A Project Report

On

"Predict diseases (heart attacks) with AI"

by

Bhushan B. Yadav

1. Introduction:

An estimated 17 million people die of CVDs (Cardiovascular disease), particularly heart attacks and strokes, in the world every year. Cardiac ailments killed more Indians in 2016 (28%) than any other non-communicable disease, said a new study published in the September 2018 issue of health journal, The Lancet. These are double the numbers reported in 1990 when heart disease caused 15% of deaths in India. Today we will try to build a heart attack predictor. Based on some diagnostically measured parameters we will predict who among the subjects under consideration, are on high risk of heart attack. This can revolutionize the healthcare system and help save many many lives.

(taking help from others report)

2. Objective:

To build a heart attack predictor based on some diagnostically measured parameters.

3. Problem Definition:

Predicting heart attacks using predictive analysis models of machine Learning like decision tree, logistic regression, random Forest, KNN

4.Related Theory:

4.1.Decision Tree:

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.
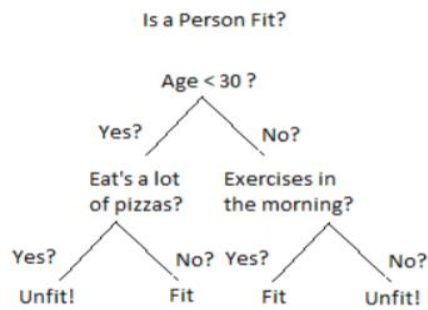
An example of a decision tree can be explained using above binary tree. Let's say you want to predict whether a person is fit given their information like age, eating habit, and physical activity, etc. The decision nodes here are questions like 'What's the age?', 'Does he exercise?', 'Does he eat a lot of pizzas'? And the leaves, which are outcomes like either 'fit', or 'unfit'. In this case this was a binary classification problem (a yes no type problem).

4.2. Logistic Regression:

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.
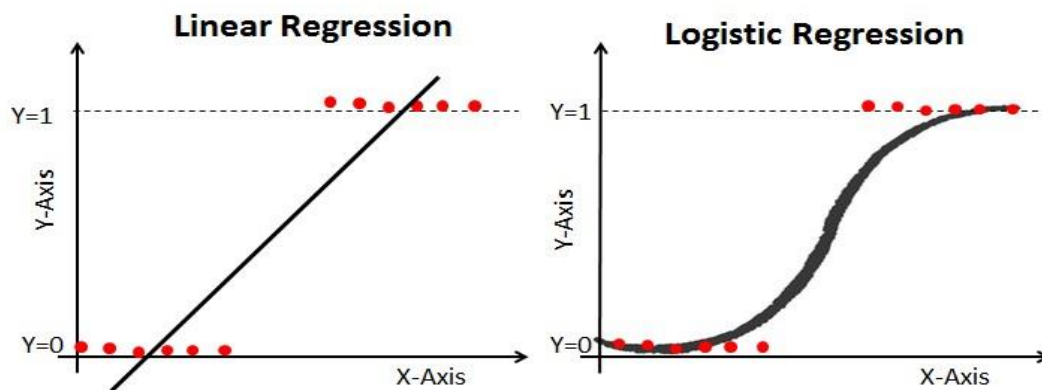

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

4.3 KNN:

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.
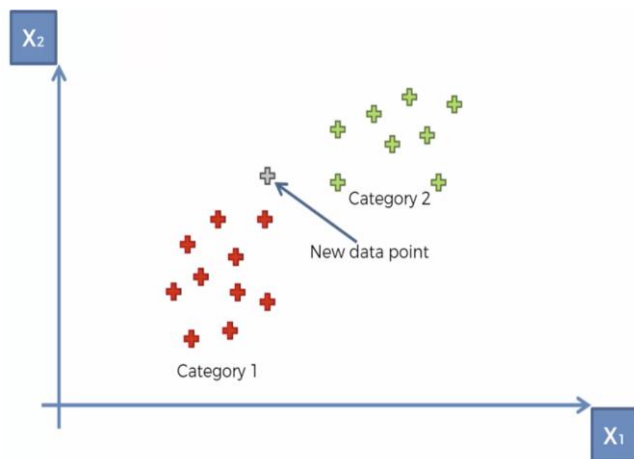
Notice in the image above that most of the time, similar data points are close to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity with some mathematics, calculating the distance between points on a graph.
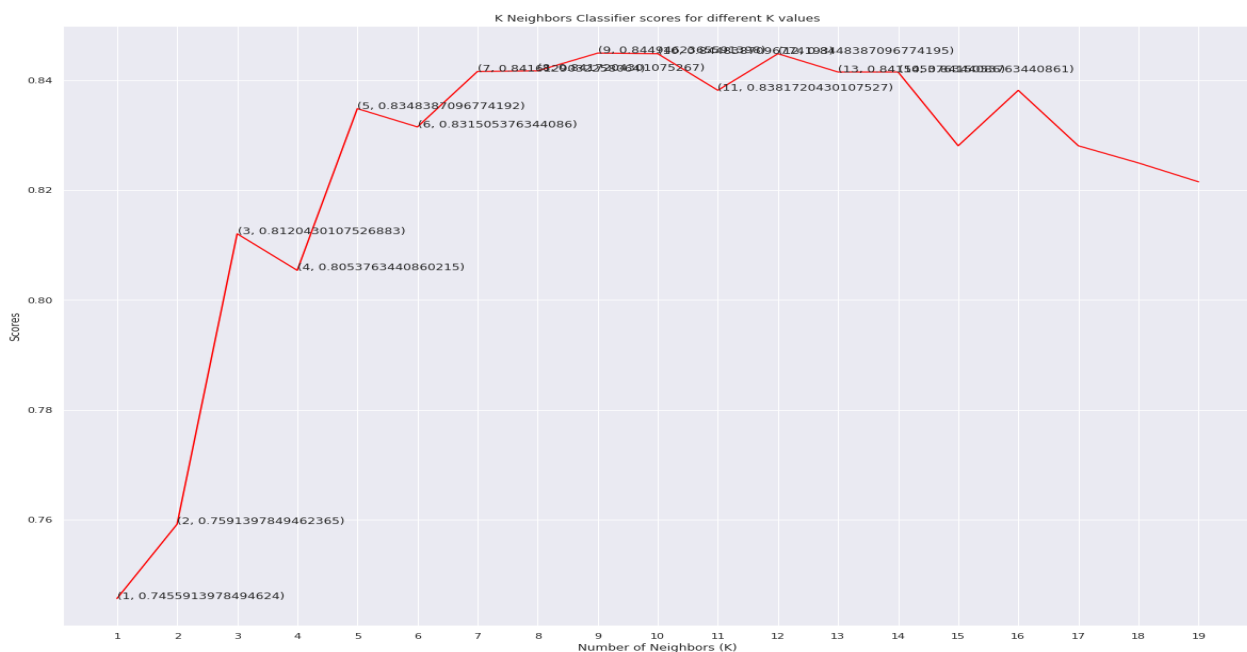
Observation : we take 1 to 30 iterations and apply KNN as we can see for the first KNN accuracy which is 1.07455 and likewise for 2KNN acc - 2.0755, Lastly for 12th iteration the KNN accuracy is about 85% which very good for K.

4.4 Random Forest :

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

In the above Image simple representation of random forest with neat code and I would the accuracy – 0.8215 which is not bad accuracy.

5.Dataset -

| 4 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |