IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9666

Development of Machine Learning Model for Diseases Prediction with an EMR Tool

Ranjitha Kulkarni[1], Prof. Seema Nagaraj[2]

[1]Student, Department of MCA, Bangalore Institute of Technology

[2]Assistant Professor, Department of MCA, Bangalore Institute of Technology

Abstract: Machine learning is gradually becoming more prevalent in the realm of medical diagnostics. Designing a Disease Prediction Model with the help of rapid development Application in the field of medical science with using machine Learning, would be helpful for the Doctor to predict and diagnose the diseases at the initial stage to take the necessary precautions. This Prediction is possible with the use of Machine learning techniques, libraries, Statistics and the EMR tool. The Calculation of the Prediction is carried out by implementing various Supervised Machine Learning Algorithms like Decision Tree, Random Forest and KNN. A large set of data can be processed in this system for improvising the accuracy level of Prediction. The data are split. An EMR tool records the data of the patients. The trained dataset were made to compare with test cases. To avoid multicollinearity, a feature has to be extracted and a Statistical Analysis must be carried out. We produce an accuracy level of 90% through this model. It can only predict whether the patient is effected or not and a precautionary steps to be taken. The Predicted results are sent to the patient's Mail ID. The Data are plotted by making us of confusion Matrix with a Binary data Labels i.e., Quantity Analysis.

Keywords: EMR, Disease Prediction, DT, RF, KNN.

I. INTRODUCTION

The Emergence of Artificial Intelligence has made the computer to be more intelligent than earlier [1]. Machine Learning is a part of an Artificial Intelligence program that enables the machine to examples of "self-learn" [2] by training data and improving it over a period of time, without being coded explicitly by the programmer. There are numerous kinds of Leanings in Machine Learning like supervised, semi-supervised, unsupervised, deep learning, and Reinforcement [3], this help to classify a huge data quickly.

Machine Learning, has become one of the prominent responsive tools for many applications and has created a great impact in several fields [4] like E-Commerce, Healthcare, Cloud Computing, and Big Data.

Machine learning may be used to diagnose, detect, and forecast a variety of disorders in the medical industry [5]. It makes use of different algorithms, models, and sample data which in turn helps in decision making.

This paper's primary objective is to provide tools i.e., Electronic medical records (EMR) for doctors to detect various diseases like COPD, Diabetes, and Heart (Cardiopathy) at an earlier stage. A disease prediction model is been designed for ill predictions and it makes use of Supervised Machine Learning Algorithms like Decision Trees, Random Forest, KNN. An EMR approach is also been used, which is a digital version of collecting the patient data and analyzing it. These EMRs have a set of recorded data and have advantages over paper records. As a result, patients will receive effective care and severe consequences will be avoided.

Data Processing and splitting are done for the accurate prediction of the data [6] in an effective manner.

Some of the Python Libraries are also used, like SKLearn, Numpy, Pandas, and Matplotlib for easier implication, Prediction, and Visualization [7] of the Diseases.

Only the introduction part is explained briefly in section I, section II represents the literature survey, section III represents the Proposed Schema, System Architecture, and Algorithms used, section IV represents Results / Discussions and section V represents the conclusion [8].

II. LITERATURE SURVEY

Numerous studies have been conducted to focus on the diagnosis of several diseases [9]. As the lifestyle of the people is been changing and because of all these, the people are affected with various diseases. These diseases have to be prevented and have to be detected in the initial stage so that they won't cause serious problems in the future.

The World Health Organization has made a list of the top 10 life-threatening diseases based on Global Health Estimates. There is a gradual increase in 4 of 10 leading diseases caused in 2000 [10]. They have highlighted a need for an immediate precaution for intensifying global affected diseases and treatment of Cardiopathy, respiratory diseases (COPD), Diabetes, and Cancer for sustainable development.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9666

Hence, most of the life-threatening diseases are highlighted and been Predicted in this paper using Machine Learning Technique and EMR approach is also used.

- **Prediction of COPD**

COPD (Chronic Obstructive Pulmonary Disease) is a disorder that affects and blocks the respiratory system. This disease is also commonly seen in people and is one of the third most lives threatening diseases. COPD population has increased in recent times and can cause a burden in the future [11]. The precautionary measure has to be taken on this disease and a prediction has to be done immediately.

- **Prediction of Diabetes**

Diabetes Mellitus or Diabetes has been described as equal to life-threatening diseases because even small children are also been affected by these. It is increased, when a person has high glucose levels over a protracted period in his body. Recently, it's been noted as a risk for developing Alzheimer's and a variety one causes of blindness & nephritis [12]. Prevention of the disease is additionally an important topic for research within the healthcare community people.

- **Prediction of Cardiopathy**

Health care has enormous data of information, and we need to process the data using certain techniques, data processing is one of all the techniques often used in machine learning algorithms, and cardiopathy is one of the leading explanations for death worldwide. Using this machine learning algorithm technique predicts the emerging possibilities of cardiopathy. The results of these methods provide the chances of occurring cardiopathy in terms of percentage [12]. The datasets used are classified in terms of the medical field of attributes like input and output data to split using the model selection method, this technique evaluates those parameters using the processing classification technique. [13] The datasets will predict the output using python programming.

III. PROPOSED SCHEMA

A. System Architecture

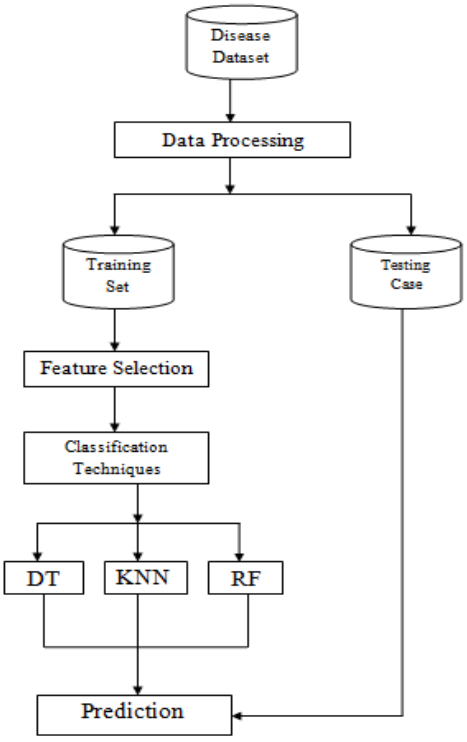

Fig 3.1 System Architecture

B. Proposed Model

- **Dataset**

Machine Learning Model is provided with a dataset as well as the data dictionary of the properties involved, which are studied in the Python environment.

- **Data Processing**

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9666

Data mutilation takes place which refers to the estimation of missing values in some variables, which is important because incomplete data prevents most interpretations. In the case of a continuous variable, mean value of the variables are replaced; and for the categorical variable, they are replaced with the mode value.

- **Data Splitting**

The data's are split into two, Training Set and Testing Cases, once data is processed. A Trained data contain all the data that is given as input for machine for learning and producing the prediction. Test Data contain all the EMR recorded which needs to be compared with trained for generating the prediction of the disease.

- **Feature Selection**

A Feature must be selected, which is essential for any predictive modeling and is done to avoid multicollinearity, remove redundant characteristics that are highly associated with one another, and improve the model's performance. We used the backward selection strategy to exclude qualities that aren't important for disease diagnosis.

- **Calculation Indices**

We start with all of the model's attributes and then eliminate them based on the p-value. When running a null hypothesis test in statistics, this aids in determining the significance of the results. The remaining variables were re-fitted into the model after the attributes with p-values greater than 0.05 were eliminated. This technique was repeated until each of the model's existing variables reached a meaningful level.

After each repetition, the updated R square value was recorded to calculate the fraction of variance explained by only those independent variables that have a real impact on the target variable's prediction.

- **Fitting and Testing**

Finally, the model is checked for its Fittings, Comparisons, and Tested, using the following feature selection, 3 classification algorithms were employed with the selected feature, including Decision Trees, Random Forest, and KNN, their prediction accuracy was compared using the Train/Test split approach with an EMR Tool. The test size for comparison was set at 0.1, which means that 90% of the dataset was used for classifier training and 10% was used for testing.

C. Algorithms used,

Machine Learning uses the algorithms which are Mathematical model mapping methods to perform classifications, pattern recognition, and prediction by using the training sets.

- **Decision Tree**

Finding the impurity of the input set in the Decision Tree is done via Information Gain. Information gain calculates the difference between the dataset's average entropy before and after splitting depending on the values of the provided attributes.

Formula:

$$\text{Info(D)} = -\sum_{i=1}^{m} pi \log_2 pi$$

A decision tree algorithm's fundamental principle is as follows:

Step 1: Sort the records by choosing the best attribute using Attribute Selection Measures (ASM).

Step 2: Create a decision node for that attribute and split the dataset into smaller parts.

Step 3: Recursively repeats this method for each kid to begin growing the tree until one of the conditions is met:

1. Each and every tuple has the same attribute value.
2. There are no more qualities left.
3. There are no longer any occurrences.

- **Random Forest**

Technically, it is an ensemble method of decision trees created on a randomly divided dataset (based on the divide-and-conquer strategy). Each tree casts a vote in a classification problem, and the class with the most votes wins out in the end.

