correlation between our independent variables**

Application, table, Excel

Description automatically generated

**Training and testing sets**

Data set was separated into training and testing sets for evaluation process. This has been done using scikitlearn library.



Splitting the data set into testing and training data set

Text

Description automatically generated

**4.Experiments**

**Confusion Matrix Outcomes**

This has used to indicate the summary of prediction results including correct and incorrect on a classification problem. Further, this was used to not only errors but also types of errors. The segments of the confusion matrix indicate the following parameters.

• True Positives (TP): cases which are predicted yes (they have the disease), and they do have the disease.

• True Negatives (TN): cases which are predicted no, and they do not have the disease.

• False Positives (FP): cases which are predicted yes, but they do not actually have the disease (Type I error).

• False Negatives (FN): cases which are predicted no, but they actually do have the disease (Type II error). The following outcome indicates the matrix of the dataset.

The following outcome indicates the confusion

Chart

Description automatically generated

According to the outcome of the confusion matrix, Correct predictions (56751+830) =57581

Incorrect predictions (5195+644) =5839

Therefore, True Positives:830

True Negatives:56751

False Positives:644

False Negatives:5195

The accuracy of the model is 0.9079

**ROC Curve**

The ROC Curve is a simple plot which used to visualize the performance of a binary classifier. Further, this shows the tradeoff between the true positive rate and the false positive rate of a classifier for various choices of the probability threshold.

Graphical user interface, application

Description automatically generated

Good classification accuracy models should have significantly more true positives than the false positives at all thresholds. Area Under the Curve(AUC) quantifies the model classification accuracy

**Recall:**

Recall is the ratio (TP/(TP+FN))

0.1377

**Precision:**

TP/(TP+FP)

0.56309

**F1 Score:**

0.22136

F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall)

**Conclusion:**

All attributes selected after the elimination process show P-values lower than 0.5% and thereby suggesting significant role in the Heart disease prediction.

According to the logistic regression outcome, men are more suspectable to heart disease than women Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease

• Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of 'good cholesterol(HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%)

• The model predicted with 0.9079 accuracy using logistic Regression . The model is more specific than sensitive. Overall model could be improved with more data

**References**

1. http://www.ijarp.org/published-research-papers/aug2019/Predicting-Heart-Diseases-In-Logistic-Regression-Of-Machine-Learning-Algorithms-By-Python-Jupyterlab.pdf

2. Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In NIPS 14, pp. 841–848.

3.https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression

4. https://towardsdatascience.com/logistic-regression-explained-593e9ddb7c6c