Visualization and Prediction of Heart Disease Using Interactive Dashboard

Visualization of data

Data visualization is a crucial phase in the data science process that enables teams and individuals to communicate data to coworkers and decision-makers more effectively. Teams that oversee reporting systems frequently use predefined template views to keep an eye on efficiency. However, performance dashboards aren't the only applications for data visualization. For instance, while text mining unstructured data, an analyst might employ a word cloud to identify important ideas, patterns, and undiscovered connections. As an alternative, they can show the connections between things in a knowledge graph using a graph structure. It's crucial to keep in mind that there are numerous methods to represent various sorts of data, and that this is a set of abilities that should go beyond your core analytics team.

In order to make the dataset we are working with easier to understand, we will first take a look at data Wrangling.It would enable us to make better use of the data. We have to import pandas, matplotlib and seaborn. We can now conduct exploratory data analysis after finishing data wrangling.

A timely diagnosis of heart disease (HD), one of the most prevalent diseases today, is essential for many healthcare professionals to protect their patients from the condition and save lives. To accurately classify and/or predict HD cases with few variables, a comparison analysis of various classifiers can be done for the classification of the heart disease dataset. Accurate decision-making and ideal therapy are needed to address cardiac risk. Five machine learning models were utilized in a Canadian study to examine 1-month mortality among hospitalised patients with congestive heart failure. Several tests, including auscultation, blood pressure, cholesterol, ECG, and blood sugar, are carried out prior to the diagnosis of a condition. These tests assist in identifying the patient's medication requirements. In this work, the predictive accuracy of various machine learning methods is investigated to calculate cardiovascular risk. The performance comparison of the most recent REP Tree and Random Tree machine learning algorithms in terms of cardiovascular disease prediction is innovative.

Pre-processing of Data:

The real-world data has many missing and noisy values in the initial stage of data mining. To avoid these issues and produce precise predictions, these data have been pre-processed. The raw data is unreliable and insufficient. You can eliminate the missing numbers or use the mean value in their place. Therefore, employing a filtering strategy, the data obtained must be slightly adjusted in order to conduct a good study. Here, the multifilter method is applied.

Extraction of Features:

Reduce the quantity of input attributes prior to data processing. Not all characteristics affect prediction success in the same way. Multiple attributes lead to increased complexity and worse performance . It is necessary to carefully extract features without sacrificing system performance as a result.

Methods of Machine Learning:

REP Regression tree: the tree generates many trees over various iterations. As a sample of all the created trees, it selects the best tree. Think about removing predictions from the tree using the mean square error. Reduced Error Pruning (REP) quickens learning and creates decision trees based on the knowledge acquired. As a result, even when working with vast volumes of data, REP offers a simpler and more precise classification tree.

M5P Tree: For numerical prediction, the M5P model tree is utilised. Each layer creates a linear regression model that maintains its predictions for the class value of instances. Splitting the training data's T part yields the best characteristics.

K-means Clustering

The K-means clustering algorithm was chosen for the web dataset due to its scalability, efficiency, ability to construct populations of equal size, and simplicity. In order to classify groups of data points, the K-means algorithm requires a minimum total of squares. The dataset in question has 209 observations for 7 variables. The steps listed below are used to calculate the cluster's initial center..

1) Detect K random clusters

2) Find the important clusters iteratively

3) If the distance between the observation and its nearest cluster centre is greater than the distance between the other closest cluster centres, the observation is substituted with the nearest centre by computing the Euclidean distance between the cluster and the observation.

4) The total of squares within a cluster is determined as follows:

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

where S_k is the set of observations in the *k*th cluster and $\bar{x}_{kj}$ is the *j*th variable of the cluster center for the *k*th cluster.

The Final Cluster Centres are reached when the difference between the sum of the squares in two successive iterations is at a minimum.

Age, maximal heart rate, the nature of the chest pain, and the condition are all factors taken into consideration when predicting heart disease. The findings are addressed individually after taking into account four different forms of chest discomfort.

