**PROJECT: Heart Disease Prediction Model using Machine Learning(ML) Algorithms**

**Abstract**

The most decisive work in the field of healthcare is diagnosis of disease. If a disease is diagnosed at early stages then the chances of saving my lives will be increased. Machine learning(ML) algorithms can assist the medical field by providing a correct and rapid diagnosis of diseases. Hence, it will save time of both doctor and patient. Heart disease is the major cause of death in the world today, it becomes the most crucial disease to diagnose. In the report a survey of the machine learning algorithms has been put forward to help healthcare professionals in diagnosing heart disease. Report will start by explaining machine learning and briefly describing most commonly used classification techniques to diagnose heart disease. Then, we evaluate representable research works on using techniques of machine learning classification in this field.

**1. Introduction**

Artificial Intelligence(AI) is a segment of Computer Science(CS) that has the ability of making computers more brilliant. Among all the requirement the most basic requirement of artificial intelligence is learning so the subfield of artificial intelligence is called machine learning(ML). Machine learning is of great importance in the healthcare field since it is an intelligent tool to examine data and the medical field has lot of data. Due to the digital revolution in recent years lot of data was collected and stored. Data collection devices and monitoring devices that are available in contemporary hospitals are being used on daily basis and a lot of data have been collected. It is very difficult or nearly impossible for humans to manually analyze this collected data. Machine learning algorithms learns from previously detected cases of patients and it helps doctor to diagnose new patients with efficiency and higher speed.

Since in today's world there are numerous amount of medical dataset for which machine learning is required. Machine learning has the ability to discover hidden patterns and provide beneficial data from it. Although machine learning has many uses, it is majorly used in medical field for disease prediction.

Heart disease is also known as cardiovascular disease. It encloses variety of heart diseases that affect the heart. Cardiovascular diseases are the number 1 reason for death globally because more people die annually from this than any other cause.

**2. Background**

This background section provides descriptions of the topics that are related to this report like machine learning techniques with its techniques, data preprocessing, performance evaluation metrics

Machine Learning (ML) is a field of artificial intelligence that includes constructing algorithms that learns from past experience. Machine learning algorithms detect hidden patterns in the input dataset and build models accordingly. They can make correct predictions for new dataset that are entirely new for the algorithms and build model accordingly. This way through learning the machine becomes more intelligent, so it can identify hidden patterns that are very hard for humans to detect by themselves. Machine learning algorithms can work with large datasets and make predictions and decisions.

**2.1 Machine Learning Techniques**

The Machine Learning Techniques can be classified as follows:

**2.1.1 Supervised Learning**

In supervised learning technique a dataset is present with examples and their output i.e. response. Its algorithm learns from the dataset through a process named as training. It then responds to any new input based on its training.

**Examples of supervised learning includes:** Classification and Regression.

Input Data

Classifier

Dataset

Model

Prediction

**Fig.1: Simplified Representation of Machine Learning**

**Fig.1** depicts working of machine learning. In the above figure it is shown that it is the dataset which is process first. The patient's database is present in this dataset. This database is processed first. This first step named preprocessing is an important one and as it prepares the data for machine learning algorithm by cleaning it. The second part i.e. model can consist both single algorithm or it can multiple algorithm working together in a hybrid approach. The output of the model is called a classifier. Classifier is the place where the intelligence is and it will make a prediction. When the classifier receives input, it provides prediction without any manual help.

For an instance let consider model receives medical dataset as input data one of health person and one of unhealthy person's information. As the input data is entirely new to the classifier it will make prediction whether this new person is health/unhealthy based on its past data or past experience.

**2.1.2 Unsupervised Learning**

In Unsupervised Learning technique there is no dataset present with examples and their output. In this the algorithm tries to identify the similarities between values that are input. Also, it will categorize them based on their similarities.

