Heart stroke detection using orange tool

# Affiliation:

### Department of Computer Application, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India)

In the medical management sector, extraction

#### Abstract:-

Heart disease is one of the most common diseases today, many factors are the cause, including high blood pressure, diabetes, changes in cholesterol, and fatigue and early diagnosis of such diseases. It is a major cause of disability. Major factor is high BP. The most common cause is blockage of heart or coronary arteries. People also have heart attacks and the major reason is coronary artery disease that happens slowly over time. Some heart problems may happen to the valves in the heart, heart failure.Some key steps to control risk of heart disease is control the B.P, lower the cholesterol, not to smoke and get enough exercise. So based on all these symptoms and effects, we are going to predict the heart stroke disease and the data model uses for these prediction is K Nearest Neighbor (KNN), Naive Bayes, Random Forest, Linear Regression, Gradient Boosting and Decision Table classifiers to show the performance of the selected classification algorithm, and uses training and test samples for a classification and / or optimal prediction.

**Keywords:-**Data mining; orange; Classification; Prediction; KNN; Naïve Bayes; Random forest; Heart Stroke; Heart disease.

**Introduction (Data Mining):-**

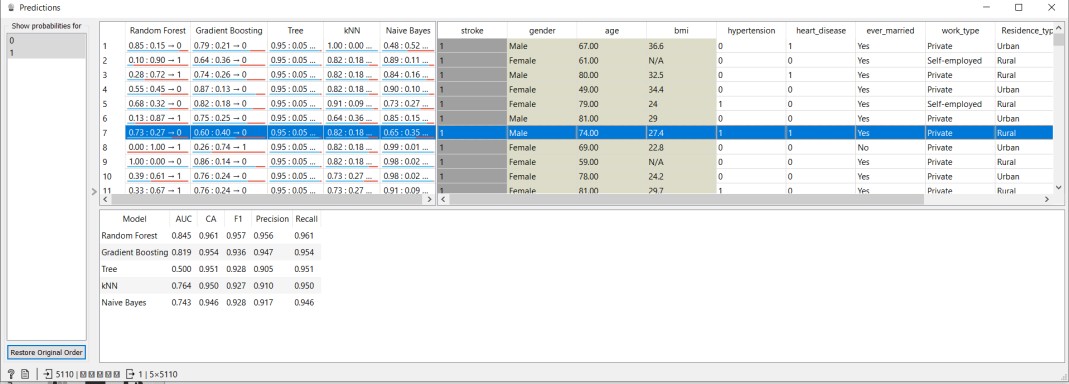
Data mining is very useful for information extraction. Data mining is used for commercial and research purposes. It has been popularly as well as broadly used next to many productions.

of data is fetching extra and admiring day by day. This mining technique can be of great advantage to most of the people engaged in the medical industry. e.g; information extraction can support the medical industry to detect phishing and misuse, manage customer relationships, functional patient supervision and leading execution, and inexpensive health protection. Maybe. The huge quantity of information caused by health concerns are so complex and enormous to be refined and analysed using traditional techniques. In this treatise, we mainly describe the behavior of information about heart stroke related to heart disease. This task will include a brief survey on the application and use of data mining on health issues. Data analysis algorithms are applied to the medical industry and help the general public find out how old, new, and common illnesses are occurring. This field can find many applications including pharmaceutical and hospital management.

This project tends to design and evolve a system for diagnosis and prediction of heart disease based on prognostic mining. Heart disease is a term assigned to many health conditions related to the heart. This disease refers to an unusual health condition that directly strikes the heart and all its parts. Heart disease is a critical health problem in just a few days. It is an open source data visualization software based on components, machine learning and data mining toolkit, suitable for machine learning and data mining

i.e. visualization. These components are called widgets and range from visualization to pre- processing of data for exploratory analysis, evaluation, and predictive modeling. Orange is presented to both experienced users and

analysts in data mining and machine learning. These users want to create and test their own algorithms while reusing code as much as possible. It has also been suggested that as many users as needed in the field can also write short python content of data analysis. [2]



#### Methodology:

The widgets that we have used for data classification are Naïve Bayes, KNN and Decision Tree and Random forest. After applying all the classifiers, we found which classifier gives the best and accurate result on the heart stroke dataset and according to that we will choose the classifier widget for the model.

