**Prediction of heart diseases using classification**

A Synopsis Submitted

in Partial Fulfillment of the Requirements

for the Course of

# Minor Project – I

In

Third year – Fifth Semester of

**Bachelor of Technology**

specialization

In

# Business Analytics and Optimization

Under

**Dr. Tanupriya Choudhury**

By

**500063452                  R103217106               Shweta Rawat**

**500062969                  R103217082               Rahul Jain**

**500062351                  R103217004               Adarsh Mundra**



DEPARTMENT OF INFORMATICS

SCHOOL OF COMPUTER SCIENCE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES, BIDHOLI, DEHRADUN, UTTRAKHAND, INDIA

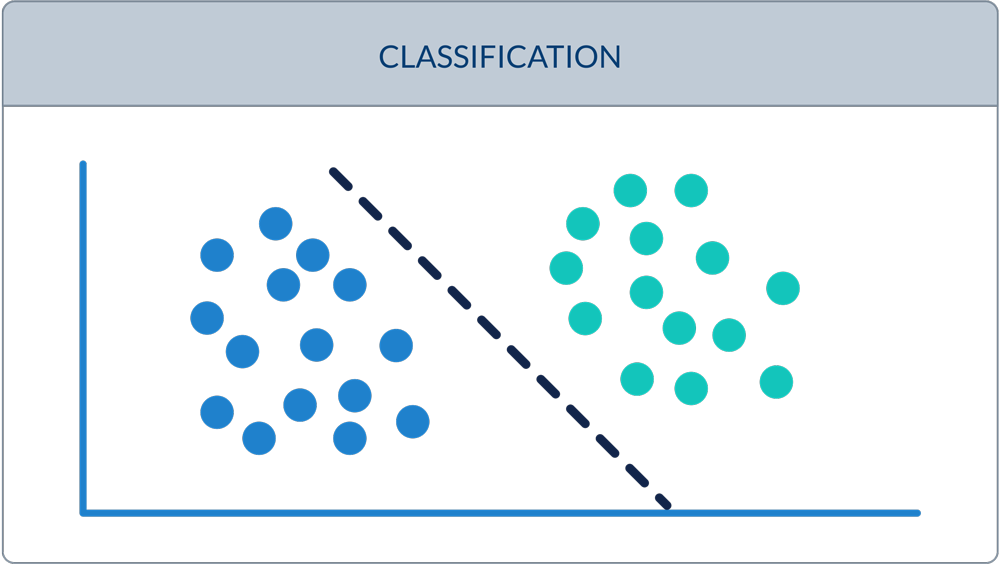
# Aug, 2019

**Synopsis**

**Introduction**

The heart disease accounts to be the leading cause of death worldwide. It is difficult for medical practitioners to predict the heart attack as it is a complex task that requires experience and knowledge. The health sector today contains hidden information that can be important in making decisions. Data mining here comes into picture. It is the process of discovering hidden patterns in large data sets according to different perspectives for categorization into useful information. Data mining enable the health sector to predict patterns in the dataset. There are a lot of data mining algorithms for classification such as Naive Bayes, CART , KNN etc .

In this project we will be applying one of the supervised technique that is classification on our dataset to find relationships between the available features and will classify the labels accordingly. Talking about classification when the true goal of our analysis is to predict to which class or group an observation belongs, the techniques we use are termed as classification techniques. There are two methods of classification : i) classification according to attributes and ii) classification according to variables. An attribute is a qualitative characteristic which cannot be expressed numerically for example gender, country etc. we can’t quantify these characteristics. The qualities of an attribute can easily be differentiated by means of some natural line of demarcation. For example if we take hair color as the basis of classification, there will be two groups one of brown haired people and other with black haired people. Next we have classification according to variables, variable are quantifiable characteristics and can be expressed numerically for example age, weight, salary etc. The data is generally shown in the form of a frequency distribution(a tabular presentation that generally organizes data into classes, and shows the no. of observations falling into each of these classes ).



