IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

Disease Prediction System Based on SVM, Random Forest and Naive Bayes

Aditya Patil[1], Shreya Wadkar[2], Amol Nidankar[3] and Prof. Aparna M. Bagde[4]
Students, Department of Computer Engineering[1,2,3]
Faculty, Department of Computer Engineering[4]
NBN Sinhgad School of Engineering, Pune, Maharashtra, India
adityapatil.nbnssoe.comp@gmail.com

Abstract: *The development and exploitation of several leading data-mining techniques in many real-world application areas (e.g., industrial, healthcare, and life sciences) has led to the use of such techniques in machine learning environments to extract important pieces of information from the specified data in health communities, biomedical fields, etc. Accurate analysis of medical database benefits in early disease detection, patient care and community services. Machine learning techniques have been used successfully in various applications, including prediction of early-stage disease prediction and diagnosis. This research work demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier.*

Keywords: Machine Learning, Data mining, Decision Tree Classifier, Random Forest Classifier, Naive Bayes classifier.

I. INTRODUCTION

Today healthcare industry has become a huge money making business. The healthcare industry utilizes and produces quite a large amount of data which can be used to extract information about a disease for a patient. This information of healthcare will further be used for effective and best possible treatments for patient's health. This area also needs some improvement by using the informative data in healthcare .But a crucial challenge is to extract the information from the data because the data is present in a large amount so some data mining and machine learning techniques are used. The expected result of this project is to predict the disease in advance, so that the risk of life can be prevented early and lives can be saved and the costs of treatment can be significantly reduced. In India, it should also adopt the non-manual system of medical treatment, which is most suitable for the improvement and understanding of human health. The main reason is to use the concept of machine learning in healthcare to improvise the treatment of patients. Machine learning has already made it much easier to identify and predict different diseases. Disease prediction analytics using many machine learning algorithms help us to predict diseases and treat patients effectively. Machine learning disease prediction also uses medical histories and health data by applying various concepts such as data mining and machine learning techniques and some algorithms. Diseases and health problems like malaria, dengue fever, impetigo, diabetes, migraines, jaundice, chickenpox, etc. have a significant impact on health and sometimes even lead to death if ignored. The healthcare industry can make an effective decision by "evaluating" the massive database they possess i.e. extracted the hidden patterns and relationships in the database. Data mining algorithms such as decision tree, random forest, and naive Bayes algorithms can help here. Therefore, they developed an automated system that can discover and extract hidden knowledge associated with diseases from a historical database (disease symptoms) according to the rule set of the respective algorithms

II. LITERATURE SURVEY

"Disease Prediction the usage of Machine Learning Algorithms",[1] have been correctly employed in numerous packages which include Disease prediction. The goal of developing classification tools for the usage of device mastering algorithms is to immensely help to solve the health-related issues through assisting the physicians to be expecting and diagnose illnesses at an early stage. Sample records of 4920 patients' statistics recognized with 41 diseases were determined for analysis. An established variable was composed of 41 illnesses. ninety five of 132 independent variables(symptoms) closely related to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

ailments were decided on and optimized. This research paintings executed demonstrates the disease prediction tool superior to the usage of Machine mastering algorithms Such as Decision Tree classifier, Naïve Bayes classifier, Random forest classifier. The paper offers the comparative have a look at of the consequences of the above algorithms used. The system getting to know techniques "Disease Prediction the use of Machine Learning Algorithms"[2], have been efficiently employed in diverse packages inclusive of Disease prediction. The intention of growing classifier device the use of system getting to know algorithms is to immensely assist to resolve the health-associated troubles through helping the physicians to predict and diagnose illnesses at an early stage. A Sample information of 4920 patients' data identified with forty one diseases have been decided on for analysis. An established variable turned into forty one illnesses. Ninety five of 132 impartial variables(symptoms) intently associated with illnesses have been selected and optimized. This studies paintings demonstrated the disorder prediction device advanced the use of Machine getting to know algorithms Such as Random forest classifier, Decision Tree classifier, and also Naïve Bayes classifier. The paper provides the comparative observe of the effects of the above algorithms used.

III. METHODOLOGY

3.1 Naive Bayes

A type method primarily based totally on Bayes' Theorem with an assumption of independence amongst predictors. In easy terms, a Naive Bayes classifier assumes that the presence of a selected function in a category is unrelated to the presence of every other function.

3.2 Bayes' Theorem

Bayes theorem gives a manner to calculate the opportunity of a speculation given our earlier knowledge. In different phrases it unearths the opportunity of an occasion happening given the opportunity of every other occasion that has already occurred. Using Bayes theorem, we are able to discover the opportunity of a happening, for the reason that B has occurred. Here, B is the proof and A is the speculation. The assumption made right here is that the predictors/functions are independent. That is, the presence of 1 specific function does now no longer have an effect on the difference. Hence it's far known as naive.

Bayes' theorem is said mathematically as the subsequent equation:

$$P(A|B) = P(A \cap B)/P(B)$$

Where in A and B are activities and $P(B) \neq 0$.Basically, we're looking for the opportunity of occasion A, given the occasion B is true. Event B is likewise termed as proof. P(A) is the priori of A (the earlier opportunity, i.e. Probability of occasion earlier than proof is seen). The proof is an characteristic cost of an unknown instance (right here, it's far occasion B).P(A|B) is a posteriori opportunity of B, i.e. opportunity of occasion after proof is seen.

3.3 Random Forest

Random Forest is a Supervised Machine Learning Algorithm that is used extensively in Classification and Regression problems. It builds choice bushes on certainly considered one among a type sample and takes their majority vote for class and is not an unusual place in case of regression.

Steps in Random Forest Algorithm:

In Random Forest Various numbers of random statistics are taken having n quantity of statistics. Individual choice bushes are constructed for each sample. Each Decision tree generates Output. Final Output is the Averaging for type and regression respectively.

A. Advantages of Random Forest

1. **Diversity:** Not all attributes/variables/capabilities are considered at the same time as making a character tree, each tree is specific.
2. **Immune to the curse of dimensionality:** Since each tree now not continues in thought all the capabilities, the feature space is reduced.
3. **Parallelization:** Each tree is created independently out of various information and attributes. This way we can make entire use of the CPU to assemble random forests.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

4. **Train-Test break up:** In a random wooded area we shouldn't segregate the information for teach and check as there will continuously be 30% of the statistics which isn't seen via the choice tree.
5. **Stability:** Stability arises because of the reality the give up end result is based mostly on majority vote casting/ averaging.

B. Hyperparameters in Random Forest

Hyperparameter are used to boost overall performance and for Predictive energy of fashions or to run fashions faster.

1. **n_estimators:** Quantity of bushes the set of rules builds earlier than averaging the predictions.
2. **max_features**: Most quantity of capabilities in a random variable considers splitting a node.
3. **mini_sample_leaf**: Determines the minimal quantity of leaves required to break up an inner node.

Support vector machines (SVMs) are effective but bendy supervised system mastering algorithms that are used each for type and regression. But generally, they may be utilized in type problems. An SVM version is largely an illustration of various lessons in a hyperplane in a multidimensional area. The hyperplane could be generated in an iterative way through SVM in order that the mistake may be minimized. The purpose of SVM is to divide the datasets into lessons to discover a most marginal hyperplane (MMH).

3.4 SVM Factors

- **Hyperplane:** As we will see withinside the above diagram, it's far a choice aircraft or area that is divided among a hard and fast of gadgets having extraordinary lessons.
- **Support Vectors:** Data Points which are closest to the hyperplane are known as aid vectors. Separating line could be described with the assistance of those records factors.
- **Margin:** It can be described as the space among strains at the closet records factors of various lessons. It may be calculated because of the perpendicular distance from the road to the aid vectors. Large margin is taken into consideration as a terrific margin and small margin is taken into consideration as an awful margin.

A. Advantages of SVM Classifiers

SVM classifiers give amazing accuracy and paintings nicely with excessive dimensional area. SVM classifiers essentially use a subset of schooling factors subsequently and the end result makes use of very much less memory.

B. Disadvantages of SVM Classifiers

They have excessive schooling time subsequently in exercise now no longer appropriate for big datasets. Another downside is that SVM classifiers no longer paint nicely with overlapping lessons.

IV. CONCLUSION

From the historical development of machine learning and of its applications in the medical sector, it can be shown that systems and methodologies have emerged that have enabled sophisticated data analyzes through the simple and direct use of algorithms of machine learning. This paper presents a comprehensive comparative study of the performance of three algorithms on a medical record, each providing up to 95% accuracy. The performance is analyzed through the confusion matrix and the precision score. Artificial intelligence will play an even greater role in analyzing data in the future due to the availability of huge data produced and stored by modern technology.

REFERENCES

[1]. Disease Prediction using Machine Learning Algorithms, Sneha Grampurohit, Chetan Sagarnal, Published in:2020 International Conference for Emerging Technology (INCET))

[2]. Machine learning equipped web based disease prediction and recommender system, Harish Rajora, Narinder Singh Pun, Sanjay Kumar Sonbhadra, and Sonali Agarwal,arXiv:2106.02813v2 [cs.CV] 4 Jul 2021March -2020

[3]. Disease Prediction System,Aditya Tomar, International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

[4]. Disease Prediction System using data mining techniques, Aditya Tomar, International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016

[5]. Qulan, J.R. 1986. "Induction of Decision Trees". Mach. Learn. 1,1 (Mar. 1986),81-10

[6]. Sayantan Saha, Argha Roy Chowdhuri et,al "Web Based Disease Detection System",IJERT, ISSN:22780181,Vol.2 Issue 4, April-2013

[7]. Shadab Adam et.al "Prediction system for Heart Disease using Naïve Bayes", International Journal of advanced Computer and Mathematical Sciences, ISSN 2230- 9624, Vol 3,Issue 3,2012,pp 290-294[Accepted- 12/06/2012].

[8]. Min Chen, Yixue Hao et.al "Disease Prediction by Machine Learning over big data from Healthcare Communities", IEEE[Access 2017]

[9]. Mr Chintan Shah, Dr. Anjali Jivani, "Comparison Of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE-31661

[10]. Palli Suryachandra, Prof. Venkata Subba Reddy, "Comparison of Machine Learning algorithms For Breast Cancer", IEEE