iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50642

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Human Symptoms Based on Diseases Predictor

Mrs. M. Premalatha[1], G M. Delipan[2], V. Kavyashri[3], S. Sanjay[4], K. Srijayakanth[5]
[1]*Assistant Professor,* [2,3,4,5]*UG Scholar, Department of Computer Science and Engineering, Nandha Engineering College (Autonomous), Erode, Tamilnadu, india*

Abstract: ***Many situations occur in day to day life which affects a human being. Many problems are happening in fast manner and new diseases are rapidly being created. The main objective of this project is to apply classification algorithm to predict model for occurrence of various diseases. This project work is aimed in identifying the best classification algorithm to identify the disease probability of patients. The identification of the possibility of diseases in patients is a tedious task for doctors and researchers because it requires experience and more medical tests need to be taken. The main objective of this project is to find the best classification algorithm suitable to provide accuracy improvement during classification of normal and abnormal persons. The project contains Naïve Bayes, Support vector machine and decision tree classification with their accuracy score calculation. The applied NBS, SVM, DT classification help to predict the disease with higher accuracy in the new data set. Python 3.9 is used as the coding language.***
Keywords: Machine Learning, Decision Tree, Support Vector Machine, Naïve Bayes Classification.

I. INTRODUCTION

Big data has its significance in each and every field in world including health care industries. It alters the way to handle doctors and patients doctors with care. From huge number of sample data, we could expect more accurate results, insights for health care industries. Like many other industries, health care industry is also a framework which contains heterogeneous multi sectors which are complex to deal with high accuracy, in which the patients demand better care with reduced cost.

Day by day, emerging technologies are being included to healthcare industry, where big data analytics have a vital role to give effective business insights to hospitals and patients. In technical world, data analysis also plays an important role in every field in the world where the data volume is so limited. But today, the world is now in big data era. Existing statistics announces that the data analytics is very important in future for health care industries and it will become very crucial in operational, clinical and banking/financial sectors. The collected data is potentially be used by Government and public organizations create or enhance procedures, policies and trainings. Over all, project will have the potential to heighten awareness for the requirements to give the best treatments in any healthcare environment. Most of the patients are uneducated and those are not familiar with precise treatments. Majority of patients/people approach private health care centers which are not able to save details of patients and their diseases. So there is a requirement for organizing health camps that educate and sensitize the communities. This study explains about diagnosis and numerous types of health hazards.

II. LITERATURE SURVEY

Nowadays, more amounts of structured, unstructured, and semi-structured data are generated by numerous institutions around the globe and, these heterogeneous data are referred to as big data. Health industry sectors have been confronted by urge to maintain the big data being produced by various sources, which are known for producing huge volumes of heterogeneous data.

More number of big-data analytics tools and techniques are developed to handle massive amounts of data for healthcare sectors. In this paper, authors discussed impact of big data in health care, and different tools available in the Hadoop ecosystem to handle it.

They also explored conceptual architecture of bigdata analytics in health care that includes data gathering history of various branches, genome database, and EHR (electronic health records), text/imagery as well as clinical DSS (decisions support system).

Every day, data is generated by a range of numerous applications, and geographical research activities for purposes of disaster evaluation, , prediction of weather, weather forecasting, crime detection, and also the heath industry, to name a few. In today scenarios, big data is associated with core technologies and numerous enterprises that include Google, Facebook, and IBM, that extract valuable information from huge volumes of data collected [3–5]. An era of open information for healthcare is now on the road. Big data is generated rapidly in all fields which include healthcare, with respect to patient compliances, care and various regulatory requirement.

As global population continues to grow with human lifespan, treatment delivery models evolve rapidly, and some decisions underly these fast changes are based on the data [6]. Healthcare shareholders promised the novel knowledge from big data, so called both for volume as well as its complexity with range.

Pharmaceutical-industry experts and shareholders started to routinely scrutiny/analyze big data for obtaining insight, but the activities are still in the beginning stages and must be coordinated in order to address health care delivery problems and enhance the health care quality. Early systems of big-data analytics for the health care informatics are established across various scenarios, for example, investigation of patient characteristics/determination of treatment cost and results to pinpoint the best and most cost effective treatments [6].

Health informatics is termed as assimilation of the health care sciences, computing and information sciences in study of healthcare information. Health informatics includes data storage, acquisition and retrieval for providing better results by health care providers.

They concluded that they have provided an in-depth description and a brief overview of big data in general and in healthcare systems, that plays a significant role in health care informatics and influences greatly health care system and big data Vs in the health care. They also proposed uses of conceptual architecture for solving health care problems in big data using the Hadoop terminologies, which involves utilization of big data, generated by diffrent levels of medical data and development of methods for analyzing this data and for obtaining answers in medical questions. The combination of the big data and health care analytics leads to treatments which are effective to specific patients to provide ability to prescribe proper medications for each individuals, than those that work with most other people.