**Fig: Classifying blue and green classes with classification**

**Motivation**

We are in the era of “Information age” where terabytes of data are produced and stored everyday because of fast growth in “Information Technology”. Huge amount of data is being generated by hospitals such as heart paining results, x-ray, chest pain results, blood pressure etc. There is no effective use of the important data that is generated from the hospitals. Heart disease has the largest proportions of deaths in the world and a lot of research is being conducted in this field.

The motivation of this project is to predict heart diseases using data mining techniques like KNN, Naive Bayes etc. and visualizing the results using one of the famous analysis tool ‘MS Excel’.

Today machine learning has no doubt occupied almost every field and is performing really well but implementing and creating models in C language proposes some challenges related to loading and preprocessing the data along with feature extraction and classification.

**Related work**

Classification techniques has been really useful in finding out the class or group to which the new observation belongs. There is a lot of data available for researchers and prediction of heart diseases has been a field of interest for many researchers and a lot of research has been done in this field . Mosima Masethe mentioned in her work

‘Prediction\_of\_Heart\_Disease\_using\_Classification\_Algorithms some classification algorithms like Naive Bayes, J48, REPTREE, CART, and Bayes Net https://www.researchgate.net/publication/289760362\_Prediction\_of\_Heart\_Disease\_using\_Classification\_Algorithms

Simge Ekiz and Pakize Erdogmus in their paper ‘ Comparative study of heart disease classification’ compared two important machine learning platform results for the same dataset and conducted the experiment to classify heart disease both in Matlab and WEKA environment using six different algorithms including Linear SVM, Quadratic SVM, Cubic SVM, Medium Gaussian SVM, Decision Tree and Ensemble Subspace Discriminant machine learning approaches.

J. Hu, G. Chen, H. Peng, J. Wang, X. Huang and X. Luo, "A kNN classifier optimized by P systems," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, 2017, pp. 432-437. In this paper they proposed the k-nearest neighbors (kNN) classification algorithm optimized by P systems, called kNN-P, which can improve the performance of the original kNN classifier. The proposed kNN-P is evaluated on 18 benchmark datasets and compared with classical kNN algorithm and 8 recently developed improved algorithms. Comparison results demonstrate the availability and effectiveness of the proposed algorithm.

Ying Li and Bo Cheng, "An improved k-nearest neighbor algorithm and its application to high resolution remote sensing image classification," 2009 17th International Conference on Geoinformatics, Fairfax, VA, 2009, pp. 1-4. In this paper Li and Bo adopted an improved KNN classification algorithm and applied it to object-oriented classification of high resolution remote sensing image . They believed KNN will increase classification error rate when training samples distribute unevenly or sample number of each class is very different. An improved KNN algorithm was introduced and used to classify the sample points respectively. Finally, classification results are compared. the improved KNN algorithm achieved higher accuracy in the classification of high resolution remote sensing image.

A. P. Pawlovsky and M. Nagahashi, "A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis," IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, 2014, pp. 189-192. This work shows the application of the k-Nearest Neighbors method to prognosis breast cancer and also proposed a method to select a good setting with the parameters that can be changed when using the classification method. They got an average accuracy of 76%, a small standard deviation, and a small difference between its maximum and minimum values.

we also did analysis on some datasets applying some classification techniques like decision trees and KNN along with logistic regression for non- linear data in some well-known data sets like farmingham data set, breast cancer data set etc.

**Proposed Method**

When performing any kind of analysis whether it is classification or regression, 70% of the job is done when you get a good dataset. The work starts from gathering a good dataset and loading it in our work space followed by data preprocessing .Performing data preprocessing in C is very challenging in itself. In the context of the language, as we all know C is kind of restricted language with no databases connectivity and networking support. Unavailability of databases in C creates many problems, because for small datasets we can use arrays but arrays are not that sufficient when it comes to analysis with large datasets. Below are the steps or modules to be followed in the work.