#### Classifiers Used:

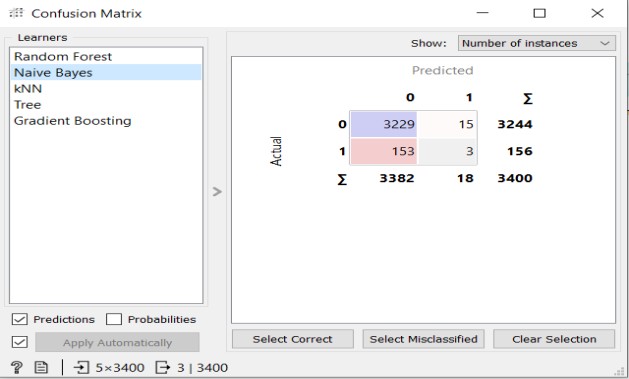
The classifiers are used such as Recall, Naïve Bayes, kNN, Precision, Decision tree algorithm, Random Forest, Gradient Boosting.

#### Recall:

In simple terms, we can say completeness or fullness. Recall finds, what percentage of positive tuples or rows did the classifier label as positive. In orange tool, Recall is represented as sensitivity.

***Recall = TP***

***TP + FN***



#### Precision:

In simple terms, Precision is also called exactness or we can say correctness. Precision finds what percentage of tuples or rows that the classifier labelled as positive are actually positive.

***Precision = TP***

***TP + FN***

***Note: Here, TP represents (True Positive), FP (False Positive) and FN (False Negative).***

## Comparative study of classification algorithm

**Validation:**

### We have used ‘confusion matrix’ widget in our workflow to validate the actual and predicted values for a better analysis and understanding.

#### Naïve Bayes:

This classifier algorithm is a contingency algorithm which evaluates the conditional likelihood of the target or class variables from the training data and approaches them to predict new data objects.. The algorithm is very vast for discrete functions, but least measurable for continuous functions. Naive Bayes data mining classifier technology has been applied, which uses the smallest training set to generate the best predictive model to predict the chances of heart disease patients. Comparative analysis for heart-stroke datasets in orange for Recall is 0.948, Precision is

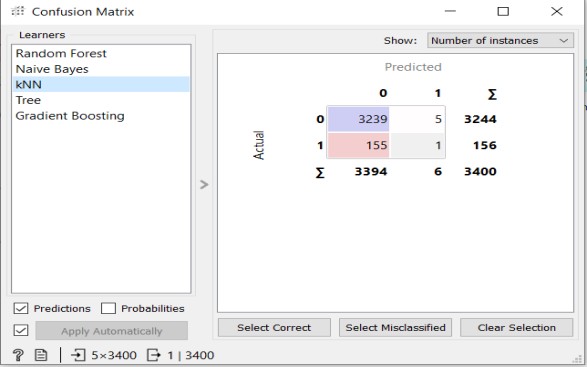
0.921 and classification accuracy is 0.932.

accuracy is 0.932.

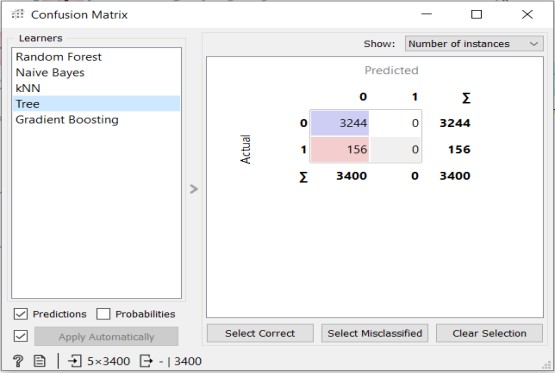
#### kNN Algorithm:

It is also called the K nearest neighbour algorithm. KNN is a strong and powerful classification algorithm which is mostly used in pattern recognition. It accepts all available cases and classifies new cases based on similarity measures like distance function. However, it is primarily used to categorize forecasting issues in industry. Comparative analysis for heart-stroke datasets in orange for Recall is 0.954, Precision is 0.929 and classification accuracy is 0.954. [3]

#### Decision Tree:



The decision tree is a regulated learning calculation, reasonable for discrete and ceaseless factors. Gap the dataset into subsets dependent on the main ascribes of the dataset. To discover the accuracy of the model, a test set comprising of test tuples and class marks is utilized. The level of test set tuples is effectively grouped by the model to distinguish



