iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6077

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: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Disease Prediction using Machine Learning

Kriti Gandhi[1], Mansi Mittal[2], Neha Gupta[3], Shafali Dhall[4]
[1, 2, 3, 4]*Department of Information Technology, Bharati Vidyapeeth's College of Engineering, New Delhi, India*

Abstract: ***Machine learning has various applications and one of them is healthcare. There should be much more advanced medical facilities so as to provide the best possible treatment for the patients[3].Also there are many machine learning algorithms (such as KNN, Random Forest and Decision Tree Classifier algorithms and many more) which were selected and on the given data many algorithms were applied so as to produce the best results. We can say that when machine learning implemented in healthcare can lead to a high increase in patient satisfaction. so this research paper, will try to implement functions of machine learning in health facilities in a particular system[8]. Instead of directly performing treatment for the patient, if the disease is predicted beforehand using certain machine learning algorithms then the entire process of treatment can be made much more efficient[12]. There are also some cases which occur when early diagnosis of a disease is not performed or carried out. Hence disease prediction is a really important step while treating the patient. As it is said "Prevention is better than cure", the right prediction of disease would definitely lead to an early prevention of that particular disease[19].***
Index Terms: KNN, healthcare, Logistic Regression.

I. INTRODUCTION

Today healthcare industry has become a big money making business. The healthcare industry uses and produces quite a large amount of data which can be used to extract information about a particular disease for a patient. This information of healthcare will further be used for effective and best possible treatment for patient's health. This area also needs some improvement by using the informative data in healthcare sciences. But a major challenge is to extract the information from the data because the data is present in a huge amount so some data mining and machine learning techniques are used. The expected result and of this project is to predict the disease beforehand so that the risk of life can be prevented at an early stage and save life of people and the cost of treatment can be reduced to a particular extent. In India also we should adopt the non-manual system of medical treatment which is the best for improving and understanding the human health. The main motive is to use the concept of machine learning in healthcare to improvise the treatment of patients. Machine learning has already made it much more easier to identify and predict various diseases[7]. Predictive analysis of the disease with the help of many machine learning algorithms helps us to predict the disease and helps in treating the patients in an effective manner. Disease prediction using machine learning also uses the patient history and health data by applying vari- ous concepts like data mining and machine learning techniques and also some algorithm. Many works have also applied data mining techniques to the pathological data for prediction of some particular diseases. These approaches were intended to before- hand predict the re-occurrence of certain diseases[15]. Also, some approaches tried to do prediction while controlling the disease. The recent work of deep learning was in disparate areas of machine learning which have driven a shift to machine learning models that can learn and understand the hierarchical representations of raw data with some pre-processing. With the development of this concept called big data technology, more attention is paid to disease prediction.

II. PROPOSED SYSTEM

A. Creating the classification model

In this paper we first build a simple classified model using the concepts like sklearn library. Scikit-learn is one of the machine learning library used for Python. It also uses various algorithms like support vector machines, random forests, and k-neighbours, and it supports Python numerical and a concept called scientific libraries which include NumPy and SciPy.

B. Compile the model

Firstly, the model is to be defined, then only we can compile it. The different algorithms used to perform different proce- dures were compiled first and the performance of different algorithms was then tested on the basis of their respective accuracy scores. Another important step that came into picture was feature selection. As the dataset comprised of a large number of features, it is equally important to only logical to select and keep the essential features. Various procedures were used for feature selection with the algorithms.

©IJRASET: All Rights are Reserved

C. Fit the Model

When we have created and compiled the model, the next step is to train it and then fit the model on our respective data by using the fit() function. First, the model will fit on the original dataset, which comprises of all the features, however that might lead to an unwanted situation like overfitting. In order to tackle this model it was then fit over datasets with some other number of features. The optimal features for different algorithms were identified in order to obtain a range of high accuracy.

D. Evaluate the Model

After we have done the above step, we can evaluate the model based on certain parameters like accuracy, precision, processing speed and many more etc. Evaluating our model means understanding and correcting the pros and cons of our model in an efficient manner. In simple terms, we need to basically critically analyze the model's performance and then we can obviously improvise it as per our requirement.

