International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Comparative study of Machine Learning Algorithms for Predicting Diseases Based on Symptoms

[1]Sneha Karanjai, [2]Shilpa M Varghese, [3]Hetvee Sanjay Patel, [4]P.K. Nizar Banu
[1234]Department of Computer Science,
CHRIST (Deemed to be University)
Bengaluru - 560029
India
Email: [1]sneha4ria@sgmail.com, [2]shilpavarghese10@gmail.com,
hetvee.patel2011@gmail.com, [4]nizar.banu@christuniversity.in

Abstract—The expediting power of Machine Learning in dis-ease diagnostics and appropriate use of the enormous medical data, will immensely empower and hasten precise decision mak-ing and timely medication. In the recent times, implementation of Machine Learning techniques in the field of healthcare, has become a dynamic area of research. The application of Machine Learning in the healthcare industry, carefully aligns the complex world of clinical medicine into components that even a machine can interpret, thus drastically reducing mundane human efforts. This paper aims at the study of binary classification of diseases and symptoms and comparative analysis of different classifi-cation algorithms. The algorithms illustrated for comparison includes Decision Tree, Random Forest Classifier, Multinomial Naive Bayes, Support Vector Machine, K-Nearest Neighbour and Multilayer Perceptron. Multiplayer Perceptron performed the best in comparison to the other algorithms under study with an accuracy of 86.41.

Keywords : Classification, Decision Tree, Random Forest, Sup-port Vector Machine, K Nearest Neighbour, Multi layer Perceptron, F1 Score, Recall and Precision Score, Machine Learning

I. INTRODUCTION

There is an alarming subset of the population that do not have access to proper healthcare and medical services. A lot of adults regularly trust the Internet to self-diagnose their ailments and find medications.This can be risky because despite the wealth of information available on the World Wide Web, very few are uncorroborated which makes the entire process confusing and deceiving. This results in people not seeking medical guidance for ailments that might require immediate assistance.

With the rise of technologies and rapid increase of data storage facilities, the healthcare industry has recorded massive amounts of medical data which can be used to further guide people in discomfort. Using algorithms, a correlation between a person's symptoms and the corresponding possible diseases can be found and hence ease triage and self diagnosis. Machine Learning Algorithms for pattern recognition and discovery of hidden relationships can be used for this purpose given that there is enough data and the permission to access it.For the analysis of high-dimensional and multimodal bio-medical data, machine learning offers a worthy approach for making disease predictions [1].

II. EXISTING WORK

Many researchers in the past have worked on various types of supervised and unsupervised machine learning algorithms to predict different diseases. Studies have been carried out in medical diagnosis to predict heart diseases, lungs diseases, and various tumors based on the historical data collected from patients [11]. This study is extremely domain specific and restricts the prediction of illness to one particular organ.

A great deal of research has been done that aims at diagnosing and monitoring of heart diseases at an early stage. [3]works on building a simple and accurate mobile application that uses machine learning algorithms to predict the risk of an heart disease which is a major cause of death worldwide. Machine Learning techniques have also been used to detect the risk of Diabetes in people using Classification Algorithms [4]. This is done by recognising and analysing the patterns found in the patient records for Diabetes through classification. The paper aims at finding solutions to diagnose the disease by

analyzing the patterns found in the data through by employing Decision Tree and Naive Bayes algorithms . Diseases like Liver ailments have also been predicted using SVM and Naive Bayes Algorithms in [5]. Another project determines the diagnostic and triage accuracy of online symptom checkers (tools that use computer algorithms to help patients with self diagnosis or self triage) [6]. In [7] the concept of machine learning based disease prediction over the big data is surveyed. Big data is useful in the Healthcare industries as it stores vast amount of medical records of patient data.

III. METHODOLOGY

The framework of this model involves the user to interact with the system like they would usually do with a doctor. From the text entered by the user, the symptoms are extracted and matched with those of the data set which has a list of diseases along with the corressponding symptoms. Further, machine learning methodologies are used to build a probabilistic model which displays a range of diagnoses that matches the symp-toms. This helps the patient to decide whether they should seek medical care at all and if so, then with what urgency [6].