Four steps make it work:

1. Choose arbitrary samples from a dataset.
2. Create a decision tree for each sample and use the results to make predictions.
3. Cast a ballot for each expected outcome.
4. As the winning prediction, choose the outcome that received the most support.

- **KNN**

Euclidean Distance method is one of the type of KNN that represents the shortest distance between two points.[Re]

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9666

Formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Four Steps are used for calculations:
Step 1: Determine the distance
Step 2: Locate the nearest neighbors
Step 3: Select labels

IV. RESULT / DISCUSSIONS

- **Dataset**

Dataset of the Patients is taken by Kaggle Website.

- **System Setup**

All the experiments carried out in this phase were implemented using Python as backend, Django as Python framework and frontend tool, and Sqlite3 as database. EMR tools were used along with numerous ML algorithms and feature extraction methods, run in an environment with an Intel Core i3, 2.10 GHz, Windows 10 (64 bit), and 4 GB of RAM system configuration.

- **Predicated rate**

The rate of accurate prediction on each disease and algorithm used is done and have obtained 90% and above.

Graph representation

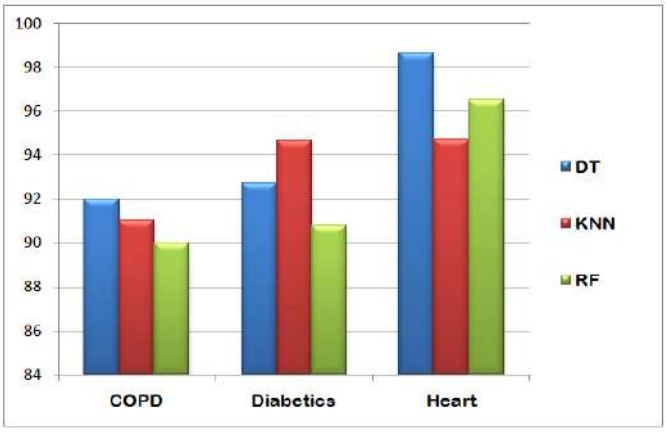

Fig 4.1 Disease Prediction with Algorithm used.

Fig 4.1 illustrates, the accuracy rate of DT (Decision Tree) is more in COPD Prediction, KNN (K-nearest neighbor) is more accurate in diabetics Prediction and the DT is more accurate and effective technique in Heart Prediction.

- **Patients Reports through Email**

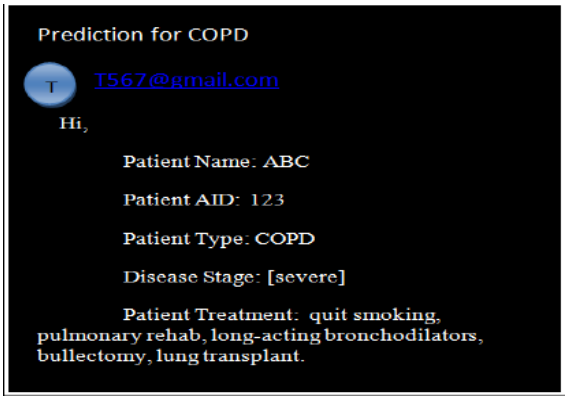

Fig 4.2 Patient Report

Fig 4.2 illustrates, A Sample of Patient report is generated, for every COPD Prediction.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9666

- **Visualization results of Prediction**

The Matplotlib Library makes use of Seaborn's Confusion matrix for Disease Prediction Visualization.