III. TECHNOLOGY USED

A. Machine learning (ML)

ML algorithm is a process or set of processes that help the model to adapt to our dataset. An ML algorithm will specify the way the data is transformed from input to output and how the model will learn the appropriate transformation from input to the final output.

B. KNN (K- Nearest Neighbors)

KNN is a non-parametric and also a lazy learning algorithm. Its purpose is to use a particular database in which the data will point to several classes to predict the classification of a new sample. When we say one technique is nonparametric, we mean that it does not make any assumptions beforehand on the underlying data. In other words, the structure isdetermined from the data only

C. Logistic Regression

Logistic regression is a regression analysis to study when the dependent variable is binary in nature. Like all regression analyses, logistic regression is also a predictive analysis. Logistic Regression is used when a particular variable or target is categorical in its type. It uses maximum estimation as a method of approximation.

D. Decision Tree

Decision tree is a supervised learning algorithm which has a predefined target variable that is used in classification prob- lems. It works for concepts like categorical and continuous input and output variables for the model. In this technique, we will split the population of sample into some homogeneous sets based on most significant input variables.

E. Naives Bayes

It is basically a classification technique based on Bayes' Theorem with some changes like as follows. In simple terms, a Naive Bayes algorithm assumes that the presence of a certain feature in a class is not related to the presence of any other feature. For example, a fruit may be called as an apple if it is red, round, and is somewhat about 3 inches in its diameter. But if these features depend on each other then all of these features will independently contribute to the probability that this fruit is an apple only and that is the reason why it is known as 'Naive'.

F. Linear Discriminant Analysis algorithm

Linear Discriminant Analysis or Normal Discriminant Anal- ysis or Discriminant Function Analysis is a dimensionality reduction algorithm or a technique which is commonly used for the supervised learning of various processes that support machine learning in our particular model.

G. Random Forest

Random decision tree is a type of ensemble learning method for classification. They are used for correction for the habit of overfitting of the training set.

H. Classification Problems

These are used for studying the differences in groups i.e. separating two or more classes. It is used to highlight the These are used for studying the differences in groups i.e. separating two or more classes. It is used to highlight the features in higher dimensions of space into a lower dimension of space.

©IJRASET: All Rights are Reserved

LDA is represented in a simple manner which can be easily understood by everyone. The model comprises of the statistical properties of our data that is calculated for each and every class of our model. The same properties are used and calculated for the Gaussian for the case of multiple variables. Thus these problems are essential to be used.

IV. DATA SET USED

The dataset comprises of 133 columns, comprising of 132 varied symptoms experienced by patients suffering from a range of ailments. A total of 40 diseases are present in this dataset. [b28]

List of diseases being predicted

Fungal infection	Malaria
Allergy	Hepatitis A
Gerd	Hepatitis B
Acne	Hepatitis C
Pneumonia	Hepatitis D
Common Cold	Hepatitis E
Arthiritis	Alcohol Hepatitis
Diabetes	Heart Attack
Psoriasis	Chicken Pox
Dengue	Typhoid
Impetigo	Paralysis
Jaundice	Asthma
Drug Reaction	AIDS
Hyper Tension	Migraine
Tuberculosis	Cervical Spondylosis
Varicose Veins	Hypoglycemia
Paroxysmal Positional Vertigo	Chronic cholestasis
Drug Reaction	Peptic ulcer disease
Hyperthyroidism	Osteoarthristis
Gastroenteritis	Fungal infection

Table I. Algorithms Used For Classification

Algorithms			
Supervised		Unsupervised	
Regression	Classification	Clustering	Dimensionality Reduction
Decision	Tree Naive Bayes	K means	PCA
Linear regressions	SVM	mean shifts	Feature Selection
Logistic regression	KNN	K medoids	LDA

Table II. Results Obtained Using Embedded Method For Feature Selection

Algorithms	Accuracy Scores	Standard Deviation	No. Of Features
LOGISTIC REGRESSION	0.988790	0.003105	124
CART	0.963035	0.006168	61
KNN	0.960116	0.008465	52
NA¨IVE BAYES	0.969986	0.004774	52
SVM	0.956250	0.006451	52
LDA	0.950578	0.006168	77
RANDOM FOREST	0.808566	0.042578	52

©IJRASET: All Rights are Reserved

A. Data Preprocessing

Machine learning always faces a challenge whenever there is an inadequate dataset. To handle the missing data is an important step to make sure that the models and algorithms of machine learning produce the most accurate results. This includes creating a category object, fitting the encoder of the system to the 'prognosis' column present, and then to apply the encoder in our column to transform these categories into final integers. We also handled the missing values of the columns with less than 50.

B. Dividing Input Data into Training and Test sets

After processing our dataset to a certain considerable level, the next step is to specify the input and target variables present. Our input will be every column except the 'prognosis' column, since that's what we're attempting to predict—hence, it's our target variable. The data is then split into training and test sets, and a random seed of 5 is specified for the purpose of reproducing the results.

V. RESULT

The main aim of was to understand and improve the process of disease prediction and do a comparative study of algorithms in order to find the best suited algorithm. Apart from the comparison of accuracy scores of algorithms, data visualization was also done so as to understand the data and its trends in detail. The final comparison of the results obtained by the different algorithms employed for prediction of disease prediction from hospital data, revealed that CART model or simply said, decision tree gave the highest performance, followed by logistic regression, KNN, Na¨ıve Bayes, SVM, Random Forest, and LDA. RFE method of feature selection even though employed on only two models, made a major impact. The second preferred feature selection method was embedded method, whereas Pearson Correlation made the least impact. While analyzing the results we must also keep in mind the extent by which the features were reduced to obtain a certain level of accuracy. If we think from that perspective then embedded method made a major difference by reducing the number of features from 132 to 52 and still obtaining relatively good accuracy for all algorithms. Pearson Correlation is a less precise method for feature selection, hence its performance was lower than that of the other two methods which have higher precision.

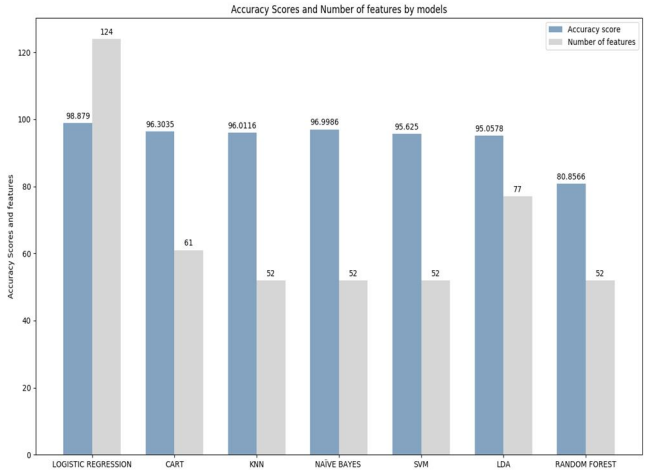


Fig. 1. Comparison of Results

The final comparison of the results obtained by the different algorithms employed for prediction of disease prediction from hospital data, revealed that Logistic Regression gave the high- est performance, followed by Naïve Bayes, CART or simply said, decision tree, KNN, LDA,SVM, and Random Forest. RFE method of feature selection even though employed on only two models, made a major impact. The second preferred feature selection method was embedded method, whereas Pearson Correlation made the least impact. While analyzing the results we must also keep in mind the extent by which the features were reduced to obtain a certain level of accuracy. If we think from that perspective then embedded method made a major difference by reducing the number of features from 132 to 52 and still obtaining relatively good accuracy for all algorithms. Pearson Correlation is a less precise method for feature selection, hence its performance was lower than that of the other two methods which have higher precision.