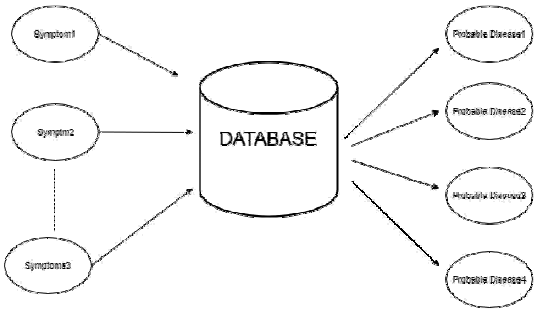


Fig. 1. MODEL FRAMEWORK

The performance metrics of different classification algorithms are compared on the test data set to evaluate the effectiveness of the algorithms. The choice of the metrics gives us a clarity on the comparison and measurement of each algorithm. The algorithms under study are :

1) Decision Tree
2) Random Forest
3) K- Nearest Neighbour
4) Support Vector Machine
5) Naive- Bayes
6) Multi layer Perceptron Classifier

The original dataset was pre-processed and converted into a categorical dataset using one hot encoding, where 1 was denoted if a symptom was present for a particular disease and 0 if the symptom is not associated with the disease.

The data set is labelled and the machine is fed with input variables (Symptoms) and output variables (Diseases) of the trained set. The machine inturn, learns the mapping function between the input and output variables. This kind of learning is called as Supervised Learning.

It is a classification problem because the symptoms are grouped together to form a particular disease. Input symptoms are matched with the group of symptoms that are already classified from the dataset and the corresponding disease is predicted. Diseases are then classified based on its associated group of symptoms. Comparison is based on F1 score, Accuracy score, Precision Score and Recall Score.

A. Decision Tree

The model uses a Decision Tree Classification Algorithm. It is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch denotes an outcome of test, and each leaf node holds a class label. The topmost node in a tree is the root node [8]. Decision tree is comparatively easier to understand and visualise compared to the other models because the functioning of the tree is visible to us. The criterion used for splitting is entropy which involves partitioning data into subsets of homogeneous instances. The root node which is a symptom is split based on the presence or absence of all the disease corresponding to that symptom. Further, that symptom is split, wherein another symptom associated or occurring along with the latter symptom, acts like the internal node. As this split is done recursively, an observation is made where the entropy or the randomness decreases at each sequential step. Thus, eventually, obtaining a leaf node or class label which is the disease predicted. It is noted that the Decision Tree Classifier takes in multiple symptoms and gives the output as the most probable disease with a probability of 1 and the others with a probability of 0.

B. Random Forests

Random forest is an ensemble classifier, which is fundamentally a collection of decision trees whose outcomes are aggregated into one final result. This kind of algorithm presents two main drawbacks : (i) the number of trees has to be fixed in prior (ii) the inter-operability and analysis capacities offered by decision tree classifiers are

lost due to the randomization principle [9]. It has a stochastic approach and creates random subsets of data with replacement. In turn, the classification instance is the modal class of all classes. Random Forest classifier which is an extension of Decision Tree, makes use of multiple Decision Trees created from randomly selected subset of symptoms from the data set to train the tree. Each of the created decision tree outputs a disease. Thus, random forest implements the concept of majority votes to each of the decision trees, and predicts the disease based on the most number of votes in each of the trees.

C. K-Nearest Neighbor Classification

Early prediction of diseases like coronary heart disease, liver ailments and diabetes detection can be done using k-Nearest Neighbor (KNN) which is a widely used lazy classification algorithm [10]. KNN is the most popular, effective and efficient algorithm used for pattern recognition. Non-parametric classification mechanism is used by k-Nearest-Neighbours (kNN) and the simple idea is based on the idea that objects" that are near each other will also have similar characteristics [11]. For a particular data point, k of its nearest neighbors are found. The algorithm is dependent on k and distance metric. Mathematically, Euclidean distance is used to calculate this distance. Parameter tuning is the process by which the value of k is determined for accurate results. A new data point is classified by initially calculating the distance of this point from all the other existing data points and then based on the value of k, it finds the "k" nearest neighbor and classifies it accordingly. The output is a class membership. This algorithm can be used to group similar symptoms together which hence predicts the corresponding disease accurately.

D. Naive Bayes

Naive Bayes is a Supervised Machine Learning Algorithm used for Classification. The algorithm is based on Bayes Theorem with a substantial (naive) assumption that every feature is independent of each other. Naive Bayes calculates the conditional probability of a disease being predicted given a symptom. It is an intuitive method that utilizes probability of each attribute belonging to each class and hence constructs a predictive probabilistic model.

Multinomial Naive Bayes(MNB) is a prominent classifica-tion approach that pleases both physicians and data scientist because it makes use of all the accessible data to justify the decision made. This explanation seems to be natural for medical diagnosis and prognosis i.e. it is close to the manner in which physicians diagnose patients [12].