In the paper [2] the authors presented a brief introduction to big data and its role in health care applications. It is seen that the use of big data architecture and the techniques are continuously assisting in managing data growth in the health care industries. Here, initially an empirical study was conducted for analyzing the big data roles in health care industries.

It is observed that significant works have been done for the big data in health care sectors. Today, it is intricate for envisioning the way machine learning and big data influences the health care industries. It is observed that most authors implemented machine learning and big data analytics use in disease diagnosis are not providing significant weightage to data privacy and security.

Here, a novel design of smart/secure health care information system with use of machine learning and also advanced security mechanism are proposed to handle big data for medical industries. The innovation lies in incorporation of optimal storage and data security layer used to maintain both security and privacy. Various techniques like activity monitoring, granular access control, masking encryption, dynamic data encryption and end-point validation are incorporated. The proposed hybrid four layer health care model seems to be more effective in disease diagnostic the big data system.

Nowadays, due to expeditious development of internet cloud computing, data grow fastly at uncontrollable rate in almost all organizations [7]. Wal-Marts imports 2.4 petabytes of data approximately each hour in to databases, Facebook handles more than 250 million photos and 9 hundred million objects each and every day etc [8].

Due to this explosive growth of data, explications are must to glean valuable insights from the datasets. The effective data utilization is important as it is deemed as the building blocks for an organization. The effective data analysis are very useful in the disease diagnosis, sale forecasting, economic analysis, social network analysis and business management, etc.

Some organizations use formerly analytics in the organized data in form of reports. The early idea of big data were introduced in the paper “Visually Exploring Gigabyte Datasets in Real Time” and were published in 1999 [9]. In 2001, Doug Laney defines the big data characteristics in 3V’s i.e. Velocity, Volume and Variety in the paper 3D Data Management: Controlling Data Volume, Velocity and Variety. Hadoop is one of the most dominant frameworks that is used for managing and analyzing the unstructured big data. In general, the big data refers to voluminous and complex amount of data collected from various sources suchas web, mobile devices, enterprise applications and digital repositories that can not be easily managed with the use of traditional tools.

Big data is not only about large data size, but it is an act of storing and managing data for eventual analysis. As the persons are getting digitized, therefore, computing embroils data with greater volume, variety, and velocity. Big data offer tremendous opportunities in health care sector for improving efficiency and quality in health care, health threats detection, managing human health by diagnosis disease in early stage and in assisting better decisions.

They concluded that they have provided an in-depth description and brief overview of th big data in general and in health care system, which plays a significant role in health care informatics and influences greatly the health care system and big data four Vs in health care.

They also proposed use of the conceptual architecture to solve health care problems in big data with Hadoop terminologies, that involved the utilization of big data, generated by various levels of medical data and the developments for analyzing this data and also for obtaining answers to medical questions.

III. EXISTING METHODOLOGY

The existing system focuses on Naïve Bayes classification algorithm to detect disease among the given data set. The dataset is taken from kaggle. Preprocessing such as null value elimination is not processing in existing system. All features are taken during classification which increases time. Confusion matrix is prepared with accuracy score calculation. The drawbacks are:

1) NBS Classification is not given more accuracy for given new test data.
2) Feature reduction before classification is not carried out.
3) Decision Tree takes more time if the data set size is growing
4) Data columns with numeric values only take for NBS classification.

IV. PROPOSED METHODOLOGY

The proposed system focuses on SVM classification algorithms along with existing system algorithms. The dataset is taken from kaggle and preprocessed such as null value elimination. Important features are selected for better classification. Accuracy Score is calculated and printed. Confusion matrix is also calculated with accuracy score. The following modules are present in the proposed application.

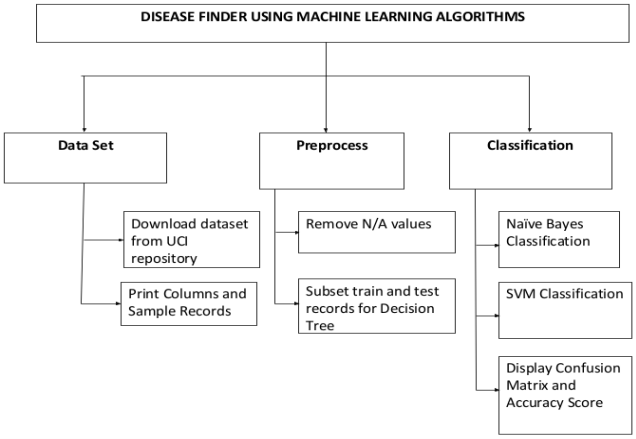


Fig 4.1

V. MODULE

A. DATA set Collection

The dataset which contains columns (itching skin_rash, nodal_skin_eruptions, continuous sneezing, shivering, chills, etc with prognosis as disease name) are saved in a single Excel workbook as records. This is the input for the project.