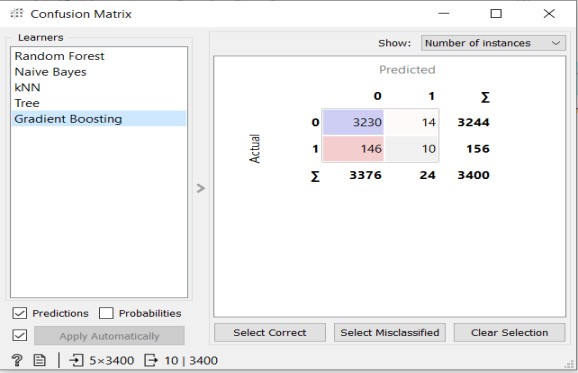
the accuracy of the model. On the off chance that the still up in the air to be exact, it is utilized to group information tuples with obscure classification marks. The choice tree assembles a grouping or relapse model inside the kind of a tree structure relapse model as a tree structure. It breaks down the dataset into more modest and

more modest subsets, while slowly creating related choice trees. Relative examination for heart-stroke datasets in orange for Recall is 0.954, Precision is 0.910 & exactness is 0.954.

[4]

#### Gradient boosting:

Gradient boosting is an AI strategy utilized for relapse and characterization issues. It creates prescient models as a bunch of frail prescient models (generally choice trees). Gradient enhancement algorithms can be used not only to predict continuous target variables (as regressors), but also to predict categorical target variables (as classifiers). When used as a regressor, the cost function is root mean square error (MSE), when used as a classifier, the cost function is log loss. Comparative analysis for



heart-stroke datasets in orange for Recall is 0.954, Precision is 0.936 and classification

accuracy is 0.954.

#### Random Forest:

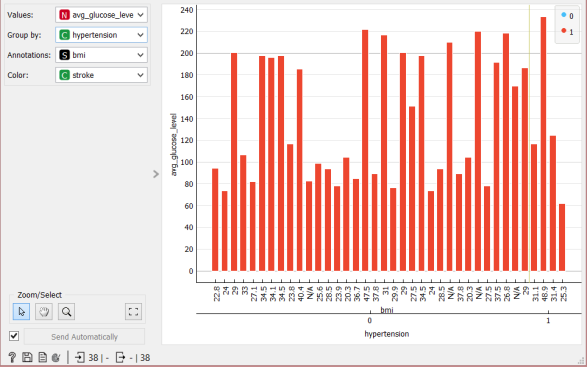
Random Forest is a managed learning calculation. It is a bunch of choice trees, which are by and large prepared utilizing the "sacking" strategy. The overall thought of the sacking technique is that the combination of learning models builds the general outcome. The arbitrary backwoods works in two phases. The primary is to make a random forest by combining N decision trees, and therefore the second is to predict each tree created within the first stage. The main reason for using Random Forest is that it requires less training time than other algorithms and can predict the output with high precision even for large data sets that are executed efficiently. Accuracy is maintained even when large proportions of data are missing. Comparative analysis for heart-stroke datasets in orange for Recall is 0.966, Precision is 0.961 and classification accuracy is 0.966.So, after comparing the datasets with all the above

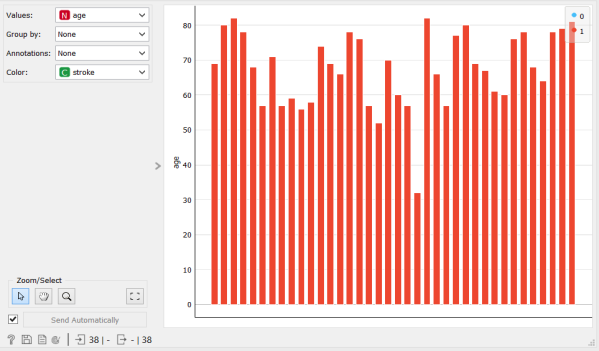
classifier, we found that Random Forest gives the best and accurate result

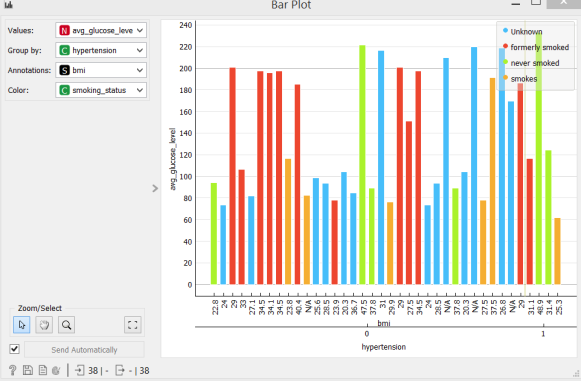
#### Comparative Analysis Chart: comparing Recall, Precision and Accuracy for heart-Stroke disease:-

