BUAN 6341.001 S22

APPLIED MACHINE LEARNING

Project Report—**Group 4**
Members: Neetu Joon, Priyanka Padmanabhan,
Snehaa Shri Hari, Mekonnen Woldehana, Bryan Xiao

Contents

Abstract

In this report, we address the problem of predicting heart disease—one of the leading causes of death for people of most races in the United States. One person dies every 36 seconds in the United States from cardiovascular disease (cdc.gov). *An estimated 16.3 million Americans aged 20 and older have Cardiovascular Heart Disease, a prevalence of 7 percent. The prevalence for men is 8.3 percent and for women is 6.1 percent* (ncbi.nlm.nih.gov). We utilize machine learning to predict if an individual would or would not have a heart disease using a few indicators.

I. Problem Statement

Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator includes diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This research makes use of heart disease dataset available on Kaggle. The proposed work predicts the chances of heart disease and classifies patient's risk level by implementing different data mining techniques such as Naive Bayes, XGBoost, Decision Tree, Logistic Regression and Random Forest. Thus, this project presents a comparative study by analysing the performance of different machine learning algorithms. The trial results verify that XGBoost algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

Question: Can we use an ML model to accurately predict cardiovascular disease with a >90% success rate?

II. Data Description

a. Data Summary

Shows the features of the dataset

Feature	Description
Heart Disease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
BMI	Body Mass Index (BMI).
Smoking	Have you smoked at least 100 cigarettes in your entire life?
Alcohol Drinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
Stroke	(Ever told) (You had) a stroke?
Physical Health	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days).
Mental Health	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
Diff Walking	Do you have serious difficulty walking or climbing stairs?
Sex	Are you male or female?
Age Category	Age category.
Race	Imputed race/ethnicity value.
Diabetic	(Ever told) (You had) diabetes?
Physical Activity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
Gen Health	Would you say that in general your health is...
Sleep Time	On average, how many hours of sleep do you get in a 24-hour period?
Asthma	(Ever told) (You had) asthma?
Kidney Disease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
Skin Cancer	(Ever told) (You had) skin cancer?

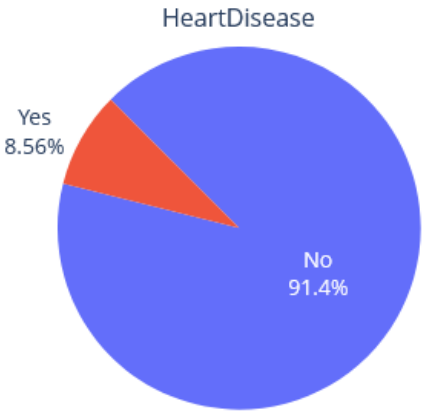
Figure a Table-1

The dataset comes from Kaggle which originally comes from the Centre of Disease Control and is a major part of the Behavioural Risk Factor Surveillance System (BRFSS), which is gathered from annual telephone surveys.

The data that is selected for this project comes from Kaggle, originally sourced from the Centre of Disease Control (CDC) that gathered responses from annual telephone surveys as part of the Behavioral Risk Factor Surveillance System (BRFSS).

- It consists of data for 48 states (excluding Alaska and Hawaii), collected annually from 1982 to 1988
- Data covers 319,795 individuals and has 18 key indicators.
- Dependent variable: HeartDisease – respondents that have ever reported having coronary heart disease or myocardial infarction
- Independent variables (18 total): BMI, Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Sex, AgeCategory, Race, Diabetic, Physical Activity, GenHealth, Sleep Time, Asthma, Kidney Disease, Skin Cancer

b. Pre-processing



We identified that our dataset is imbalanced. The variable "HeartDisease" had only 8.65% 'Yes'. Therefore, we used Python's imblearn to balance our dataset and consider it performing the under-sampling method.

```
from imblearn.under_sampling import NeighbourhoodCleaningRule

ncr = NeighbourhoodCleaningRule(n_neighbors=20, threshold_cleaning=0.5) X_ncr, y_ncr =ncr.fit_resample(X,y)
```

