JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Disease Prediction using ML

Sanket Muchhala
sanket.muchhala@gmail.com
B.E. I. T Student.
Thakur College of Engineering and Technology.

Hardik Sodhani
hardik.sodhani@gmail.com
B.E. I. T Student,
Thakur College of Engineering and Technology.

Shreeram Geedh
shreeramgeedh36@gmail.com
Associate Data Analyst,
Ugam Solutions
Mumbai, India.

Abstract:
Healthcare is a vast field in which computer technology is steadily being incorporated into various technologies, mainly Machine Learning algorithms and hospital-generated datasets. Supervised Machine Learning algorithms are Vindication in the healthcare industry. With the help of this project, we will detect the disease at the earliest stage and apply the necessary treatment. We are testing the accuracy of various models using the given dataset. To our knowledge, in large scale medical data analysis no prior work has addressed both types of data. Our proposed algorithm is more accurate with 94.8% calculation accuracy and faster convergence speed than other typical estimation algorithms as compared to CNN based unimodal disease prediction.

Index Terms **– ML, Decision Tree, Prediction, supervised learning.**

I. INTRODUCTION

Machine learning is computer programming to optimize performance using sample data or past data. Machine learning is the study of computer systems that learn from data and experience. The machine-learning algorithm has two parts: training, testing. Predict disease using symptoms and patient history Machine learning technology has been striving for decades. Machine learning technology provides an immeasurable platform in the medical field for health issues to be effectively resolved. We apply machine learning to keep complete hospital data. leading to the reference in the current text must match the list of references at the end of the document.

Abbreviations and Acronyms

ML- Machine Learning,

II. IMPLEMENTATION

2.1 Population and Sample

The dataset we're using is from Kaggle. Kaggle is leading platform for open-source free data sets for learning purposes and the datasets are accurate for models. Later, we decided to approach several hospitals for their datasets and then try our models on their dataset as their dataset would be accurate and real in life. Even surveys which asks people about their experience to the diseases can help us predict and would make our model more accurate. Right now, we are using sample dataset of 10k patients for disease, diabetes and COVID prediction.

2.2 Data and Sources of Data

MEDLINE, EMBASE, CINAHL, ProQuest, Scopus, Web of Science, Cochrane Library, INSPEC and ACM Digital Library were searched on 1 November 2021.

We are predicting disease on the bases of symptoms entered by the user. Then the symptoms are used as input in the ML models we are using. For our Project we are using Decision Tree Prediction, Random Forest Algorithm along with Naïve Bayes Algorithm. There are certain steps we followed while building our model.

2.3 Methodology

i. Collecting Data.

We get data in form of .xlsx files in Microsoft excel or text files. But are gathering the data from various sources for our model for a better prediction and better accuracy of our model. The more the data the higher the accuracy of the model and the accuracy of the model depends on the quality of data.

ii. Preparing the data.

The quality of the data used in any analytical procedure is critical. It is imperative to allot time and effort to establish data qualities and then prepare a model to address problems of missing data and outlier handling. Exploratory analysis is one approach for delving deeper into the subtleties of the data and so expanding the nutritional content**.**

iii. Training a model,

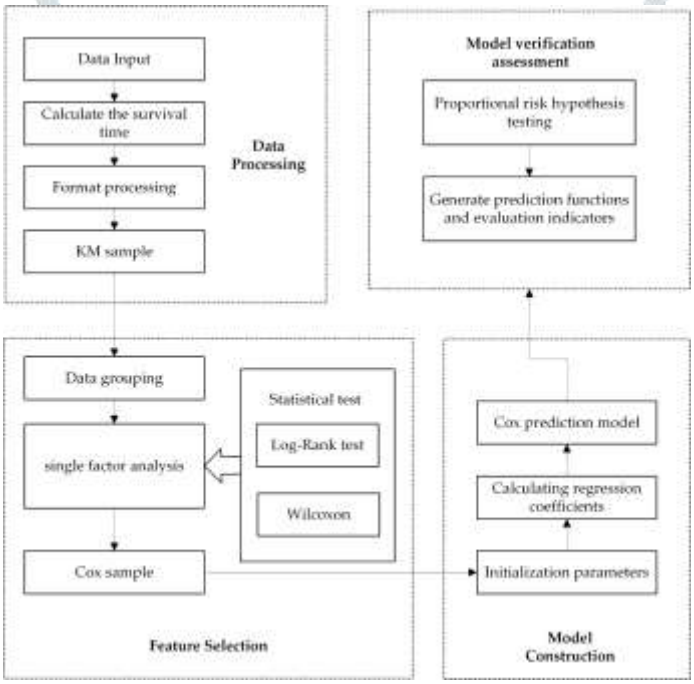
This step entails determining the best method and data representation in the form of a model. The cleaned data is divided into two parts: train and test (proportions vary depending on prerequisites), with the first (training data) being used to create the model. The second section (test data) serves as a guide.