In this paper, Multinomial Naive Bayes Algorithm (MNB) is applied for disease prediction. Multinomial Naive Bayes is a case of Naive Bayes Classifier that makes use of Multinomial distribution for each feature in the dataset. The algorithm calculates the prior probability of the class and the conditional probability which in turn, outputs the class or disease with the highest probability.

A Multinomial Naive Bayes, unlike Decision Tree helps in predicting multiple probable outputs.

E. Support Vector Machine

Support Vector Machine (SVM) can be applied to both regression as well as classification problems but is widely used in classification problems, hence an appropriate choice for the particular disease-symptom dataset. Support Vector Machine finds a hyperplane or a line that divides the data points in a way to form classes. Support Vector Machine aims to find an optimum decision boundary. An optimum decision boundary is the one that maximizes the distance between the nearest data points in all classes. The data points of each classes that are closest to the decision boundary are called support vectors. If a line cannot separate the data points then the help of transformations is used to visualise the data points on another dimension.Such transformations are termed as kernels.These kernels are derived by the Support Vector Machine library.

SVM has a very clever way to use large number of features without requiring nearly as much computation as seems to be necessary which makes it appropriate for high dimensional medical datasets [13]. To work with classification datasets in Support Vector Machine, one needs to pre-process the dataset using one encoding which gets the data in the form that Support Vector Machine can work on. We used the sigmoid kernel as it is binary classification. Support vector machines focuses only on the points that are the most difficult to distinguish, whereas other classifiers pay attention to all of the points.

F. Multilayer Perceptron

An Artificial Neural Network (ANN) mimics the function-ing of the human brain. Neural networks, as used in artificial intelligence (AI), have traditionally been viewed as simplified models of neural processing in the brain [14]. ANN's has a capability to predict all possible interactions between the predictor variables which in this case is the symptoms. In cases where there is no one to one

connection between the input and the output, such that one symptom does not directly result in a disease, it results in a non-linear pattern. Such a scenario would be overwhelming for a single neuron model i.e a perceptron, to handle. It needs multiple hidden layers to combine symptoms together which results in a probabilistic output of diseases and each and every layer of a Multi Layer Perceptron is fully connected with each other.

Decision Making has to be highly accurate and precise in the field of healthcare, especially when dealing with patients. Such a precision can be achieved by the use of high complexity algorithms like Multi Layer Perceptron. The use of neural networks in medicine, normally is linked to disease diagnostics systems [15]. However, neural networks are not only able to recognize examples, but maintain very important information. For this reason, one of the main areas of application of neural networks is the interpretation of medical data.

IV. RESULTS AND DISCUSSIONS

A. Data set

This paper uses knowledge database of disease-symptom associations which has been generated by an automated method based on information in textual discharge summaries of patients at New York Presbyterian Hospital admitted during the year 2004.

There are 404 unique symptoms pertaining to all the diseases in the database. The database was compiled by Friedman C, a member of biomedical informatics at Columbia University. The first attribute lists the most frequent 150 diseases, the next attribute tells us the occurrence of the disease and the final attribute describes the relevant symptoms given a particular disease. The attributes are labelled as ”Disease”, ”Count of Disease Occurrences” and ”Symptoms”.
The disease and symptoms are coded with UMLS(Unified Medical Language System) that helps in world-wide access and understandability of the data set in the biomedical field.

B. Metrics under consideration

Various classification algorithms were compared to predict the most probable diseases for a set of symptoms. Performance is typically estimated on the basis of synthetic one-dimensional indicators such as the precision, recall or F-score [17]. Confusion Matrix is the summarization of accuracies in the model. All other are metrics are derived from the confusion matrix. The diagonal of true positive plus true negatives tells us how many times diseases were correctly classified. On the other hand, False positive plus false negative tells us collectively how many diseases were misclassified. The algorithms were compared based on their Precision Score, Recall Score, F1score and Training time.
Accuracy Score was not used as metric of measurement as it is an imbalanced dataset where the occurrence of a disease is predominantly more than non occurrence of a disease. In this case the positive class of a group of symptoms resulting in a disease is outnumbered by the negative class.
Recall on the other hand can help us identify the possible scenarios in the data set. Recall is defined as the total number of true positives divided by true positives + false negatives.
For the data set, true positives are the diseases that is predicted by the model for a group of symptoms and is actually the disease that is resulted by that particular set of symptoms whereas false negatives are the diseases that the model does not predict for a group of symptoms but should be the probable output.

