International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 12 | November - 2020 ISSN: 2582-3930

Computer-Aided Diagnosis for Diseases using Machine Learning

Keerthi Vuppula, Software Developer, Scrutiny software solutions, Hyderabad, India
Dr. Narsimha Reddy, Professor, Department of CSE, CMRIT, India

Abstract

The world is moving with a quick speed and to keep up ourselves with the entire world we overlook the symptoms of disease which can influence our wellbeing at a huge degree or may result to death. Great Wellbeing assumes perhaps of the main part in human existence. Thus, we have planned a disease expectation framework utilizing numerous ML algorithms. That dataset involved that had in excess of 230 diseases for handling. In view of the symptoms, age, and orientation of an individual, the determination framework gives the result and foresee what he/she might have. We utilized different algorithms like decision tree, random forest, knn, naive byes with a precision of 94%.

Keywords: Symptoms, Prediction, Machine Learning, Diseases

I. Introduction

It has been demonstrated that consistently more than 65% of the populace in India has a propensity towards general disease which incorporates cough, cold, fever, Headaches. In any case, heaps of individuals don't understand that these symptoms can be more hazardous for them in future, even absence of obliviousness might result into death. Subsequently the ID of these disease at beginning phase is extremely important. In this way, the motivation behind our venture is to make expectations for the usually happening symptoms that were unrestrained and later become the lethal disease. We had applied algorithms like decision tree algorithm, random forest algorithm, naive byes algorithm [1]. These algorithms will foresee the disease in view of the symptoms' that client enter. At present the accuracy of the task is 94%.After running numerous algorithmic rule the gathered outcome and while breaking down the outcome, I used to be found out that the two algorithms, for example KNN and supply regression was having the best accuracy result contrasted and elective leftover algorithms and out of those 2 algorithms, supply regression was the best and very responsive in term of genuine positive rate or review. From the disarray matrix KNN accuracy comes undaunted be 73.97% and subsequently the SVM accuracy comes fearless be 71.97%. Thus, the best model for disease expectation is supply regression since it offers the best accuracy among every single elective algorithm. ANN is being utilized to characterize the named pictures in light of the assurance of genuine positive and misleading positive identification rates [2]. The ANN is portioned into two methodologies, at first they applied the classifier to the picture information with district of interest (return for money invested) and second incorporates the ANN gain the highlights from pre-handled picture signals [3]. The preparation information is 70% unique information and testing information is 30%, for productive information examination.

II. System Design

The arranged framework centers around anticipating the diseases utilizing machine learning algorithmic program by essentially entering 5 symptoms by breaking down the symptoms the frameworks foresee the pertinent result [4]. First the client can enter the name of patient so pick symptoms with regards to the patient's condition so click subsequently on marks name expectation one, forecast two, forecast three and forecast four. This data will be handled utilizing machine learning algorithms so the framework will be prepared to anticipate the disease upheld the information given by the client.

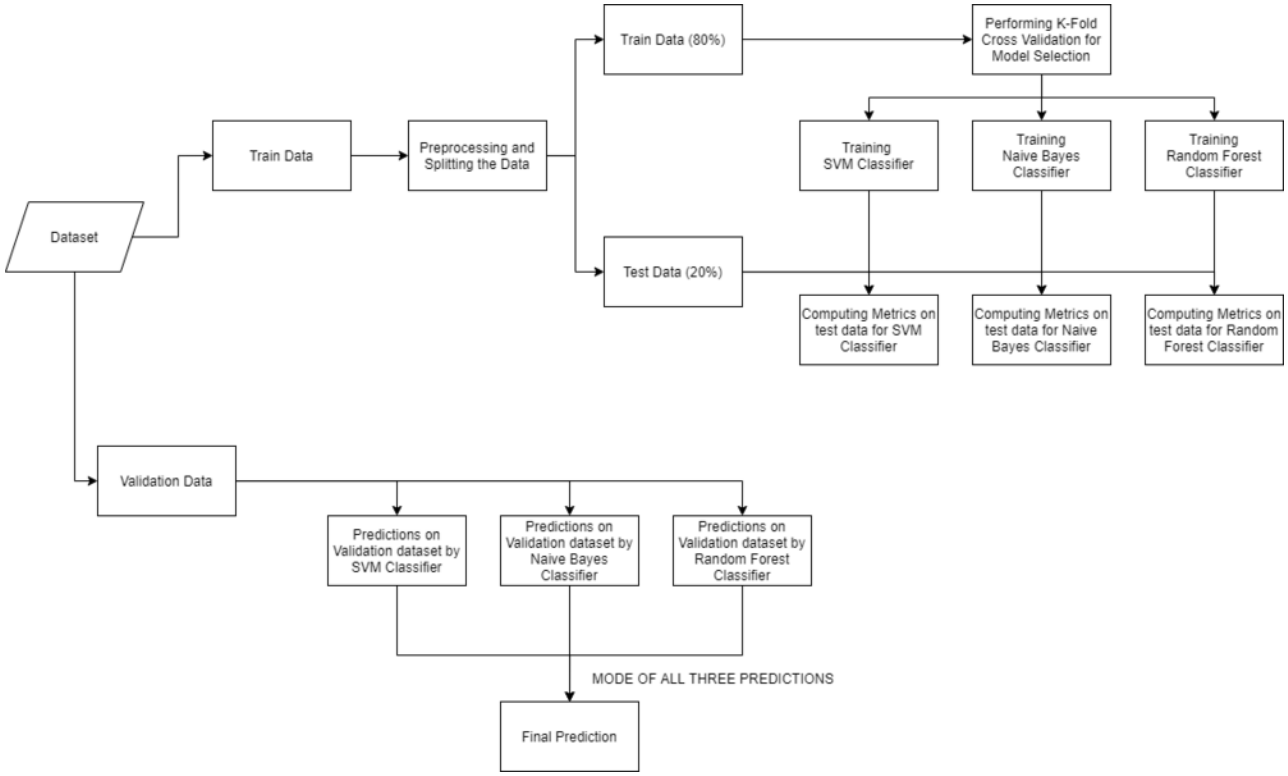

Figure 1: Flow chart of how disease prediction works using different Machine learning algorithm

1. Random forest

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, a few trees make a forest. Every individual tree in random forest lets out a class assumption and the class with most votes transforms into a model's estimate [5]. In the random forest classifier, the more the quantity of trees higher is the accuracy. It is utilized for classification too as regression task, yet can do well with classification task, and can beat missing qualities. Also, being delayed to get predictions as it requires enormous informational indexes and more trees, results are un accountable [6].

2. Decision tree

Decision tree is a classification algorithm that deals with unmitigated as well as mathematical information. Decision tree is utilized for making tree-like designs it is not difficult to execute and analyze the information in tree-shaped graph. This algorithm divides the information into at least two analogous sets in light of the main markers [7]. The entropy of each characteristic is determined and afterward the information are separated, with predictors having most extreme data gain or least entropy The outcomes acquired are simpler to peruse and decipher. This algorithm has higher accuracy in contrast

with different algorithms as it analyzes the dataset in the tree-like graph. Nonetheless, the information might be finished characterized and just a single property is tried at a time for decision-making [8].

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$IG(Y,X) = E(Y) - E(Y|X)$$

3. Naïve Bayes

Naïve Bayes classifier is a supervised algorithm. It is a basic classification technique utilizing Bayes hypothesis. It expects freedom among credits. Bayes hypothesis is a numerical idea that is utilized to get the likelihood. The predictors are neither connected with one another nor have relationship to each other. Every one of the properties autonomously add to the likelihood to expand it. Numerous mind boggling certifiable circumstances utilize Naive Bayes classifiers [9].

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$IG(Y,X) = E(Y) - E(Y|X)$$

P(X/Y) is the posterior probability, P(X) is the class prior probability, P(Y) is the predictor prior probability, P(Y/X) is the likelihood, probability of predictor [10].

4. K-Nearest Neighbor

The K-Nearest Neighbor algorithm is a supervised classification algorithm strategy. It arranges objects dependant on nearest neighbor. It is a kind of occurrence based learning. The estimation of distance of a characteristic from its neighbors is estimated utilizing Euclidean distance [11]. It utilizes a gathering of named places and uses them on the most proficient method to mark another point. The information are grouped in light of comparability among them. K-NN algorithm is easy to complete without making a model or making different presumptions. This algorithm is flexible and is utilized for classification, regression, and search. Despite the fact that K-NN is the least difficult algorithm, loud and superfluous highlights influence its accuracy [12].

III. Experimental Results

Dataset used in this strategy is during an organized organization. The dataset that is utilized contains the sickness name with its all symptoms. As our framework depends on supervised learning machine algorithms, the dataset has the name with nothing or one. Then we will more often than not partition the dataset into a preparation dataset and testing dataset. The model is prepared by a preparation dataset contains 4920 rows. Classification algorithms and in doing so see as the most precise algorithm for anticipating regardless of whether a patient would create and coronary illness[13]. This examination was finished utilizing techniques of Calculated Regression, Naïve Bayes, Backing Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, and XGBoost on the UCI dataset. Dataset was parted into preparing and test information and models were prepared and the accuracy was noted utilizing Python. The general disease expectation framework performs supervised learning algorithm which is the core of the framework [14].

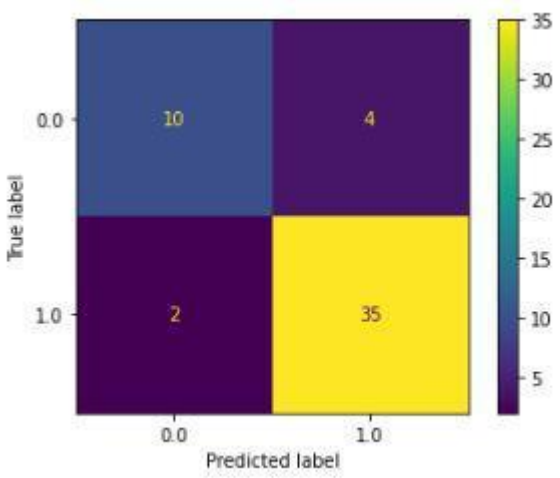

Figure 2: confusion matrix

The client needs to enter the name first then the five symptoms. Disease will be tossed as result in light of those symptoms; Rundown of symptoms is separated into levels in view of the nature of disease. Model is rotated around four unique algorithms Decision tree, Random Forest, KNN and Naïve bayes.

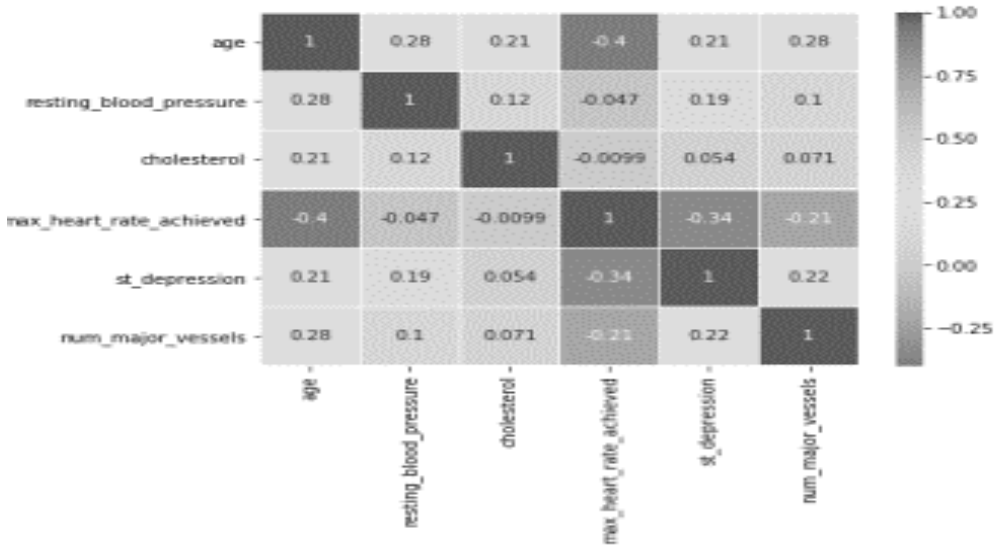

Figure 3: Correlation Matrix

Algorithm will be called when client hit the anticipate button, anticipate 1, 2, 3 and 4 button is for Decision tree, Random Forest, KNN, Naive byes individually. There to be somewhere around two symptoms expected to get the expectation, Disarray matrix and accuracy score predicts the result of our model for every algorithm and separate table is made which stores the info and result information. Correctness's given by decision tree, Random Forest, KNN and Naive bayes are 95.6%,93.66%,92% and 95.1%.

IV. Conclusion

The general point is to characterize different data mining techniques valuable in compelling coronary illness expectation. Proficient and precise expectation with a lesser number of characteristics and tests is the objective of this examination. The data were pre-handled and afterward utilized in the model. Random Forest with 86.89% and XGBoost with 78.69% are the most proficient algorithms. Nonetheless, K-Nearest Neighbor performed with the most obviously terrible accuracy with 57.83%.the more serious diseases which could be treatable whenever recognized at beginning phase and have high casualty rate, such a different kind of malignant growths so early recognition would assist the patient with counseling the specialist and get clinical consideration a whole lot earlier and a sensor can be added which in future assist the patient's family with checking the patient health.

References:

[1]. Das, M., Manmatha, R., Riseman, E.: 'Indexing flower patent images usingdomain knowledge', IEEE Intell. Syst. Appl., 1999, 14, (5), pp. 24–33[3]
[2]. Larson, R. (Ed.): 'Introduction to floriculture' (Academic Press, San Diego,CA, USA, 1992, 2nd edn.)
[3]. Chi, Z.: 'Data management for live plant identification', in Feng, D.,, Siu,W.C.,, Zhang, H.J. (ED.): 'Mutimedia information retrieval and Management'(Springer, Berlin Heidelberg, 2003), pp. 432–457
[4]. Nilsback, M., Zisserman, A.: 'Automated flower classification over a largenumber of classes'. Proc. Sixth Indian Conf. Computer Vision, Graphics &Image Processing, Bhubaneswar, India, December 2008, pp. 722–729
[5]. Nilsback, M., Zisserman, A.: 'A visual vocabulary for flower classification'.Proc. IEEE Conf. Computer Vision and Pattern Recognition, New York, NY,June 2006, 2, pp. 1447–1454[
[6]. Zou, J., Nagy, G.: 'Evaluation of model-based interactive flower recognition'.Proc. Int. Conf. Pattern Recognition, Cambridge, UK, August 2004, 2, pp.311–314
[7]. Yang, M., Zhang, L., Feng, X., et al.: 'Sparse representation based Fisherdiscrimination dictionary learning for image classification', Int. J. Comput.Vis., 2014, 109, (3), pp. 209–232
[8]. Khan, F., van de Weijer, J., Vanrell, M.: 'Modulating shape features by colorattention for object recognition', Int. J. Comput. Vis., 2012, 98, (1), pp. 49–64
[9]. Xie, L., Wang, J., Lin, W., et al.: 'Towards reversal-invariant imagerepresentation', Int. J. Comput. Vis., 2017, 123, (2), pp. 226–250[
[10]. Hsu, T., Lee, C., Chen, L.: 'An interactive flower image recognition system',Multimedia Tools Appl., 2011, 53, (1), pp. 53–73
[11]. Mottos, A., Feris, R.: 'Fusing well-crafted feature descriptors for efficientfine-grained classification'. Proc. IEEE Int. Conf. Image Processing, Paris,France, October 2014, pp. 5197–5201
[12]. L. Yang, G. Yang, K. Wang, H. Liu, X. Xi and Y. Yin, "Point Grouping Method for Finger Vein Recognition," in *IEEE Access*, vol. 7, pp. 28185-28195, 2019.
[13]. Vijay Reddy, Madireddy (2020), "A Review on architecture and security issues Cloud Computing Services", Journal For Innovative Development in Pharmaceutical and Technical Science (JIDPTS) Oct-2020,pp 1-4
[14]. S. Ramana, S. C. Ramu, N. Bhaskar, M. V. R. Murthy and C. R. K. Reddy, "A Three-Level Gateway protocol for secure M-Commerce Transactions using Encrypted OTP," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 1408-1416, doi: 10.1109/ICAAIC53929.2022.9792908.
[15]. N.Bhaskar, S.Ramana, & M.V.Ramana Murthy. (2017). Security Tool for Mining Sensor Networks. International Journal of Advanced Research in Science and Engineering, BVC NS CS 2017, 06(01), 16–19. ISSN Number: 2319- 8346
[16]. Karunakar Pothuganti, (2018) 'A comparative study on position based routing over topology based routing concerning the position of vehicles in VANET', AIRO International Research Journal Volume XV, ISSN: 2320-3714 April, 2018 UGC Approval Number 63012.

[17]. K. Pothuganti, B. Sridevi and P. Seshabattar, "IoT and Deep Learning based Smart Greenhouse Disease Prediction," 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2021, pp. 793-799, doi: 10.1109/RTEICT52294.2021.9573794.

[18]. I. Ahmad and K. Pothuganti, "Smart Field Monitoring using ToxTrac: A Cyber-Physical System Approach in Agriculture," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 723-727, doi: 10.1109/ICOSEC49089.2020.9215282.

[19]. Poornachander Vadicherla, Dhanalakshmi Vadlakonda,"Study on energy efficient routing protocols scheme in heterogeneous wireless sensor networks (network & mobility)", Materials Today: Proceedings,Volume 47, Part 15, 2021, Pages 4955-4958, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.04.173.

[20]. V. Poornachander and V. Dhanalaxmi, "Scalable, Opportunistic, Energy Efficient Routing (SOEER) - A Novel Clustering Approach for Wireless Sensor Networks," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 1256-1264, doi: 10.1109/ICAAIC53929.2022.9792656.