iv. Evaluating the model.

We need to test our model based on algorithms. Working and using algorithms and testing them for using the maximum efficient algorithm on our data set depending on our use and requirement.

v. Improving the performance.

Using the algorithms which give maximum efficient, best results and highest accuracy for given dataset. Thus, the developed model would be efficient and adaptive for the data which is going to be stored and processed.



Algorithms:

1. Decision Tree

It's a supervised learning algorithmic software that's mostly used to solve classification problems. It works for both categorical and continuous dependent variables, which is surprising. We tend to divide the population into two or many homogenized groupings in this algorithmic method. This is done with the help of the most important attributes/freelancer factors to create as many separate teams as possible. In the actual world, a tree has various analogues, and it appears that it has affected a vast area of machine learning, including classification and regression. A choice tree is commonly used in call analysis to portray selections and higher cognitive processes visually and explicitly. It employs a decision-tree-like model, as the name implies. Though it's most typically employed in data mining to come up with a plan for achieving a certain objective, it's also widely utilized in machine learning. We'll use the trained model to predict whether the balancing scale will tip to the right, left, or be balanced once we've finished modelling the Decision Tree classifier.

Formulae:

P=nA/(nA+nB)P=nA/(nA+nB)

2. Random Forest

Random Forest is a fantastic method to train early in the model creation process to assess how it works, and due of its simplicity, it's difficult to design a "poor" Random Forest. This rule is also a good option if you need to construct a model in a short period of time. On top of that, it presents a reasonably accurate representation of the weight it gives to your selections. Random Forests are quite difficult to hammer down in terms of performance. On top of that, they'll deal with a plethora of different feature types, including as binary, categorical, and numerical. Random Forest is a (largely) rapid, simple, and adaptable tool, although it has limits. Random forests are an ensemble learning method for classification, regression, and other tasks that work by constructing many decision trees during training and then outputting the class that is the mode of the categories (classification) or the mean prediction (regression) of the individual trees. Random call forests correct for the characteristic of call trees overfitting to their training set.

3. Naïve Bayes Algorithm

The Naive Bayes algorithm is an algorithm that learns the likelihood of an item with given characteristics belonging to a specific group/class. For example, if you're looking for a fruit based on its color, shape, and flavor, an orange-colored, spherical, and tangy fruit is most likely an orange. All of these characteristics add to the likelihood that this fruit is an orange, which is why it is referred to be "naive." The "Bayes" section alludes to statistician and philosopher Thomas Bayes and the theorem named after him, Bayes' theorem, which serves as the foundation for the Nave Bayes Algorithm. In more formal terms, Bayes 'theorem is expressed as the following equation:

P(A/B) = (P(B/A) *P(A)) / P(B)

Equations

Confusion matrix:

It is used for evaluating performance of a classification model. It uses an N X N matrix for the purpose, where N is the number of target classes.

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.
Let's understand TP, FP, FN, TN in terms of pregnancy analogy.

True Positive:
Interpretation: You predicted positive and it's true.
You predicted that a woman is pregnant, and she is.

True Negative:
Interpretation: You predicted negative and it's true.
You predicted that a man is not pregnant, and he is not.

False Positive: (Type 1 Error)
Interpretation: You predicted positive and it's false.
You predicted that a man is pregnant, but he is not.

False Negative: (Type 2 Error)
Interpretation: You predicted negative and it's false.
You predicted that a woman is not pregnant, but she is.
Just Remember, we describe predicted values as Positive and Negative and actual values as True and False.