**One example of unsupervised learning is:** Clustering method.

**2.2 Heart Disease Dataset**

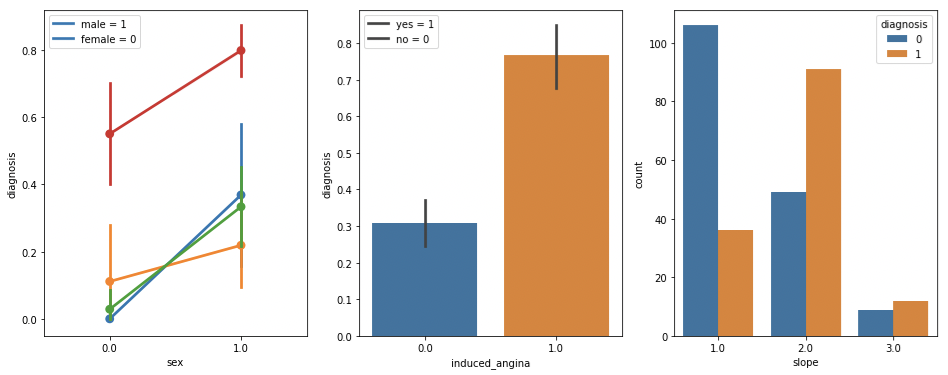
There is various heart disease dataset but one of the dataset that is used in majority is the dataset obtained from the University of California, Irvine C.A(UCI) Center for intelligent systems and machine learning. This dataset contains four databases from four hospitals. Each of these 4 databases have 14 features but have different numbers of records. The Cleveland dataset is mostly used by machine learning researchers as it contains less missing values than other datasets and also it has more records. The last field i.e. num depicts the presence of heart disease in the patient. It is integer valued in the range of 0-4 i.e. 0, 1, 2, 3, 4. In this 0 depicts no presence of heart disease. This dataset contains 303 records.

**3. Dataset Attributes**

|  |  |  |
| --- | --- | --- |
| Number | Attribute | Description |
| 1 | Age | Age in years |
| 2 | Gender | Male/female |
| 3 | cp | Chest pain type |
| 4 | trestbps | Resting blood pressure |
| 5 | chol | Scrum cholesterol in mg/dl |
| 6 | fbs | Fasting blood sugar |
| 7 | restecg | Resting electrocardiographic results |
| 8 | thalach | Maximum heart rate achieved |
| 9 | exang | Exercise induced angina |
| 10 | oldpeak | Depression induced by exercise relative to rest |
| 11 | slope | The slope of peak exercise segment |
| 12 | ca | Number of major vessels colored by fluroscopy that ranged between 0 and 3. |
| 13 | thal | Normal, Fixed defect, Reversible defect |
| 14 | num | Angiographic disease status (Diagnosis of heart disease) |

**Table 1.** depicts the 14 features/attributes that are present in the dataset with the description of each attribute.

**4. Explore the Data**



**The above output includes pairplot and 2 barplots (left - right)**

Men are much more prone to get a heart disease than women. The higher number of vessels detected through fluoroscopy, the higher risk of disease. While soft chest pain may be a bad symptom of approaching problems with heart (especially in case of men), strong pain is a serious warning! Risk of getting heart disease might be even 3x higher for someone who experienced exercise-induced angina. The flat slope (value=2) and downslope (value=3) of the peak exercise indicates a high risk of getting disease

**5. Modelling, Predicting and Comparison with Machine Learning Algorithms**

**5.1. K- Nearest Neighbours (KNN)**

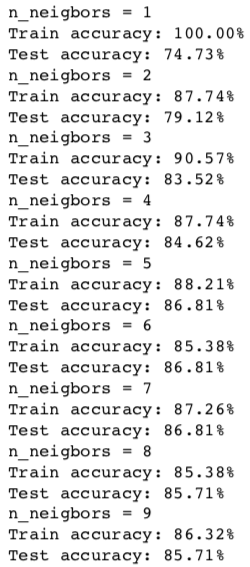
K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. The principle behind nearest neighbour methods is to find a predefined number of training samples closest in distance to the new point and predict the label from these.

**OUTPUT:**

****

Despite its simplicity, the result is very promising. Let's see if KNN can perform even better by trying different 'n\_neighbours' inputs.

**OUTPUT:**



It turns out that default value of n\_neighbours (5) is optimal.