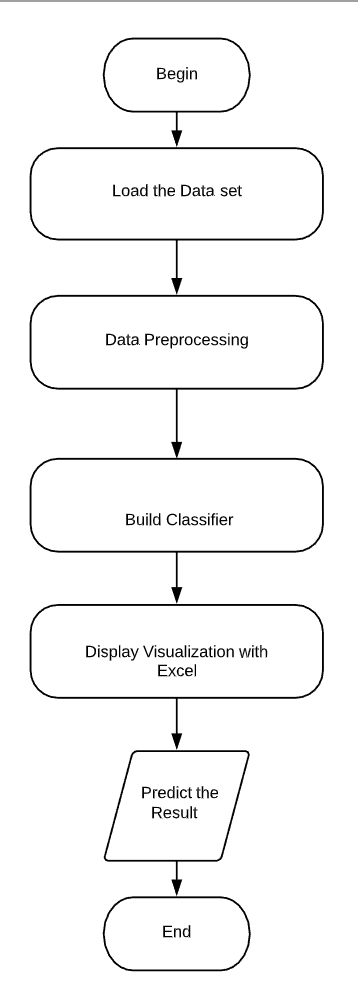
**1. Load the dataset :**  Before any calculations it is necessary to load the dataset in our workspace. We used arrays for this purpose as these provide a good way for doing necessary calculations in the data.

**2. Data Preprocessing :**  Data preprocessing or data cleaning steps involves handling the along with removal of unnecessary features . Null values are replaced by appropriate mean, median and mode values.Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. This is done because there are a large number of variables that require a lot of computing resources to process.

**3. Splitting the data into train and test sets :** oncewe have trained our classifier we can’t test it on the same data set we used for training, thus before training a classifier we split our dataset into train and test sets.

**4. Build the classifier :** we will use K Nearest Neighbors algorithm for classifying the observations. KNN works by selecting majority value from the k nearest data points and classify the new observation .

**5. Prediction and visualization:** Once the classifier is build predictions are made along with calculation of accuracy. The final step includes performing visualizations and displaying the results.Tableau is a very famous tool used in Analytics. It supports lots of data analysis features like regression, classification and lot more.



**Objective**

In the endwe conclude our objective being effectively using the important data that is generated from the hospitals to find patterns using a well known Data mining algorithm known as KNN . Heart disease is no doubt has the largest proportions of deaths in the world and researchers are continuously working in this field to find a way out to solve this problem. We discovered some new methods to load the data sets, pre process the data and build the classifier in C language and also tried to minimize the time complexity of our code.

**Methodology**

K Nearest Neighbors is one of the most essential classification algorithm and is widely used with small data sets to classify the observations. The algorithm belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. After splitting the dataset into train and test sets build KNN classifier using the algorithm below

**Algorithm:**

Let number of training data samples = n

new observation = m

1. Data points are stored in arr[]

2. for i= 0 to n calculate euclidean distance between m and other data points

sqrt((arr[i].x -m.x)\*(arr[i].x - m.x) + (arr[i].y - p.y) \* (arr[i].y - m.y));

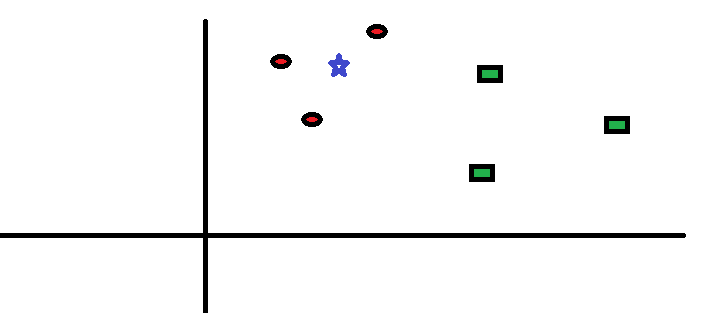
3. Make set A of K smallest distances obtained.

4. Return the majority label among A.

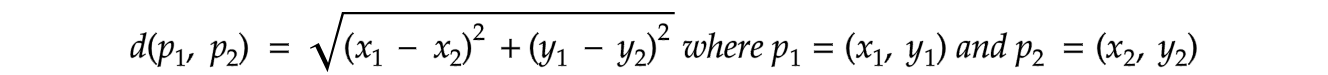
A distance function provides distance between the elements of a set. If the distance is zero then elements are equivalent else they are different from each other. The KNN algorithm uses euclidean distance as a distance metric , we have other distance measures in data mining such as Minkowski Distance, Manhattan Distance , and Chebychev Distance etc.

