Atherosclerosis risk estimator app

BY GIGACHADS

# Overview

## Project Background and Description

|  |  |
| --- | --- |
|  | *Atherosclerosis is a disease in which plaque builds up in the arteries, making it difficult for blood to flow through. This can lead to serious problems like heart attack and stroke.*  *It is a leading cause of death and disability worldwide.*  *Early detection and management of atherosclerosis can greatly improve outcomes for patients.*  *Our bioinformatics app aims to make this process more efficient and accessible by using machine learning to analyze a person's DNA and predict their risk of developing the disease.* |

## Project Scope

|  |  |
| --- | --- |
|  | In this project, first we gathered the data and find out which genes can be related with atherosclerosis and then build a Machine learning model that checks the accuracy of the data. After that we developed an App, and this app can help individuals and healthcare providers take proactive measures to reduce the likelihood of serious complications from atherosclerosis. |

## High-Level Requirements

|  |  |
| --- | --- |
|  | Our App uses state-of-the-art machine learning techniques to analyze large amounts of genomic data. |

The system include the following,

* The app is designed to be user-friendly and accessible
* The app also provides personalized recommendations for lifestyle changes and medical interventions that can help reduce the risk of developing the disease.

## Deliverables

|  |  |
| --- | --- |
|  | This complete project is delivered by Team Lead: Ilya & Team Members: Daria, Asif, Tayyab, Hamna |

## Affected Parties

|  |  |
| --- | --- |
|  | If we build an app B2B we may face more competition from established companies in the healthcare industry |

Machine LEARNING IN GENOME DATA

**Abstract:**

The identification of patterns has a significant impact on the accuracy of disease prediction during the disease diagnosis process. We require an advanced illness prediction system that can accurately diagnose the condition by evaluating a variety of characteristics and symptoms. This system must be able to learn from previously diagnosed diseases and be flexible enough to adapt to new techniques. To provide the medical association for the analysis of syndrome among patients, a graphical user alliance will be improved. In this paper, various techniques available for disease prediction such as SVM (Support Vector Machine), Gradient Boosting Model, logistic regression. An overview is done on the prevailing and anticipated models for human disease prediction. The Machine learning model that was utilized incorporate SVM, Gradient Boosting model. The framework was executed utilizing every model and their exhibition was assessed. Finally a model has been developed in which the person can insert his/her symptoms and the disease identification is done that can give the result based upon the inputs.

# Introduction:

Machine learning is a subfield of artificial intelligence. The main aim of machine learning is to comprehend the structure of data and fit it into models that people can comprehend and use. Instead, machine learning techniques enable computers to train on data inputs and make use of statistical analysis to produce values that fall inside a given range. Machine learning enables the systems to learn without being adjusted remotely. Algorithms for classification are used to select from a set of options that best matches a variety of perceptions. Supervised learning is one of the most common and important paradigms in pattern recognition, with pattern classification being one of the most important tasks. Pattern classification has become very important for decision making in many areas of human activity and the medical area is no exception.

The primary focus of this work is on classification algorithms that are beneficial for speedy disease diagnosis in the medical field. This topic is crucial since a correct diagnosis will greatly prolong the patient's life. In this work, dataset have been chosen from the (Kaggle: cardio disease dataset). In these repositories there are datasets of patterns of medical diseases, and they offer balanced and unbalanced data.

The classification used in the present work are: SVM (Support Vector Machine, Gradient boosting algorithm. In addition, we tested several measure of data complexity. The different machine learning models include:

* **Support Vector Machine (SVM):**

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problem. An SVM model can classify new text after being given sets of labelled training data for each category. The scikit-learn python bundle is used to implement SVM. SVM requires the data points and generate the hyperlane (which in two dimensions is simply a line) that separates the lables at best.

* **Gradient Boosting Model:**

Predictive analytics play an important role in clinical research. If a clinical condition or outcome can be accurately predicted, interventions can be delivered to the patient population who will benefit the most. Thus, numerous data mining algorithms have been developed for clinical prediction in nearly all subspecialties. Machine learning methods can be used in this setting as we can rely on the predictive algorithm to discover the nonlinearity and interaction structure in the data instead of hoping that the investigators will be wise enough to include such structure explicitly in their models.

Among the many machine learning techniques, gradient boosting is a particularly attractive approach. In the clinical literature, gradient boosting has been successfully used to predict, among other things, cardiovascular events (2), development of sepsis (3), delirium (4) and hospital readmissions etc.