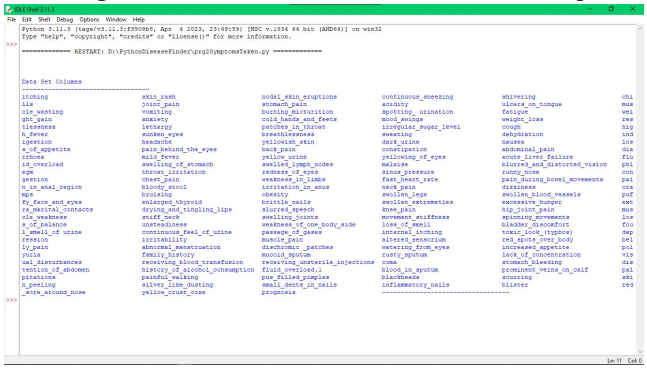


Fig 5.1

B. DATA set Subsetting

The dataset which contains columns (itching skin_rash, nodal_skin_eruptions, continuous sneezing, shivering, chills, etc with prognosis as disease name) are saved in a single Excel workbook as records. This is the input for the project in which 4920 (collectively (120 records for each disease) for training records and 15% of which is taken for testing records are split and given for classifications.

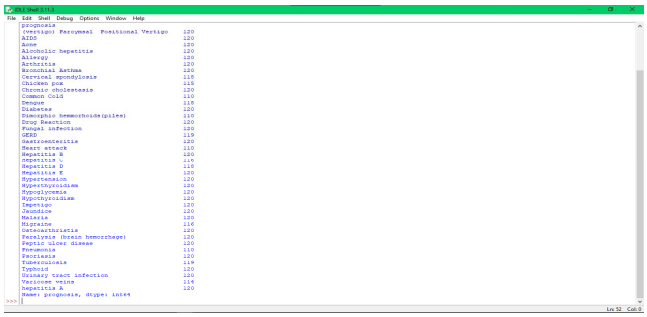


Fig 5.2

C. NBC Classification

In this module, Naive Bayes Classification is used in which 85% of the data in given data set is taken as training data and 15% of the data is taken as test data. The text (categorical) columns are converted into numerical values. Then the model is trained with training data and then predicted with test data. Of which, most of the disease are classified as itching present or not.

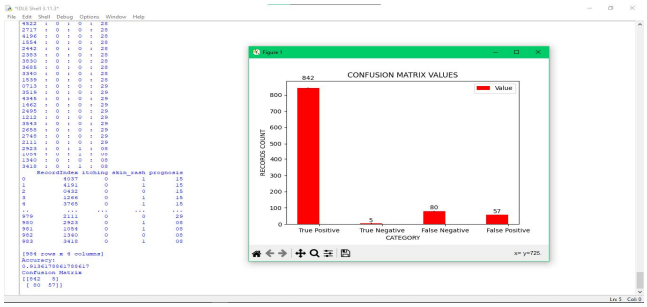


Fig 5.3

D. DT Classification

In this module, Decision Tree Classification is used in which 85% of the data in given data set is taken as training data and 15% of the data is taken as test data. The text (categorical) columns are converted into numerical values. Then the model is trained with training data and then predicted with test data. Of which, most of the disease are classified as itching present or not.

E. SVM Classification

In this module, Support Vector Machine Based Classification is used in which 85% of the data in given data set is taken as training data and 15% of the data is taken as test data. The text (categorical) columns are converted into numerical values. Then the model is trained with training data and then predicted with test data. Of which, most of the disease are classified as itching present or not.