SVM Algorithm:

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

We use a standard machine learning procedure. First, we gathered reliable ECG data. The data was then preprocessed, with selected attributes that will influence the target and outliers removed, and the data was split into test set and train set, with train set used for training algorithms and test set used to test the models.

Data Collection and Processing Data Collection

The dataset was compiled from multiple patient records that included 14 attributes such as restagc, fbs, thal, ca, cp, sex, age, thalach, chol, trestbps, slope, oldpeak, exangand target. The table below contains a description of the attributes used for analysis. The visualization shows that age does not play a significant role in predicting heart disease because the same age groups have an equal number of people with and without heart disease. Outlier detection and data preprocessing For each attribute or feature, we calculate the Z-score of each individual value of that attribute in relation to the column mean and standard deviation. After that, take the magnitude or absolute value of the obtained z score. If the z score is less than a certain threshold, a particular row or record is outlier and is removed.

Data Transformation: (Standardization)

It is a technique used to scale where the values are scaled around the mean with standard deviation as unity.

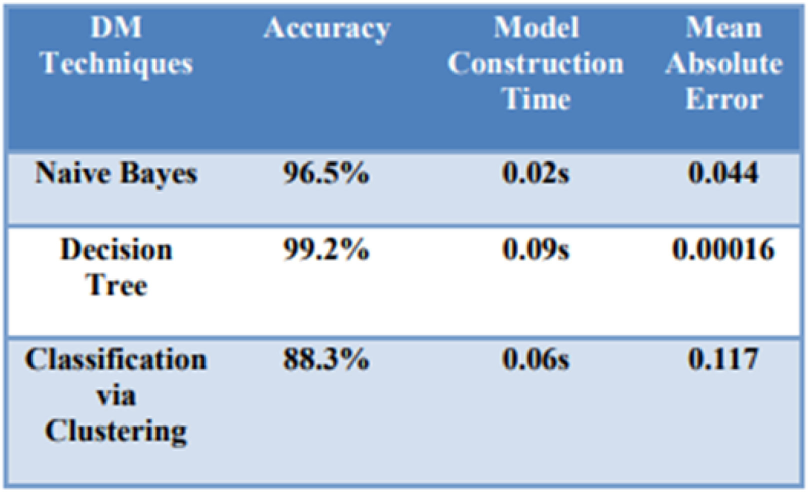
$$\mathbf{X=X-\mu\sigma}$$

Attribute Selection (Pearson Correlation)

This measures the relationship between two sets of data. The value of the correlation varies from −1 to 1. Here+1 denotes a strong positive correlation, a strong negative correlation is denoted by −1, and no correlation is zero. If the absolute value of the correlation is less than 0.2, then it means the two datasets are not correlated and can omit during analysis. If an attribute having correlation closer to zero, then that attributes can be eliminated. is the graph showing importance of each feature and it is clear that ‘thal’ is a more important feature and ‘age’ is a less important feature.

Genetic Algorithm

Natural evolution theory is incorporated into the Genetic Algorithm . The beginning population of the genetic search has randomly generated rules and no starting traits.A new population was created to match the fittest rules in the existing population and any progeny of these rules based on the concept of survival of the fittest. By using the genetic operators of cross over and mutation, offspring were produced. The generation process lasted until a population P developed that satisfied all of its rules. In addition to the genetic algorithm, the CFS Evaluator is also used. Weka 3.6.0 tool is used to conduct the observations.909 records with 13 attributes made up the first data set .Consistencies were eliminated by categorizing all attributes, and resolved for ease of use. after being reduced from 13 to 6 qualities. On the dataset, different classifiers are applied. matching these 6 characteristics of heart disease prediction. There is a performance analysis of these classifiers. It can be perceived from the table that Decision Tree has outperformed with highest accuracy and least mean absolute error.



DM Techniques	Accuracy	Model Construction Time	Mean Absolute Error
Naive Bayes	**96.5%**	**0.02s**	**0.044**
Decision Tree	**99.2%**	**0.09s**	**0.00016**
Classification via Clustering	**88.3%**	**0.06s**	**0.117**