©IJRASET: All Rights are Reserved

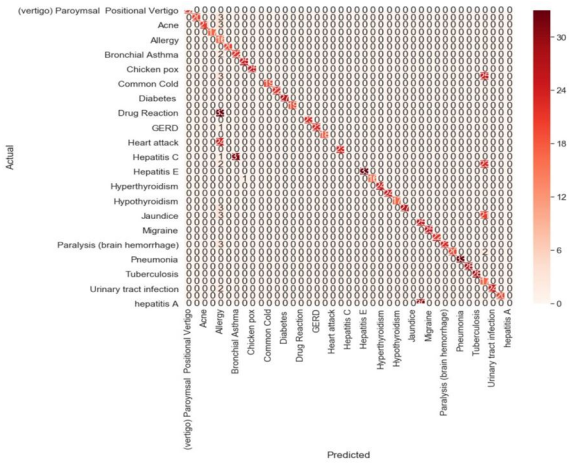


Fig. 2. Correlation Matrix

VI. CONCLUSION

With our proposed system, comparatively a good and higher accuracy is achieved. This is then used by researchers, physicians or doctors in order to provide the best treatment and medical care for the patients. Hence machine learning when used in healthcare can lead to an effective treatment and the patient is also well taken care of. Here we try to implement some of the functions of machine learning in healthcare into our system. Instead of direct diagnosis, when a disease is predicted for a patient then machine learning is implemented using certain machine learning algorithms and then healthcare can be made smarter and better. When we compare the different algorithms used for disease prediction from our dataset and the output we expect we get the best accuracy with Logistic Regression algorithm and KNN algo- rithm, whereas LDA algorithm had the lowest performance when compared to the other algorithms. Machine Learning (ML) provides methods, processes and certain techniques that can help solving the issue of diagnostic problems in a simpler and modernized variety of medical domains. ML is nowadays being used for the prediction and analysis of the clinical works. ML is also currently being used for the process of data analysis, such as detection of errors in the dataset and for dealing with incorrect data present in our system. It is a debatable topic that the perfect use and implementation of ML algorithms can act as great source of help in the integration of computer systems in the field of healthcare to facilitate and enhance the work of doctors and finally leading to improve the efficiency and quality of our medical care for the respective patients.

VII. FUTURE SCOPE

As nowadays we can clearly witness the increase in use of computers and technology to consider a huge amount of data, computers are being used to perform various complex tasks with commendable accuracy rates. Machine learning (ML) is a collection of multiple techniques and algorithms which permit computers to execute such complex tasks in an simplified manner. It is also used in both academics which is for students or learners and also in industry to make accurate predictions and use these diverse sources of dataset and information. Till date we can say we have grown in the fields of big data, Machine learning, and data sciences etc and have been a part of one of those industries which were able to collect such data and the staff to transform their goods and services in a desired manner. The learning methods developed for these industries and researches offer excellent potential to further improvise medical research and clinical care for the patients in the best possible manner. Machine learning uses mathematical algorithms and procedures which are used to describe the relationship between variables used in the model and the others. Our paper will explain the process of training the model and learning a suitable algorithm to predict the presence of a particular disease from the sample of the tissue based on its features. Though these algorithms work in different and unique manners depending on the way in which they are developed and used by the researchers. One way is to consider their supreme goals. The goal of our paper and statistical methods is to reach to a conclusion about the data which are collected from a wide variety of samples from our population. Though many techniques, like linear and logistic regression, are able to predict the diseases. For example, consider a case where, if we can create a model which described and understood the relationship between clinical variables and their transience then we can follow the organ transplant surgery i.e. we would need the factors and features which differentiate low mortality rate from high if we can develop such outcomes and reduce mortality rate to a desired rate in the near future also nothing can be said to be better than such situations.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

REFERENCES

[1] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.
[2] K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, ”Performance comparison of three data mining techniques for predicting kidney disease survivabil- ity”, International Journal of Advances in Engineering Technology, Mar. 2014.
[3] Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Anal- ysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018.