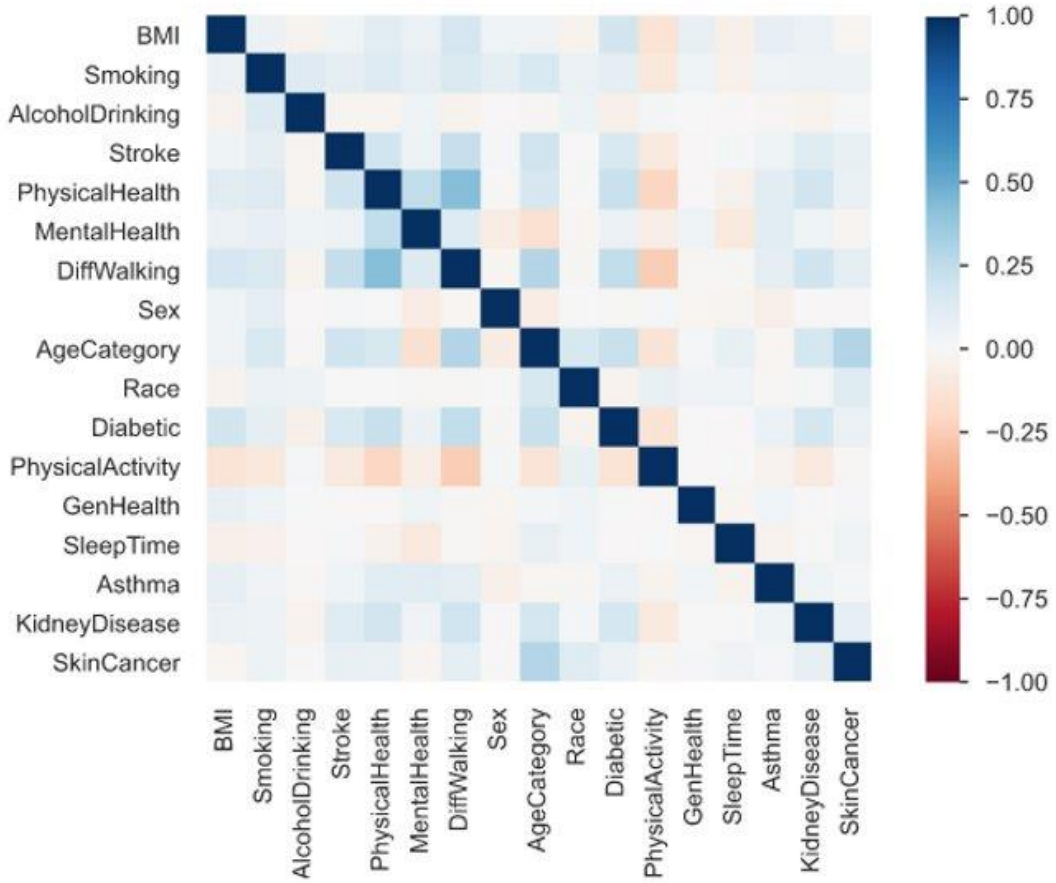
Resampling methods are designed to change the composition of a training dataset for an imbalanced classification task. The Neighborhood Cleaning Rule, or NCR for short, is an under-sampling technique that combines both the Condensed Nearest Neighbor (CNN) Rule to remove redundant examples and the Edited Nearest Neighbors (ENN) Rule to remove noisy or ambiguous examples.