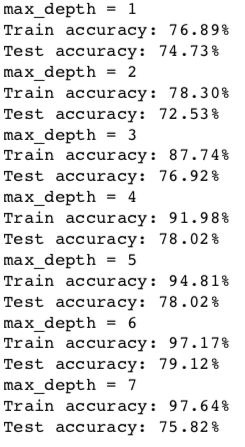
**5.2. Decision Trees**

DT algorithm creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It is simple to understand and interpret and it's possible to visualize how important a particular feature was for our tree.

**OUTPUT:**

****

**OUTPUT:**

****

**5.3. Logistic Regression**

Logistic regression is a basic technique in statistical analysis that attempts to predict a data value based on prior observations. A logistic regression algorithm looks at the relationship between a dependent variable and one or more dependent variables.

**OUTPUT:**

****

Negligible difference between train and test score tells us as that model performs at the optimal level. Although the result itself is slightly lower than KNN, yet is still satisfactory.

**5.4. Gaussian Naive Bayes**

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

**OUTPUT:**

****

This model produced the same result as leading KNN algorithm. While it slightly underfits the data, this model doesn't offer any hyperparameters for tuning and improve overall performance.

**5.5. Support Vector Machines**

Support Vector Machines are perhaps one of the most popular machine learning algorithms. They are the go-to method for a high-performing algorithm with a little tuning. At first, let's try it on default settings.

**OUTPUT:**



With impressive accuracy of almost 88%, Support Vector Machines are taking the lead!

**5.6. Random Forest**

Random 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.

**OUTPUT:**

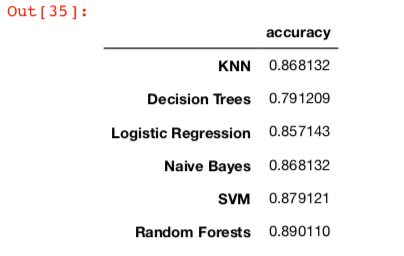


While it is typical for Random Forests to perfectly learn and fit into training data, the test accuracy achieved outstanding 89%!

**6. Conclusion**

The goal of the project was to compare different machine learning algorithms and predict if a certain person, given various personal characteristics and symptoms, will get heart disease or not. **Here are the final results:**

**OUTPUT:**

****

It does not come as a surprise that the more complex algorithms like SVM and Random Forests generated better results compared to the basic ones. It is worth to emphasize that in most cases hyperparameter tuning is essential to achieve robust results out of these techniques. By producing decent results, simpler methods proved to be useful as well.

Machine learning has absolutely bright future in medical field. Just imagine a place where heart disease experts are not available. With just basic information about a certain patient's medical history, we may quite accurately predict whether a disease will occur or not.

**References:**

<https://svitla.com/blog/data-science-vs-machine-learning>

<https://docplayer.net/12221294-A-prototype-of-cancer-heart-disease-prediction-model-using-data-mining.html>

<https://www.dezyre.com/article/top-10-machine-learning-algorithms/202>

<https://blog.codecentric.de/en/2017/09/data-science-fraud-detection/>

<https://www.researchgate.net/publication/322992160_Prediction_of_Heart_Diseases_using_Data_Mining_and_Machine_Learning_Algorithms_and_Techniques>

<https://www.sciencepubco.com/index.php/ijet/article/view/15714>

<https://www.expertsystem.com/machine-learning-definition/>

<https://www.omicsonline.org/open-access/heart-disease-diagnosis-using-data-mining-techniques-2162-6359-1000415.php?aid=88837>

<https://blog.goodaudience.com/what-is-machine-learning-simple-explanation-5afda4230ce9?gi=89a40951e7ed>

<https://docplayer.net/12221294-A-prototype-of-cancer-heart-disease-prediction-model-using-data-mining.html>

https://pdfs.semanticscholar.org/9a4f/97b89e2d8312200c00cc0cc0bc0e55b6335a.pdf

http://www.delhibusinessreview.org/v\_8n1/v8n1i\_word.pdf

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.739.9232&rep=rep1&type=pdf