Precision is the ability to identify only the relevant instances. Precision is the number of true positives divided by the number of true positives plus the number of false positives. False positives are those diseases that the model has predicted for a set of symptoms but it is not actually a disease for that given set of symptoms.

F1 score is a harmonic mean of both precision and recall. F1 is more easy to understand than accuracy in the case of imbalanced data. The closer the F1 score is to 1 the better the accuracy of the model.

If the problem statement reads that a disease should be present and it is present (True Positive) then we consider precision as the measure. But on the other hand, if the problem statement reads that a True Positive is given by a person not suffering a disease, we use recall.

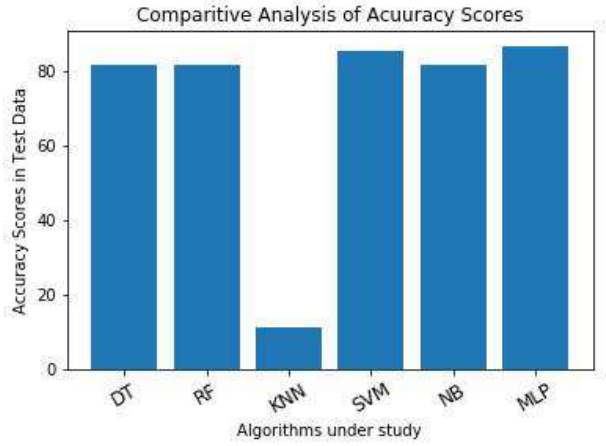
C. Comparative Analysis

The paper illustrates comparison of six Machine Learning Algorithms which are used for disease prediction.

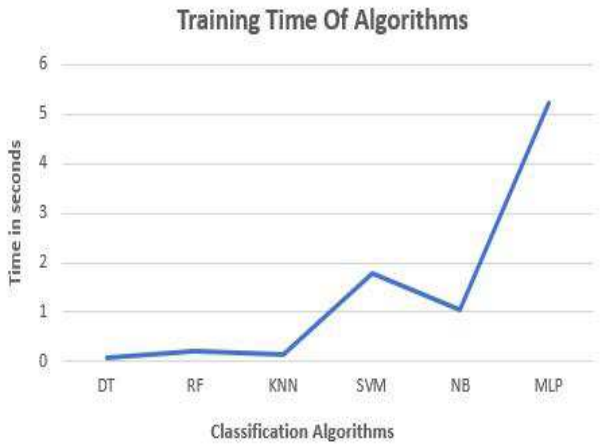
Metrics	DT	RF	NB	KNN	SVM	MLP
Precision	0.7962	0.8024	0.7962	0.0987	0.8395	0.8518
Recall	0.8148	0.8148	0.8148	0.1111	0.8518	0.8641
F1 Score	0.8024	0.8065	0.8024	0.1028	0.8436	0.8559

TABLE
I
COMPARISON OF DIFFERENT CLASSIFICATION METRICS

On differentiation between F1 score of Decision Tree 0.8024 and f1 score of Random Forest Classifier, it is noted that the metrics of classification for Random Forest is higher than that of Decision Tree. Also, an observation is made that since, Random Forest decreases the variance part of error in comparison to the bias part of error, the accuracy of training data for DT (85.0) is better than the accuracy of training dataset for RT (84.78). But the accuracy on the test dataset of Decision Tree (81.4814) is lower than the accuracy on the test dataset of Random Forest (82.7160). On comparison between Random Forest Classifier (RF) and Mutinomial Nave Bayes (MNB), the model showcases a higher classification metrics for Random Forest. Support Vector Machine (SVM) algorithm has a higher classification metrics than Random Forest (RF). Amongst all the algorithms applied on the dataset, it is found that MultiLayer Perceptron Classifier (MLP) provides the highest classification metrics. This can be justified, as an MLP classifier takes in multiple symptoms as the input layer and then has an arbitrary number of hidden layers which contains numerous coherent symptoms. The hidden layer known for being the computational complex engine of MLP then provides an output layer of probabilistic diseases. k-Nearest Neighbour (kNN) is found to have the lowest classification metrics amongst all the algorithms used for prediction of disease. kNN fails to work in an optimized way for categorical features as it is unable to find the distance between categorical data points. Since, our data set comprises of only categorical data points where a symptom exists (1) or does not exist (0) for a particular disease, kNN fails to show optimum results.



The computational time taken by the algorithm to train the data is termed as the Training Time.
Training time is the maximum in Multi layer perceptron because the complexity of this Deep Learning algorithm is maximum when compared to others. It trains the data according to the set number of epochs and hence takes the maximum time. For real time, larger data set, Multi Layer Perceptron might be time consuming. The least training time is for Decision Tree.