1. COPD

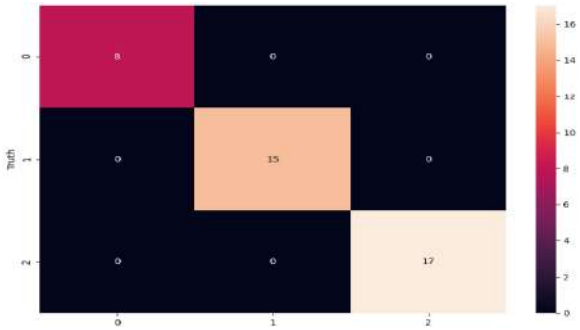

Fig 4.3 COPD Prediction

2. Diabetes

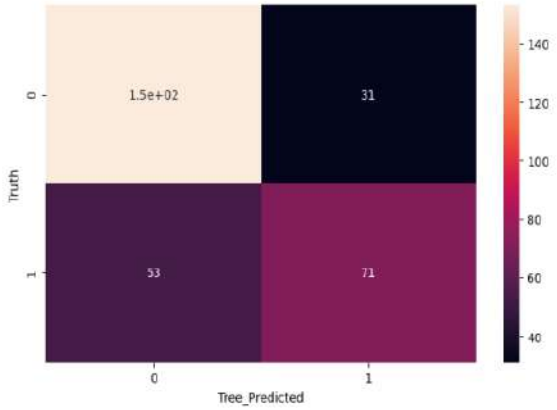

Fig 4.4 Diabetes Prediction

3. Heart

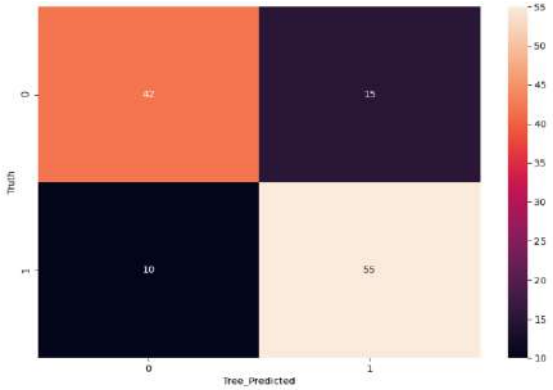

Fig 4.5 Heart Prediction

In most of the prediction problem, the prediction models results are stored in the confusion matrix. This help to count the value with Quantity Analysis. Two forms of Confusion matrix are used for plotting the Diseases:

1. A 2*2 Confusion matrix is used for plotting for Diabetics and Heart Disease.
2. A 3*3 Confusion matrix is used for plotting for COPD Disease.

It represents binary value that's true and false. False represents people are not affected and true represents people affected from disease.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9666

- **Comparison Results**

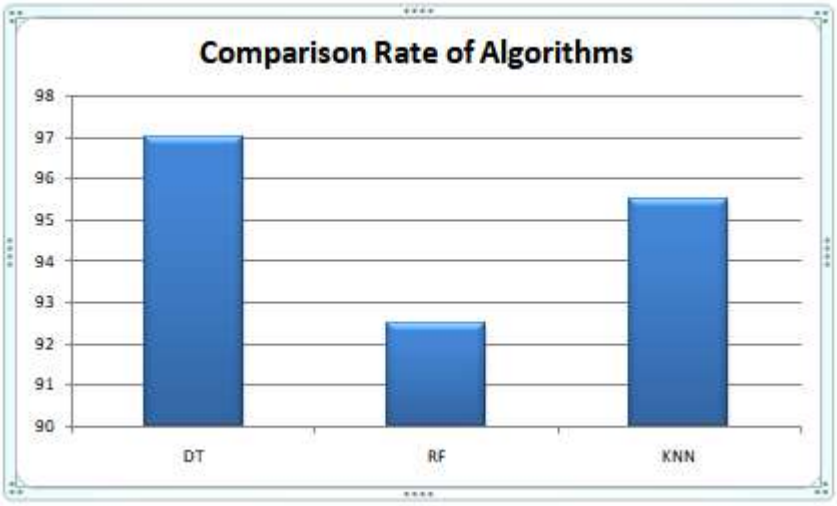

Fig 4.6 Comparison of Algorithms

Table 4.1 represents the comparison of Algorithms proposed in the system. Decision Tree is shown to be more appropriate Technique when compared to other two techniques.

V. CONCLUSIONS

A Disease Prediction Model is proposed for various diseases by utilizing three Supervised ML Algorithms namely, DT, KNN, and RF. A Better and a new Approach is been introduced and carried throughout the prediction. The results are shown as an Output Predicts the rate of accuracy of each ML Algorithms on Prediction Model. This system leads to less time consumption in a /and minimal cost. A comparison in the rate of accuracy is also figured out.

REFERENCES

[1]Marouane Fethi Ferjani,computing Department Bournment,England, "Disease Prediction Using Machine Learning", ResearchGate.
[2] Rinkal Keniya, Aman Khakharia Ninad Mehandale "Disease Prediction From Symptoms Using Machine Learning Algorithm", IEEE Paper
[3] Shagan Sah, "Machine Learning: A Review of Learning Types"
[4] "Disease Prediction Using Machine Learning Algorithm", IEEE Paper
[5] Virender Kumar Verma, Savita Verma, Machine Learning applications in healthcare section: An Overview.
[6] "Design and Analysis of Large Processing Techniques" ResearchGate
[7] V.Human Kumar, "Python Libaries, Development Frameworks and Algorithms for Machine Learning Application", IJERT
[8] Dhiraj Dahiwade, Gajanan Patle, Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach", IEEE Paper.
[9] Kennedy Ngure Ngare, "Heart Disease Prediction System", ResearchGate.
[10] Global Health Estimates- WHO, www.who.int
[11] "Prediction of COPD using Electronic Medical Records", researchgate
[12] Anand, A. and Shakti, D., 2015. Prediction of diabetes based on personal lifestyle indicators." IEEE.
[13] Pattekari, S.A. and Parveen, A., 2012. Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences.
[14] "Machine Learning Methods"-Ravil