WE converted the variable "Age" which was a categorical variable to a continuous variable. Utilized Label Encoder to encode the labels for preprocessing.

```
encode_AgeCategory = {'55-59':57, '80 or older':80, '65-69':67,
                      '75-79':77,'40-44':42,'70-74':72,'60-64':62,
                      '50-54':52,'45-49':47,'18-24':21,'35-39':37,
                      '30-34':32,'25-29':27}
heart['AgeCategory'] = heart['AgeCategory'].apply(lambda x: encode_AgeCategory[x])
```

We used StandardScaler to normalize and scale our data to run the algorithms efficiently.

c. Variable Summary



We see a positive correlation between Physical Health and Difficulty in Walking, Age and Skin Cancer. We notice a negative correlation between Physical Activity and Difficulty in walking. These variables are not highly correlated, we just noticed a small correlation.

III. Methodology and Implementation

a. Classifiers

The supervised algorithms that we had run for the purpose of this project are as follows:

1. XG Boost
2. Support Vector Machines (SVM) with Grid Search
3. Gaussian Naïve Bayes
4. K-Nearest with Grid Search
5. Logistic Regression

We have also performed Decision Tree, Linear SVM and AdaBoost with Decision Tree, but we are sticking to the above five models for the purpose of this project.

b. Evaluation Metric

We used classification report available in sklearn to summarise and evaluate our model performance. Classification report gives a perspective of your model performance. The 1st row shows the scores for class 0. The column 'support' displays how many objects of class 0 were in the test set. The 2nd row provides info on the model performance for class 1.

This table contains the accuracy as well as the precision, recall and F1 scores for class 0 and class 1 respectively.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1 Score**
XG Boost	0.94	0.95, 0.84	0.97, 0.73	0.96, 0.78
SVM	0.91	0.81	0.64	0.71
Naïve Bayes	0.86	0.93, 0.56	0.90, 0.66	0.92, 0.61
KNN w/ GS	0.92	0.81	0.66	0.73
Logistic Regression	0.92	0.94, 0.81	0.97, 0.66	0.95, 0.73
Decision Tree	0.90	0.93, 0.74	0.96, 0.61	0.94, 0.67
AdaBoost with Decision Tree	0.87	0.89, 0.74	0.98, 0.35	0.93, 0.48

c. Model Summary

The attributes mentioned in Table 1 are provided as input to the different ML algorithms]. The input dataset is split into 80% of the training dataset and the remaining 20% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analysed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further. The different algorithms explored in this paper are as follows:

Model 1 – XGBOOST

XGBoost stands for Extreme Gradient Boosting, and it is an implementation of gradient boosted decision trees designed for speed and performance. In this algorithm, decision trees are created in sequential form. Weights play a key role in XGBoost; weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted incorrectly by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

Both Bagging and Boosting techniques were applied on XGBoost Model – Both gave the same accuracy level of 93.6% and a precision value of 0.95, 0.84, recall score of 0.97, 0.73 and f1 score of 0.96, 0.78 for class 0 and class 1.