To diagnose any disease identification is a vital task. Sometimes, various signs and symptoms can not be identified, therefore diagnose is the difficult task. In our dataset, there are 3 types of input features:

* Objective: factual information;
* Examination: results of medical examination;
* Subjective: information given by the patient.

Features in our dataset includes:

1. Age (in days) 4. Gender (categorical)
2. Height (cm) 5. Smoking (binary)
3. Weight (kg) 6.Cholestrol(0:normal, 1:above normal, 2:well)

**Simulation of an Existing model:**

The SVM(Support Vector Machine) and Gradient Boosting model over the cardio disease have been simulated and calculated the Recall, Accuracy, F1\_measure, Precision and ROC\_Auc score. These quantitative measurements are used to evaluate and compare the efficiency of models.

Simulation Parameters:

A True Positive (TP): the symptoms that cause disease is treated as the symptoms

that play a role in disease by the system.

True Negative (TN): the symptoms which do not play a role to cause disease is

treated as same.

False Positive (FP): the symptoms which not play a role in disease is treated as

the symptoms which act in disease prediction.

False Negative (FN): the symptoms which cause disease but were wrongly

classiﬁed as symptoms which do not cause disease by the system

A True Positive (TP): the symptoms that cause disease is treated as the symptoms

that play a role in disease by the system.

True Negative (TN): the symptoms which do not play a role to cause disease is

treated as same.

False Positive (FP): the symptoms which not play a role in disease is treated as

the symptoms which act in disease prediction.

False Negative (FN): the symptoms which cause disease but were wrongly

classiﬁed as symptoms which do not cause disease by the system

**True Positive (TP):** the symptoms that cause disease is treated as the symptoms that play a role in disease by the system.

**True Negative (TN):** the symptoms which do not play a role to cause disease is treated as the same.

**False Positive (FP):** the symptoms which not play a role in disease is treated as the symptoms which act in disease predictions.

**False Negative (FN):** the symptoms which cause disease but were wrongly classified as symptoms which do not cause disease by the system.

1. **Accuracy:** It is the fraction of symptoms that had been efficiently labeled. It is one of the maximum effective and typically used assessment notations.

**Accuracy = (TP + TN ) / (TN + FN + FP + TP)**

1. **Recall:** It is the fraction of symptoms (the symptoms that have maximum chances of being wrong) correctly classified by the system.

**Recall = TP / (TP + FN)**

1. **Precision:** It is the number of symptoms either cause disease or not that was correctly classified. It is also known as detection rate.

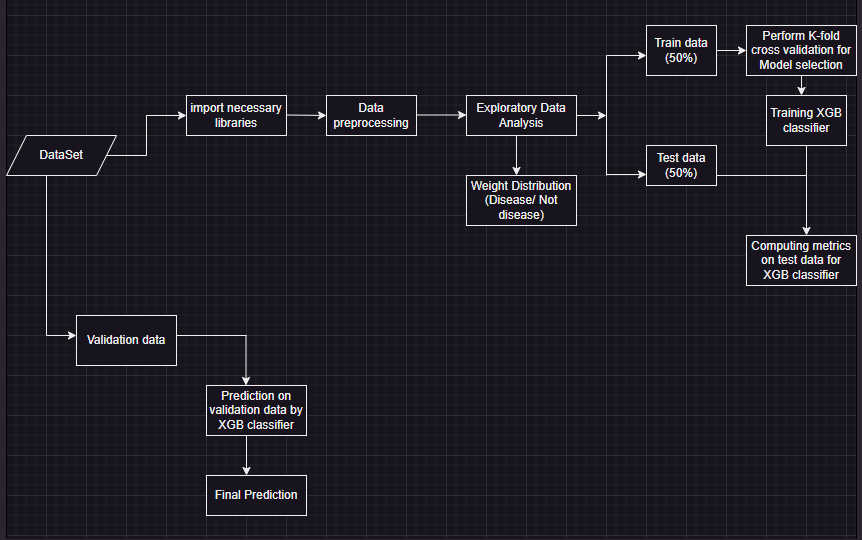
**Precision = TP / (TP + FP)**

1. **F1\_measure:** It is the harmonic mean among precision and take into account. High precision but decrease take into account, gives a really accurate; however, it then misses a big variety of times which can be difficult to categorize.

**F1\_Score = (2 \* precision \* recall) / (precision + recall)**

1. **ROC\_Score:** ROC score is the receiver operating characteristic and the graph is plotted against True Positive Rate and False Positive Rate for different threshold values. ROC\_AUC is area under the curve.