©IJRASET: All Rights are Reserved

[4] Boshra Brahmi, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journals of Multidisci- plinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164- 168.
[5] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration," Journal of biomedical informatics, vol. 53, pp. 220–228, 2015..
[6] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53- 60,March 2016.
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: Univer- sity Science, 1989.
[8] Hanka, R., Harte, T.P., Dixon, A.K., Lomas, D.J., and Britton, P.D. "Neural networks inthe interpretation of contrast-enhanced magnetic res- onance images of the breast". InProceedings of Healthcare Computing, Harrogate, UK, 275-283, 1996.
[9] Hau, D., and Coiera, E. "Learning qualitative models of dynamic systems". MachineLearning, 26, 177-211, 1997.
[10] . Ifeachor, E.C., and Rosen, K. G. (eds.) Proceedings of the International Conference onNeural Networks and Expert Systems in Medicine and Healthcare, Plymouth, UK, 1994.
[11] Innocent, P.R., Barnes, M., and John, R. "Application of the fuzzy ART/MAP andMinMax/MAP neural network models to radiographic image classification". ArtificialIntelligence in Medicine, 11, 241-263, 1997.
[12] Jankowski, N. "Approximation and classification in medicine with IncNet neuralnetworks". In [38].
[13] Karkanis, S., Magoulas, G.D., Grigoriadou, M. and Schurr, M. "De- tecting abnormalities incolonoscopic images by textural description and neural networks". In [38].
[14] Karkanis, S., Galoussi, K. and Maroulis, D. "Classification of endoscopic images based ontexture spectrum".
[15] Pouloudi, A. "Information technology for collaborative advantage in health care revisited".Information and Management, 35, 6, 345-357, 1999.
[16] Pranckeviciene, E. "Finding similarities between an activity of the different EEGs bymeans of a single layer perceptron". In [38].
[17] Prentza, A. and Wesseling, K.H. "Catheter-manometer system damped blood pressuresdetected by neural nets". Medical and Biological Engi- neering and Computing, 33, 589-595, 1995.
[18] Reategui, E.B., Campbell, J.A., and Leao, B.F. "Combining a neural network with case-based reasoning in a diagnostic system". Artificial Intelligence in Medicine, 9, 5-27, 1996.
[19] Ridderikhoff, J. and van Herk, B. "Who is afraid of the system? Doctors' attitude towardsdiagnostic systems". International Journal of Medical Informatics 53, 91-100, 1999.
[20] Ruseckaite, R. "Computer interactive system for ascertainment of visual perceptiondisorders".
[21] Schurr, M. "The Role of Machine Learning Methods in Endoscopic Techniques".
[22] Strausberg, J. and Person, M. "A process model of diagnostic reasoning in medicine".International Journal of Medical Informatics, 54, 9-23, 1999.36. Zupan, B., Halter, J.A., and Bohanec, M. "Qualitative model approach to computerassisted reasoning in physiology". In Proceedings of Intelligent Data Analysis in Medicineand Pharmacology-IDAMAP98, Brighton, UK, 1998.37
[23] M Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, "Disease prediction by machine learning over big data from healthcare communities", IEEE Access, vol. 5, no. 1, pp. 8869-8879, 2017. B. Qian, X. Wang, N. Cao, H. Li, Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction", Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070-1093, 2015.
[24] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system", IEEE Commun, vol. 55, no. 1, pp. 54-61, Jan. 2017.
[25] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data", IEEE Syst. J, vol. 11, no. 1, pp. 88-95, Mar. 2017.
[26] . Qiu, K. Gai, M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing", Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), pp. 184-189, Nov. 2016.
[27] S. Leoni Sharmila, C. Dharuman, P. Venkatesan, "Disease Classification Using Machine Learning Algorithms - A Comparative Study", Interna-

