# Capstone  Project-3

# Supervised Learning [Classification]

# Cardiovascular Risk Prediction

## INTRODUCTION TO SUPERVISED MACHINE LEARNING

Supervised learning is the types of machine learning in which machines are trained using well "labeled" training data, and on basis of that data, machines predict the output. The labeled data means some input data is already tagged with the correct output. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

## Insights of the project

We are provided with a data set containing relevant data required to predict the demand for bike. Utilising the concepts of EDA(Exploratory Data Analysis),supervised learning algorithms ,Data Visualisation we will be using the data available to analyze the factors affecting the demand . The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

## Algorithm followed:

1.Importing the necessary packages and libraries.

2.Mounting the drive for importing the data.

3.Checking for missing, Nan values, Null values.

4.Observing the datatypes and features.

5.Observing the correlation among independent variables.

6. Exploring the categorical values, numerical features from data set.

7.Exploring the target variable.

8.Handling imbalanced dataset with SMOTE.

9.Splitting the data and training the data.

10.Observing the results.

he results.

## ***1.Exploratory Data Analysis***

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

.

## ***2.Feature Engineering***

Feature engineering is a machine learning technique that leverages data to create new variables that aren’t in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy**.** Feature engineering is required when working with machine learning models. Regardless of the data or architecture, a terrible feature will have a direct impact on our mode.

## ***3.Variance Inflation Factor***

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

## ***4.Multicollinearity***

Multicollinearity, or collinearity, is the existence of near-linear relationships among the independent variables. During regression calculations, this relationship causes a division by zero which in turn causes the calculations to be aborted. When the relationship is not exact, the division by zero does not occur and the calculations are not aborted. However, the division by a very small quantity still distorts the results. Hence, one of the first steps in a regression analysis is to determine if multicollinearity is a problem.

### ***Effects of Multicollinearity***

Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false, nonsignificant, pvalues, and degrade the predictability of the model (and that's just for starters).

Now we will switch to building different types of models. Firstly we will train it by feeding and then will find the accuracy.Supervised machine learning Classification models are:-

## ***Following models have been used for predictions:-***

- Logistic Regression Classifier
- K-Nearest Neighbors(KNN Classifier)
- Support Vector Machine(SVM Classifier)
- XGB Classifier
- Decision Tree Classifier
- Random Forest Classifier

## ***Model Development***

Now its time to implement the Machine Learning models and check the accuracy of each model to point out the best one out of all. In this project we are implementing 8 machine learning algorithms to predict the target variable and also we'll apply optimization techniques to get the best resulting accuracy.

Now that the Dataset is cleaned and we have added all the neccessary features along with some conversions of categorical features. Its time to split the data into training and testing sets. 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. We can clearly see, the class are imbalanced and it'd result the model to be more biased towards '0' class (people with no Risk of CHD). We need to find a way to train a model in such a way that it can take some risks and give more of '1' class results. The reason behind that is we have a make a model that can predict a risk of CHD. If, based on patient's health stats, there's even small chance that a person could have a risk of heart disease, the model should be able to predict the risk. Surely this would lead our model to attain a lower accuracy value since its taking more risks for 'at risk' patients, but in a real world scenario this model is more useful because its highly likely for such model to be able to predict if a patient is at risk. If we would have used a normal approach here, it might give us a better accuracy thanks to the class imbalancy, but in real world such model isn't a ideal solution, its expected that it'd mostly fail to predict if the patient is at risk, which takes away the sole purpose of the project.

## ***Conclusion***

We've noticed that XBG Classifier is the stand out 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.KNN gave us recall of 86 % and with gridsearch CV it gave a recall of 93%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 dont want to mispredict 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.

## ***References:-***

- https://pandas.pydata.org/pandasdocs/stable  Python MatPlotLib
- Documentation https://matplotlib.org/stable/index. html  Python Seaborn Documentation
- https://seaborn.pydata.org  Python SkLearn Documentation
- https://scikit-learn.org/stable