**Workflow for the Implementation:**

****

This workflow diagram is providing the information of our Machine learning coding implementation. We can easily understand the concept of our model and its performance.

**Prototype:**

A model for diagnosing human disease to aid in the identification of common disorders, medical expert systems have been developed. This will aid the doctor in making a few diagnoses. The framework undergo the following steps:

**Step 1:** Adding the required libraries (Pandas, Numpy, KNeigh-borsClassifier, SVM, make pipeline, XGB Classifier) is the first step.

**Step 2.** Retrieve the dataset

**Step 3:** Pre-process the dataset using the given procedures.

**Step 4.** Divide the dataset into training and testing categories.

**Step 5:** Create the model by utilizing several methods.

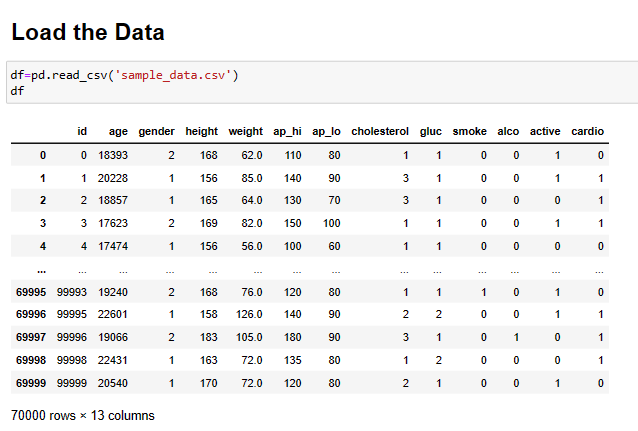
**Step 6**. Examine the model.

**Step 7.** If the outcome (score) indicates that the model is effective, continue to step 8; if not, go to step 3 and attempt a different model.

**Step 8:** Predict the outcome using the user's input.

1. **Reading the dataset:**

First, we will use the pandas library to load the dataset from the folders. We shall remove the null column when reading the dataset. This dataset is free of null values, and each feature only contains 0s and 1s. Every time we complete a classification operation, we must determine whether or not the target column is balanced.



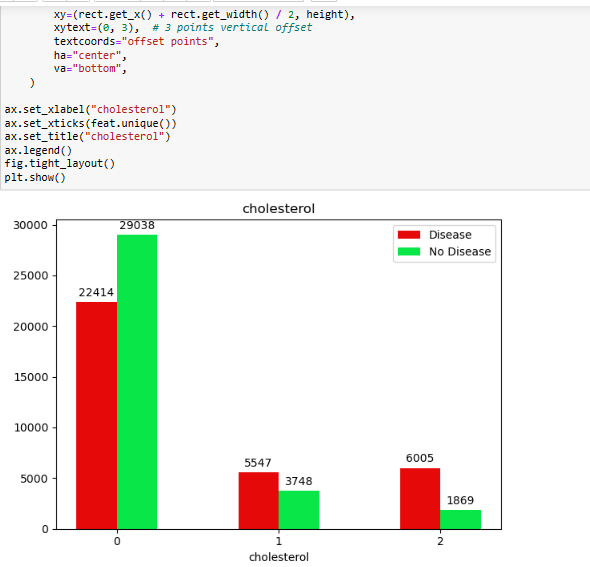
1. **Processing:**

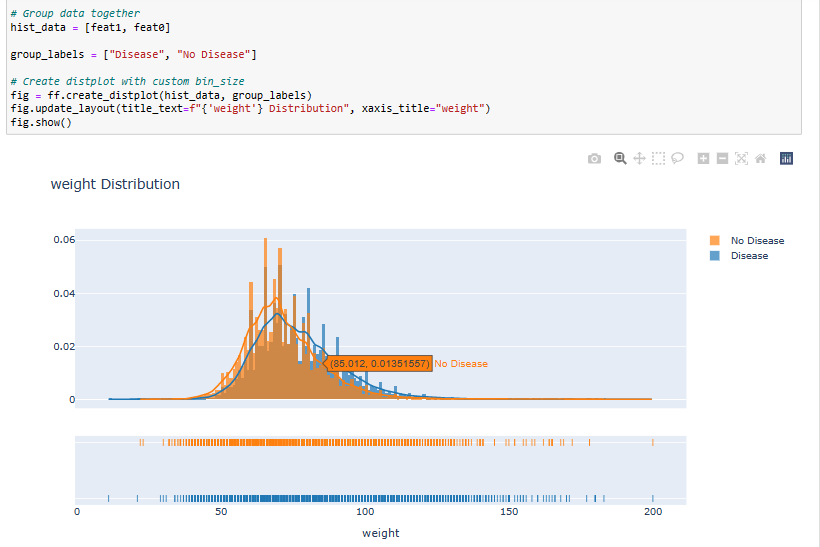
The first step in solving any machine learning challenge is to preprocess the data. Before doing any machine learning calculations on the data, a significant percentage of it needs to be cleaned, or preprocessed. The segments' mean estimates are used to replace the portions that have invalid attributes. Our model is trained on the SVM, Gradient boosting algorithm. These models have been chosen as they are the best among all, especially for medical field. Based on the trained model, our system is tested by taking the input and the model predict the result according to the inputted value whether the person is suffering from the disease or not and gives the accuracy ratio that how much sure the person is suffering from particular disease.