Training Time Of Algorithms

V. CONCLUSION AND FUTURE WORK

For the training purpose, the dataset was split into 80:20 train and test sets. The model is fit into training set and hence we check the accuracy of the test set and the predicted variables to determine the count of misclassified diseases. Multi Layer Perceptron gives the highest accuracy for a classification data set of purely categorical features followed by Support Vector Machine. K Nearest Neighbour falters in the comparison because of its limitations on categorical data. The most important parameter is not the way of functioning of the algorithms, but the result information, accuracy and operation speed.

With exponential growth in machine learning and artificial intelligence, the initial stage of disease diagnosis can be made simpler and quicker benefiting patient care and early diagnosis. However, the accuracy of the analysis of prediction primarily depends on the availability of

medically relevant data [16]. With a more real time approach towards this model and a wide variety of diseases covered in the data along with the symptoms, the models will show a more appropriate result. Ideally the thumb rule for any Machine Learning algorithms for accurate training suggests that the the number of rows should be 10 times more than the number of columns, but in this scenario the number of symptoms will always increase with the increase in the number of diseases. The data set is also purely categorical because of which algorithms failed to show optimum desirable results.

REFERENCES

[1] Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algo-rithms for Disease Diagnostic, Journal of Intelligent Learning Systems and Applications
[2] Shee, Hassan W Cheruiyot, Kipruto Kimani, Stephen. (2014). Appli-cation of k-Nearest Neighbour Classification in Medical Data Mining.
[3] A. F.Otoom, Emad E. Abdallah, Y. Kilani, A. Kefaye and M. Ashour, Effective Diagnosis and Monitoring of Heart Disease International Journal of Software Engineering and Its Applications Vol. 9, No. 1 (2015),
[4] Iyer, Aiswarya Jeyalatha, S Sumbaly, Ronak. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining Knowledge Management Process. 5. 1-14. 10.5121/ijdkp.2015.5101.
[5] S. Vijayarani, S.Dhayanand Liver Disease Prediction using SVM and Nave Bayes Algorithms International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015
[6] Hannah L Semigran,Jeffrey A Linder,C. Gidengil,A. Mehrotra, Evalu-ation of symptom checkers for self diagnosis and triage: audit study, BMJ 2015
[7] K.DeepthiKrishnan, B.Senthil Kumar A Survey on Disease Prediction by Machine Learning over Big Data from Healthcare Communities [IOSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719]
[8] B. N. Patel, Satish G. Prajapati and K. I. Lakhtaria Efficient Classifi-cation of Data Using Decision Tree, Bonfring International Journal of Data Mining, Vol. 2, No. 1, March 2012
[9] S. Bernard, L. Heutte and S. Adam, ”On the selection of de-cision trees in Random Forests,” 2009 International Joint Confer-ence on Neural Networks, Atlanta, GA, 2009, pp. 302-307. doi: 10.1109/IJCNN.2009.5178693
[10] Jabbar MA, Prediction of heart disease using k-nearest neighbor and particle swarm optimization, Biomedical Research 2017; 28 (9): 4154-4158
[11] Shee, Hassan and W Cheruiyot, Kipruto Kimani, Stephen. (2014). Application of k-Nearest Neighbour Classification in Medical Data Mining. 4.
[12] K.M. Al-Aidaroos, A.A. Bakar and Z. Othman, 2012. Medical Data Classification with Naive Bayes Approach. Information Technology Journal, 11: 1166-1174.
[13] P Janardhanan, Heena L., and F Sabika, Effectiveness of Support Vector Machines in Medical Data mining JOURNAL OF COMMUNICA-TIONS SOFTWARE AND SYSTEMS, VOL. 11, NO. 1, MARCH 2015
[14] O.G Nayeem, M N Wan,K. Hasan, Prediction of Disease Level Using Multilayer Perceptron of Artificial Neural Network for Patient Monitor-ing, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-5 Issue-4, September 2015
[15] E. Xhumari, P. Manika, Application of artificial neural networks in medicine, University Of Tirana,RTA-CSIT, 2016
[16] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, ”Disease Prediction by Machine Learning Over Big Data From Healthcare Communi-ties,” in IEEE Access, vol. 5, pp. 8869-8879, 2017. doi: 10.1109/AC-CESS.2017.2694446
[17] C. Goutte and E. Gaussier, A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation. European Confer-ence on Information Retrieval, 2005