**How KNN works:**

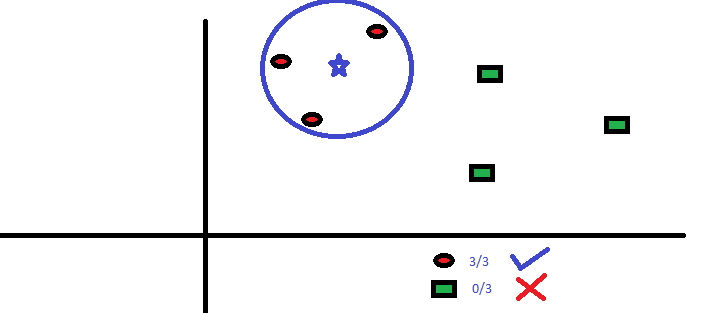
**1.** Load the data and initialize K to your chosen number of neighbors



2. For each example in the data calculate the distance between the query example and the current example from the data and add the distance and the index of the example to an ordered collection.

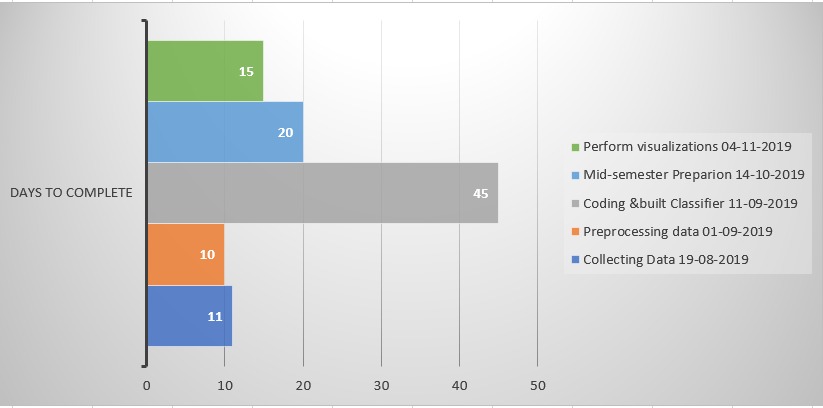


3. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances and pick the first K entries from the sorted collection.



4. Get the labels of the selected K entries . If regression, return the mean of the K labels. If classification, return the mode of the K labels.

**Plan of work**

** Gantt Chart**

**References**

* Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.
* K. Sudhakar, Dr. M. Manimekalai, “Study of Heart Disease Prediction using Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, pp.1157-60, January 2014.
* C. Ordonez, “Association rule discovery with the train and test approach for heart disease prediction,” IEEE transactions on Information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society, vol. 10, no. 2,pp.334–43,Apr.2006.
* H. Almarabeh and E. Amer, “A study of data mining techniques accuracy for healthcare,” International Journal of Computer Appli-cations, vol. 168, no. 3, pp. 12–17, Jun 2017.
* [N. Amma, “Cardiovascular disease prediction system using genetic algorithm and neural network,” in International Conference on Computing, Communication and Applications. Dindigul, Tamilnadu, India: IEEE, Feb 2012, pp. 1–5.](https://ieeexplore.ieee.org/abstract/document/5569740)
* S. Amin, K. Agarwal, and R. Beg, “Genetic neural network based data mining in prediction of heart disease using risk factors,” in IEEE Conference on Information Communication Technologies. Thuckalay, Tamil Nadu, India, April 2013, pp. 1227–1231. https://doi.org/10.1109/CICT.2013.6558288.
* B. Venkatalakshmi and M. Shivsankar, “Heart disease diagnosis using predictive data mining,” International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, no. 3, pp. 1873–1877, 2014.
* H. Masethe and M. Masethe, “Prediction of heart disease using classification algorithms,” in Proceedings of the world congress on Engineering and Computer Science, San Francisco, USA: Interna-tional Association of Engineers (IAENG), 2014, pp. 22–24.
* S. K. Yadav and Pal S., “Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification”, World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.
* Vikas Chaurasia, and Saurabh Pal, Data Mining Approach to Detect Heart Dieses, International Journal of Advanced Computer Science and Information Technology, Vol. 2, No. 4, pp. 56- 66, 2013.