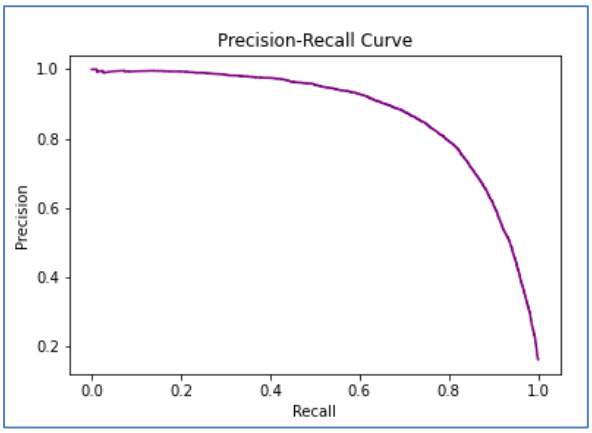


Figure c Precision-Recall for Model 1

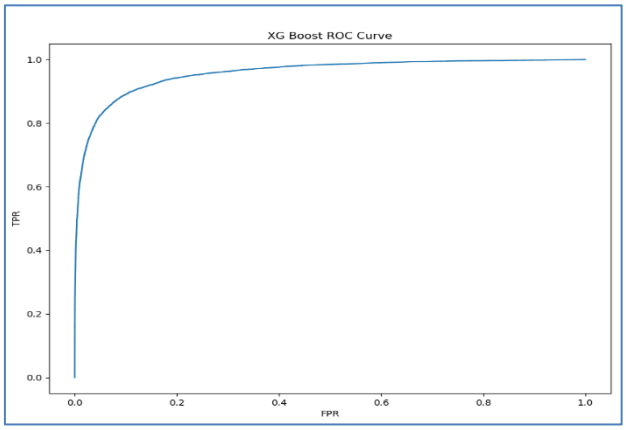


Figure b ROC Curve for Model 1

Model 2 – SVM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. Support Vector Machine is a supervised classification algorithm where we draw a line between two distinct categories to differentiate between them. The goal of SVM is to widen the 'lane' between these categories as much as possible. SVM is also known as the support vector network.
Accuracy attained with SVM - 91.83%, with a precision score of 0.81, 0.64, recall score of 0.64 and f1 score of 0.71.

Model 3- Gaussian Naïve Bayes

Naïve Bayes algorithm is based on the Bayes rule. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B). Accuracy attained was 0.86 and a precision value of 0.93, 0.56, recall score of 0.90, 0.66 and f1 score of 0.92, 0.61 for class 0 and class 1.

Model 4 – KNN with Grid Search

KNN algorithm can be used for both classification and regression problems. The KNN algorithm uses 'feature similarity' to predict the values of any new data points.
Accuracy achieved through KNN – 92.14%, with 0.81 precision, 0.66 recall and 0.73 recall scores.

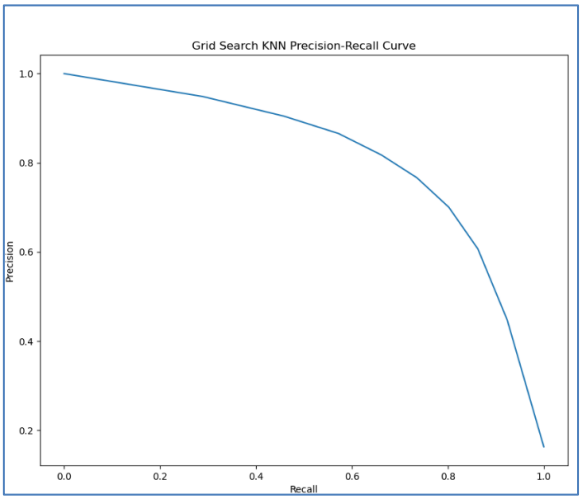


Figure e Precision Recall Curve for Model 4

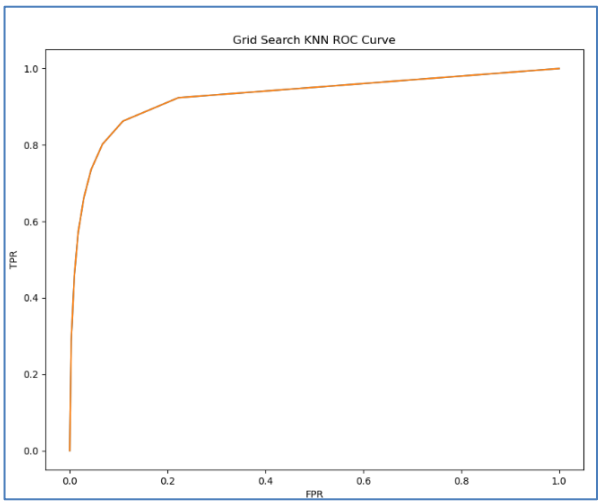


Figure d ROC Curve for Model 4

Model 5 – Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 18 independent variables which makes logistic regression good for classification.

Accuracy achieved using Logistic regression was 91.93%, with a precision score of 0.94, 0.81, recall of 0.97, 0.66 and f1 score of 0.95, 0.73.