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Recall | Precision | Accuracy |
| KNN | 0.954 | 0.929 | 0.954 |
| Tree | 0.954 | 0.910 | 0.954 |
| Random Forest | 0.966 | 0.961 | 0.966 |
| Naïve Bayes | 0.948 | 0.921 | 0.948 |
| Gradient Boosting | 0.954 | 0.936 | 0.954 |

**Comparative bar chart of avg\_glucose\_level and hypertension by taking x and y axis respectively**





**Heart Disease:** Hypertension is a significant danger factor for coronary illness and stroke since it harms the covering of the supply routes, making them more defenceless against the building up of plaque, which limits the corridors bringing about the guts and mind. [5]

**Average glucose level:** The average blood glucose of diabetes can damage neurons that govern blood vessels and nerves. Over time, these damages can lead to heart disease and it affects person’s cardiovascular system. So that, this attribute leads to be considered as major point.

## Discussion:

#### Data set used:

We have created a dataset (training data) of real time data that we collected from various sources. The file contains heart disease data for 5111 different patients.

Another dataset (test data) of 510 data that is to be used for predicting the number of strokes of heart disease patient in the model. We have taken the attributes such as gender, hypertension, heart disease, work type, Average\_glucose\_level, BMI, smoking\_status, and number of strokes to predict the heart stroke detection value (Target Variable) of heart stroke where **1 means has a stroke and 0 means not has any stroke.**

#### Attributes Taken:

**Gender:** On whom we will predict the heart- stroke disease so, as we have taken the gender attribute and classified this attribute as male and female for prediction.

**Hypertension:** Hypertension is another basic condition generally occurred because of hypertension. A heart that works with such high pressing factor can make different heart infections. Hypertensive coronary illness incorporates cardiovascular breakdown, thickening of the heart muscle, coronary conduit infection, and different conditions, all of which can prompt a respiratory failure. In our model we have taken the hypertension attribute in the form of binary as (0 and 1).

**BMI:** BMI is a gauge of muscle versus fat and is a decent pointer of the danger of ailment that can happen with more muscle versus fat. The higher your BMI, the higher your danger of getting certain sicknesses, like coronary illness. It was tracked down that the higher the BMI, the higher the troponin level. The most obese people in 12 years (BMI 35 and above) had the most heart failure.

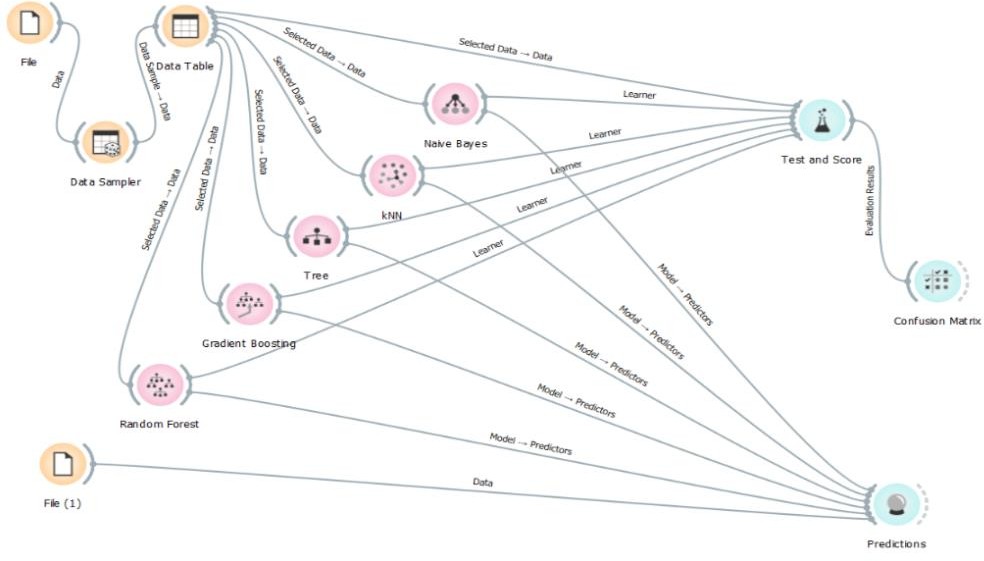
**Smoking status:** Smoking is a well-known risk factor for atherosclerosis. However, there are insufficient data on the predictive effect of smoking status on the risk of accidental heart failure. Heart disease is responsible for 1 in 5 smoking-related deaths. People over the age of 35 who are taking smoking pills have a much higher risk of heart disease and stroke.

