Supervised ML (classification)

Cardiovascular Risk Prediction



**Technical documentation**

**\_\_**

AKSHAY FASALE

# INTRODUCTION

Cardiovascular diseases, also called CVDs, are the leading cause of death globally, causing an estimated 17.9 million deaths each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age. The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity.

# Problem statement

The project goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

# Overview of data

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

This dataset has the following 17 variables as explanatory variables:

### **Demographic:**

• Sex: male or female("M" or "F")

• Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

### **Behavioral:**

• is\_smoking: whether or not the patient is a current smoker ("YES" or "NO")

• Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

### **Medical(history):**

• BP Meds: whether or not the patient was on blood pressure medication (Nominal)

• Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)

• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)

• Diabetes: whether or not the patient had diabetes (Nominal)

### **Medical(current):**

• Tot Chol: total cholesterol level (Continuous)

• Sys BP: systolic blood pressure (Continuous)

• Dia BP: diastolic blood pressure (Continuous)

• BMI: Body Mass Index (Continuous)

• Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

• Glucose: glucose level (Continuous)

### **Predict variable (desired target):**

• 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

**Steps involved**

1. **Data Cleaning**

Data cleaning is the process of removing incorrect, duplicate, or otherwise erroneous data from a dataset. These errors can include incorrectly formatted data, redundant entries, mislabeled data, and other issues; they often arise when two or more datasets are combined. Data cleaning improves the quality of your data as well as any business decisions that you draw based on the data.

1. Determining the critical data values you need for your analysis.
2. Collect the data you need, then sort and organize it.
3. Identify duplicate or irrelevant values and remove them.
4. Search for missing values and fill them in, so you have a complete dataset.
5. Fix any remaining structural or repetitive errors in the dataset.
6. Identify outliers and remove them, so they will not interfere with your analysis.
7. Validate your dataset to ensure it is ready for data transformation and analysis.
8. Once the set has been validated, perform your transformation and analysis.
9. **Exploratory Data Analysis (EDA)**

**Exploratory Data Analysis (EDA)**is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

1. Exploring head and tail of the data to get insights on the given data.
2. Converting the data into appropriate data types to create a model
3. Creating data frames which help in drawing insights from the dataset.
4. Creating more columns in our dataset which would be helpful for creating model.
5. Creating visualization and analysis based on univariate, bivariate and multivariate.
6. **Data Transformation**

Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form.

Data transformation is an essential data pre-processing technique that must be performed on the data before data mining to provide patterns that are easier to understand.

Data transformation changes the format, structure, or values of the data and converts them into [clean, usable data](https://www.zuar.com/blog/data-cleaning-the-benefits-and-steps-to-creating-and-using-clean-data/). Data may be transformed at two stages of the data pipeline for data analytics projects. Organizations that use on-premises data warehouses generally use an ETL (extract, transform, and load) process, in which data transformation is the middle step. Today, most organizations use cloud-based data warehouses to scale compute and storage resources with latency measured in seconds or minutes. The scalability of the cloud platform lets organizations skip preload transformations and load raw data into the data warehouse, then transform it at query time.

**Data integration, migration, data warehousing, data wrangling** may all involve data transformation.

1. **Model Building and Evaluation**

Now that the Dataset is cleaned, and we have added all the necessary features along with some conversions of categorical features. Its time to split the data into training and testing sets and observe how the models are performing.

Note: - These training and testing data are going to be same for all the model we'll build such that all of the models are evaluated on a same set of parameters.

## **Models we will try to build:**

* **Logistic Regression Classifier**
* Logical regression analyses the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modelling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.
* For example, 0 – represents a negative class; 1 – represents a positive class. Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).
* **Decision Tree Classifier**
* Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.
* Decision trees are upside down which means the root is at the top and then this root is split into various several nodes. Decision trees are nothing but a bunch of if-else statements in layman terms. It checks if the condition is true and if it is then it goes to the next node attached to that decision.
* **Random Forest Classifier**
* Random Forest is a technique that uses ensemble learning, that combines many weak classifiers to provide solutions to complex problems.
* As the name suggests random forest consists of many decision trees. Rather than depending on one tree it takes the prediction from each tree and based on the majority votes of predictions, predicts the final output.
* Random forests use the bagging method. It creates a subset of the original dataset, and the final output is based on majority ranking and hence the problem of overfitting is taken care of
* **XGB Classifier**
* The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage.
* XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models.
* The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms.
* **Bagging** and **Boosting** are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees.
* **K-Nearest Neighbours(KNN Classifier)**
* KNN which stands for K-Nearest Neighbours is a simple algorithm that is used for classification and regression problems in Machine Learning. KNN is also non-parametric which means the algorithm does not rely on strong assumptions instead tries to learn any functional form from the training data.
* Unlike most of the algorithms with complex names, which are often confusing as to what they really mean, KNN is pretty straight forward. The algorithm considers the k nearest neighbours to predict the class or value of a data point.
* The K Nearest Neighbour Algorithm can be performed in 4 simple steps.
* Step 1: Identify the problem as either falling to classification or regression.
* Step 2: Fix a value for k which can be any number greater than zero.
* Step 3: Now find k data points that are closest to the unknown/uncategorized datapoint based on distance(Euclidean Distance, Manhattan Distance etc.)
* Step 4: Find the solution in either of the following steps:
* In case of classification, we assign the uncategorized datapoint to the class where the maximum number of neighbours belonged to.
* Support Vector Machine (SVM Classifier)
* SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems.
* Support Vectors: These are the points that are closest to the hyperplane. A separating line will be defined with the help of these data points.
* Margin: it is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM large margin is considered a good margin. There are two types of margins hard margin and soft margin

**Challenges involved**

* Pre-processing the data was one of the challenges we faced.
* One of the challenges which we faced was that we can’t skip the outliers as the data which we have is of patient as we can’t skip anything.
* Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

**Conclusion**

We are finally at the conclusion of our project!

We've noticed that **XBG Classifier** is the standout performer among all models with an f1-score of **0.8397**. it's safe to say that XGB Classifier provides an optimal solution to our problem.

In case of **Logistic regression**, we were able to see the maximum f1-score of **0.658**.

Out of the tree-based algorithms, the **Random Forest Classifier** was providing an optimal solution towards achieving our Objective. We were able to achieve an f1-score of **0.7703** We also noticed that in the case of **Decision-tree Classifier**, we were able to achieve an f1-score of **0.7034** for the test split.

For **SVM (Support Vector Machines)** Classifier, the f1-score lies around **0.7417**.

Finally, As in the medical domain (False negative values have importance we don’t want to mis predict a person safe when he has the risk) recall ha the most importance. KNN, XGB, Random Forest gave the best recall 0.86 ,0.80 ,0.81.