1. **Exploratory Data Analysis:**

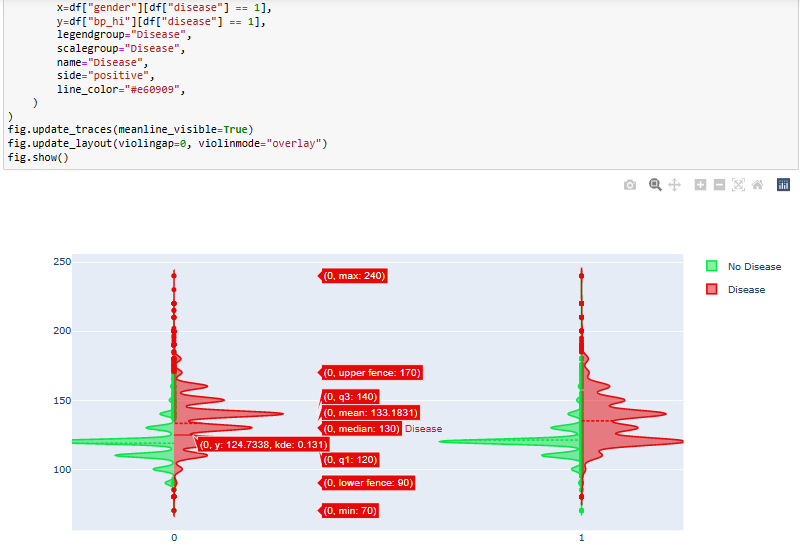
After processing the data we display data in different graphs for better understanding. In cholesterol column we represent the numerical value (0: normal, 1: above normal, 2: well above normal).

* Bar chart shows the distribution of the patients that have normal cholesterol level, above normal or well above normal with disease or no disease rate with color representing green and red. And then distribute the weight that is showing the person have a disease or not.



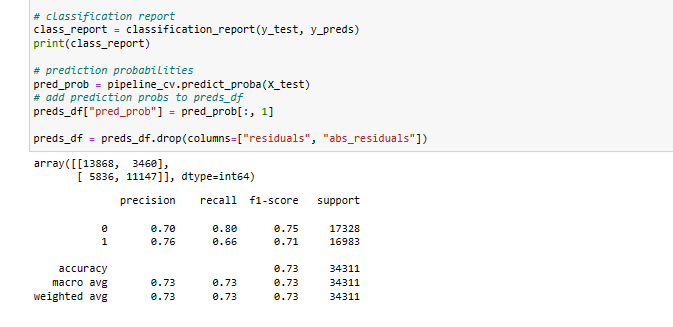


* And Violin plot summarizes a data set using 6 measures: a [boxplot](https://www.labxchange.org/library/pathway/lx-pathway:3eec864d-746b-41dc-b3e4-8bb6a03949fe/items/lx-pb:3eec864d-746b-41dc-b3e4-8bb6a03949fe:html:cf9951aa) is used to show the minimum, first quartile, median, third quartile, and maximum. There is also a probability density function (PDF), which is essentially a rotated, smoothed histogram.



1. **Splitting the data for training and testing model:**

It's time to divide the data into training and testing groups after cleaning the data by removing null values. We split the data into 50:50 format i.e. 50% of the dataset will be used for training the model and 50% of the data will be used to evaluate the performance of the models. When we predict the test data then the accuracy, recall, precision and f score of the model is shown below that display that the recall score is high as compared to other estimators.



**Predict with ROC:**