**Stroke:** In the model, Stroke is the main attribute or we can say that its target or class attribute depending on which we are predicting the heart disease. This attribute has been further classified into binary form which is 0 and 1 i.e.; 0 means has not any stroke analysed and 1 means has analysed the stroke.

#### Data mining Tool Taken:

The orange data mining tool taken for the project work. Because Orange is the best open- source data mining tool. Its user friendly and interactive. It provides data manipulation, visualization, modelling, and evaluation etc. tools for data mining. We can save the workflow and can export it as per our need on later part.

#### Orange Tool Workflow:

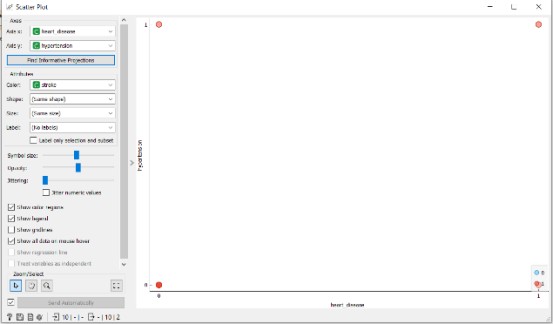


We can clearly see that the confusion matrix for gradient boosting predicts the most accurate value for target variable among all other classifiers.

## Observation:

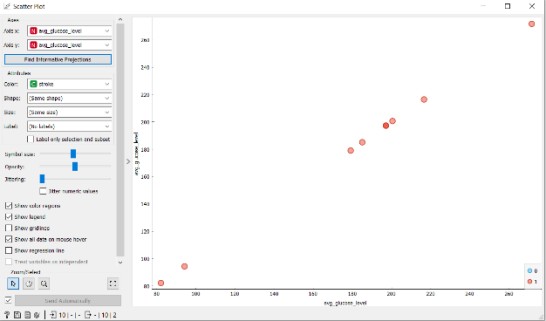
### C:\Users\Vicky-PC\AppData\Local\Microsoft\Windows\INetCache\Content.Word\c1.pngWe can see from the prediction widget and confusion matrix that the classification accuracy of the decision tree is higher among all other classifiers in predicting the portability value for our given dataset.

#### Comparing Classification Algorithms through Scatter Plot:



**Comparison by taking smoking\_status and heart\_disease factor in x and y axises respectively**

#### Comparison by taking hypertension and heart\_disease factor in x and y axises respectively



**Comparison by taking avg\_glucose\_level factor on both the axis.**

## Advantages and Limitations:

#### Advantages

Increased accuracy for effective heart disease diagnosis. Manages huge amount of data using random forest algorithm and feature selection. Decreases the time complexity of doctors. Cost effective for patients.

#### Limitations

Prediction of cardiovascular disease results is not accurate. Data mining techniques doesn't help to

supply effective deciding .Cannot handle huge datasets for patient records.

## Conclusion and Future Scope:

Information mining methods assist with tracking down the secret information during a group of illness information which will remain wont to dissect and anticipate the more drawn out term conduct of infections. Arrangement is one among the records mining technique which appointed a class mark to a gathering of unclassified cases. The primary target concerning this paper is to coordinate with the information mining devices on the possibility of theirs order precession and review. As per the aftereffect of three information mining apparatuses utilized in this paper, such has been noticed so various information mining instruments are outfitting various outcomes concerning same informational collection with various grouping calculation. WEKA and ORANGE are showing best grouping Precession and Recall. In future, more sickness dataset are frequently utilized for grouping strategies, and different information handling procedures like bunching are regularly utilized predictable with think about the presentation of shifted information preparing devices. [6]

Perhaps the most disadvantage of those works is that the primary spotlight has been on the machine of grouping methods for heart condition forecast, rather than examining different information. It is seen that an exact cleaned and pruned dataset gives preferred precision rather over messy one with missing qualities. Choice of appropriate procedures for information cleaning close by legitimate characterization calculations will cause the occasion of forecast frameworks that give upgraded exactness. In future a canny framework could likewise be created which will cause determination of legitimate treatment techniques for a patient determined to have heart condition. A great deal of work has been done effectively in making models which will foresee whether a patient is most likely going to foster heart condition or not.

There are a few treatment techniques for a patient once determined to have a particular kind of heart condition. Information mining are regularly of great assist with picking the street of treatment to be trailed by separating information from such reasonable data sets. [7]

# References

1. Dan Emerson, Orange Date Mining fruitful and fun, New York: Big Data Made Simple, 2006.
2. N Ch Sriman Narayana Iyenger, Heart Disease Prediction System Using Data Mining, August 2016.