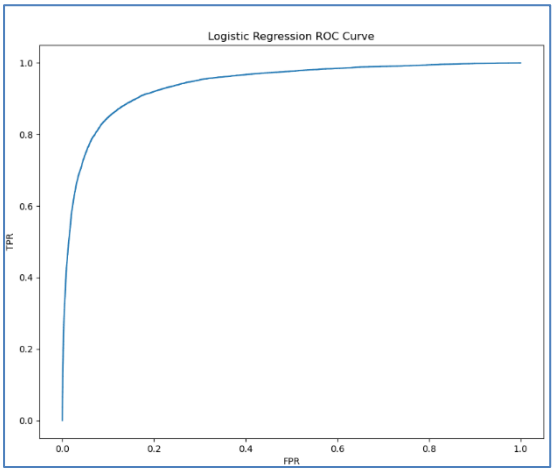


Figure f ROC Curve for Model 5

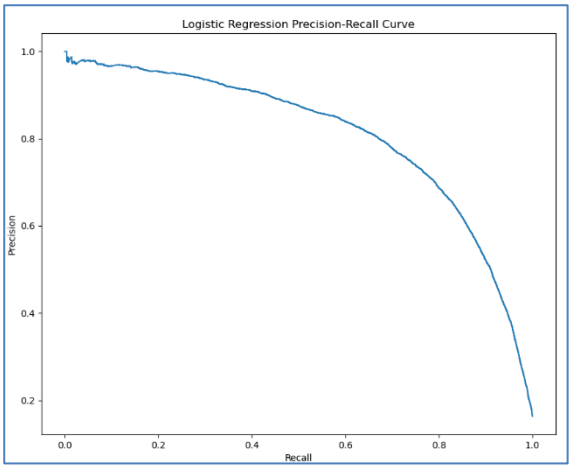


Figure g Precision Recall Curve for Model 5

IV. **Result and Analysis**

The results obtained by applying XGBoost, SVM, KNN with GridSearch and Logistic Regression are shown below. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure.

Precision metric provides the measure of positive analysis that is correct. Recall defines the measure of actual positives that are correct. F-measure tests accuracy.

Precision = (TP) / (TP +FP)
Recall = (TP) / (TP+FN)
F Measure = (2 * Precision * Recall) / (Precision +Recall)
TP True positive: the patient has the disease, and the test is positive.
FP False positive: the patient does not have the disease, but the test is positive.
TN True negative: the patient does not have the disease and the test is negative.
FN False negative: the patient has the disease, but the test is negative.

XG Boost was the best fitting model considering the evaluation metrics out of all the classification models we tested. It has the accuracy of 0.94 with precision value of 0.95-0.84 and recall is 0.97 - 0.73. The F1 score is 0.96 -0.78. This is the model that we would recommend for predicting heart disease.

V. **Conclusion**

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the evaluation score of XGBoost, Logistic Regression, SVM and KNN algorithms for predicting heart disease using Kaggle dataset. The result of this study indicates that the XG Boost was the best fitting model considering the evaluation metrics out of all the classification models we tested. This is the model that we would recommend for predicting heart disease.

VI. **Future scope**

In the future, this work can be enhanced by developing a web application based on the XGBoost algorithm, and the dataset can be expanded to be more feature-inclusive (i.e. Blood pressure, cholesterol, family medical history) as compared to the one used in this analysis. This will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

We also think that classifying the heat disease prediction problem as a multi-class problem depending on the stage of the disease can provide more personalized patient care approach.

Deep learning algorithms should be explored to find if the accuracy, precision, recall and F-measure can be improved further.

References

- Institute of Medicine (US) Committee on a National Surveillance System for Cardiovascular and Select Chronic Diseases. A Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung Diseases. Washington (DC): National Academies Press (US); 2011. 2, Cardiovascular Disease. Available from: https://www.ncbi.nlm.nih.gov/books/NBK83160/
- ncbi.nlm.nih.gov
- https://www.ijert.org/heart-disease-prediction-using-machine-learning