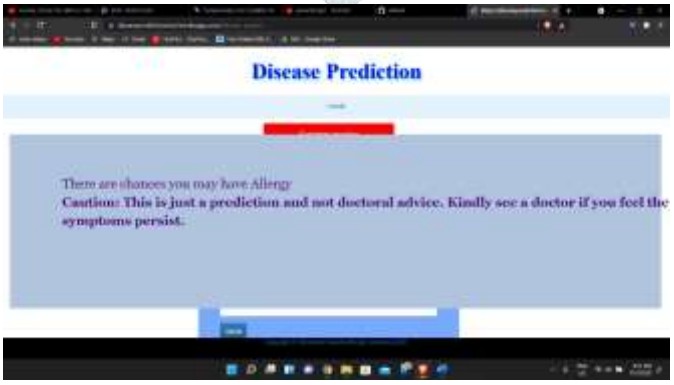
RESEARCH METHODOLOGY

The methodology section lays out the study's strategy and methods. The research's universe, sample, data and sources of data, study variables, and analytical approach are all included. Following are the specifics. We researched on models using other research papers and other projects on GitHub and other websites for selection of our algorithms and we concluded to decide 3 algorithms for our model. Decision tree classifier, Naïve Bayes algorithm and Random Forest algorithm.

RESULTS AND DISCUSSION

Check our GitHub repo link for our project.
Prediction Models

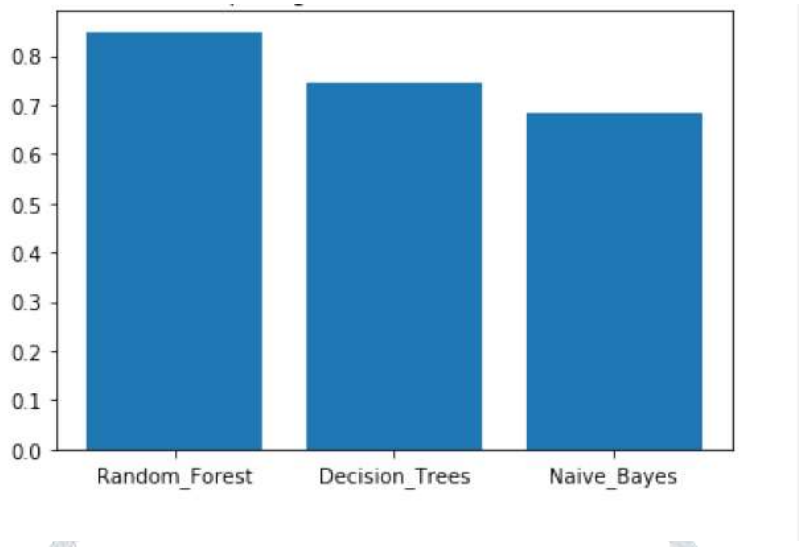
Diabetes Prediction



Disease prediction

4.1 Results of Descriptive Statics of Study Variables

Comparing the accuracy between random forest, naïve bayes and decision tree algorithm. We conclude that random forest has the highest accuracy as compared to the other 2 algorithms. But for our project all 3 models are combined to give the best accuracy output



REFERENCES

[1] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction ofheart disease using machine learning," in 2018 Second InternationalConference on Electronics, Communication and Aerospace Technology(ICECA), 2018, pp. 1275–1278.
[2] Y. Hasija, N. Garg, and S. Sourav, "Automated detection of dermatolog-ical disorders through image-processing and machine learning," in 2017International Conference on Intelligent Sustainable Systems (ICISS),2017, pp. 1047–1051.
[3] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparingdifferent supervised machine learning algorithms for disease prediction,"BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019.
[4] R. Katarya and P. Srinivas, "Predicting heart disease at early stagesusing machine learning: A survey," in 2020 International Conferenceon Electronics and Sustainable Communication Systems (ICESC), 2020,pp. 302–305.
[5] P. S. Kohli and S. Arora, "Application of machine learning in diseaseprediction," in 2018 4th International Conference on Computing Com-munication and Automation (ICCCA), 2018, pp. 1–4.
[6] M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra,"A proposed model for lifestyle disease prediction using support vectormachine," in 2018 9th International Conference on Computing, Com-munication and Networking Technologies (ICCCNT), 2018, pp. 1–6.
[7] F. Q. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in 2016 IEEE International Confer-ence on Industrial Engineering and Engineering Management (IEEM),2016, pp. 1903–1907.