©IJRASET: All Rights are Reserved

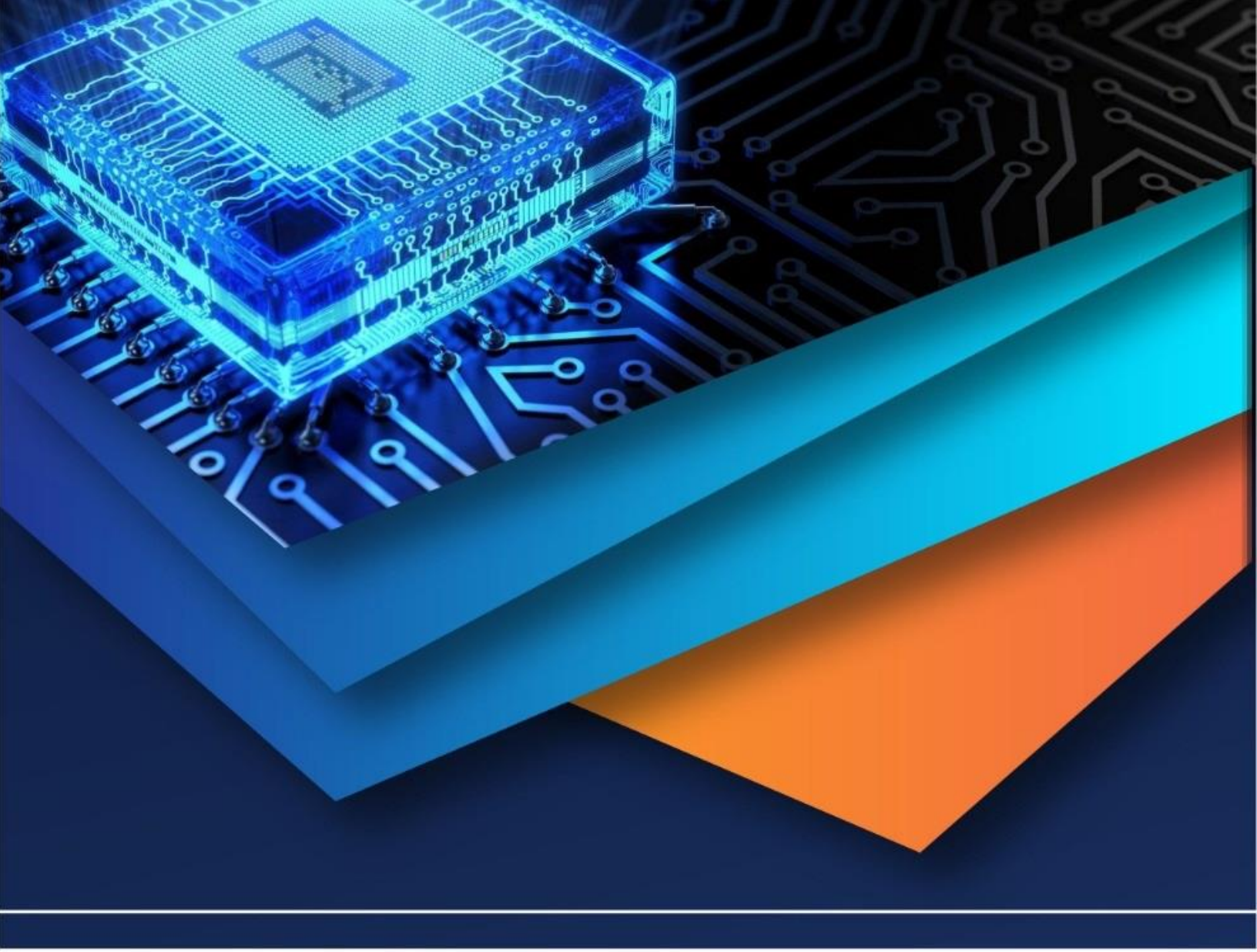
tional Journal of Pure and Applied Mathematics, vol. 114, no. 6, pp. 1-10, 2017.

[28] Allen Daniel Sunny, Sajal Kulshreshtha, Satyam Singh, Srinabh, Mohan Ba, H Sarojadevi, "Disease Diagnosis System By Exploring Machine Learning Algorithms", International Journal of Innovations in Engineer- ing and Technology (IJIET), vol. 10, no. 2, May 2018.

[29] Shraddha Subhash Shirsath, "Disease Prediction Using Machine Learn- ing Over Big Data", International Journal of Innovative Research in Science, vol. 7, no. 6, June 2018.

[30] Shubham Rathi, Mahesh Motwani, Manish Ahirwar "Data-Driven Clin- ical Decision Support System for Medical Diagnosis and Treatment Recommendation" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)