Fig 5.4

VI. FINDINGS

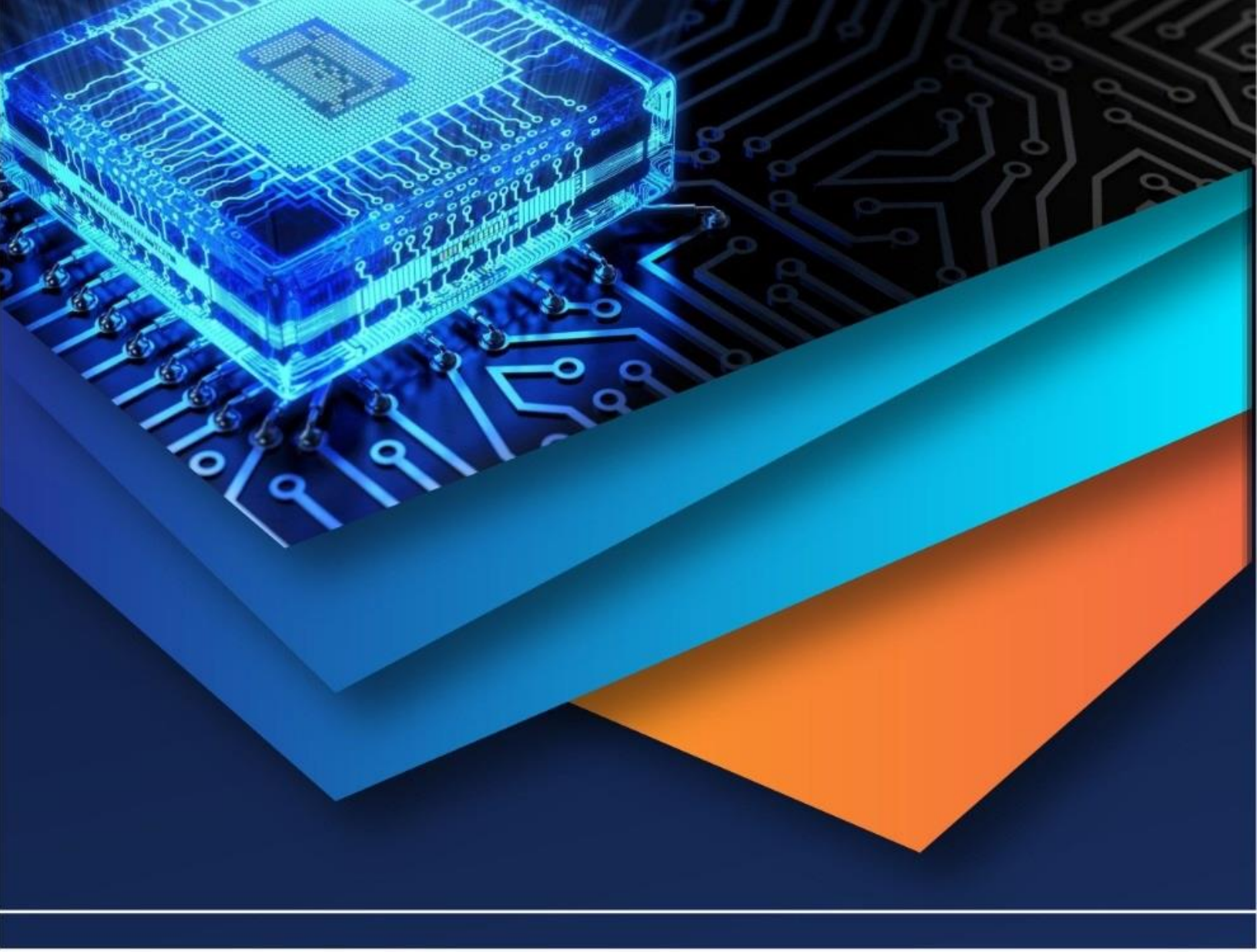
1) SVM classification is considered suitable for the given new test data.
2) SVM classification gives more accuracy.
3) SVM supports well even if the dataset size is big.
4) SVM based prediction model is worked out to find algorithm efficiency with different test data sizes.

VII. CONCLUSION

Since the finding of possibility of diseases in patients is a tough task for clinical persons and researchers as it requires more experience and medical tests need to be taken. The project finds the best classification algorithm which is suitable for providing accuracy improvement during classification of normal and abnormal persons. It contains Support vector machine, Naïve Bayes and decision tree classification with their accuracy score calculation. The applied SVM, NBS, DT classification helps for predicting the disease with higher accuracy in the new data set.

REFERENCES

[1] Gandomi and M. Haider, Beyond the hype: Big data concepts, methods and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015.
[2] O'Driscoll, J. Daugelaite, and R. D. Sleator, "Big Data", Hadoop and cloud computing in genomics, Journal of Biomedical Informatics, vol. 46, no. 5, pp. 774–781, 2013.
[3] L. P. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, vol. 275, pp. 314–347, 2014.
[4] M. Herland, T. M. Khoshgoftaar, and R.Wald, A review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 2, 2014.
[5] Ozgur C., Kleckner M. and Li Y. (2015) "Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities." Sage Open: 1-12.
[6] Prableen Kaur, Manik Sharma, Mamta Mittal,Big Data and Machine Learning Based Secure Healthcare Framework, Procedia Computer Science 132 (2018) 1049–1059
[7] Picciano A. G. (2012) "The Evolution of Big Data and Learning Analytics in American Higher Education." Journal of Asynchronous Learning Networks 16(3): 9-20.
[8] Sunil Kumar, Maninder Singh, Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools, BIG DATA MINING AND ANALYTICS ISSN222096-0654, 05/06, pp48–57, Volume 2, Number 1, March 2019, DOI: 10.26599/BDMA.2018.9020031
[9] Sagiroglu S. and Sinanc D. (2013) "Big Data: A Review. Presented in International Conference: Collaboration Technologies and Systems (CTS)." IEEE Xplore.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY