iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41860

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 10 Issue IV Apr 2022- Available at www.ijraset.com

Heart Disease Prediction Using Logistic Regression Algorithm

Bhagyesh Randhawan[1], Ritesh Jagtap[2], Amruta Bhilawade[3], Durgesh Chaure[4]
[1, 2, 3, 4]*Department of Information Technology, Parvatibai Genba Moze College of Engineering, Pune*

Abstract: ***Heart - a primary organ of our circulatory system. Which keeps blood that's full of oxygen circulating throughout your body. From past two decades Heart-disease remained as a leading cause of death at global level. Statistics illustrate the lethality of cardiovascular disease by showing the percentage of deaths caused by heart attacks worldwide. Therefore, it is crucial to predict the condition as earliest as possible time. Cardiologist have limitations, they cannot predict heart disease risk to a high degree of accuracy. So, a reliable, accurate and feasible system is required to predict such diseases in time for proper treatment. In order to automate analysis of large and complex medical datasets, Machine Learning algorithms and techniques have been applied. Machine learning techniques have been increasingly used by researchers in the health care industry and by professionals to diagnose conditions related to the heart. A quick and efficient detection technique is needed to reduce the high death rate caused by heart diseases. Here, machine learning algorithms and data mining techniques play a very crucial role. Using machine learning algorithms, this research aims to predict the occurrence of heart disease in a patient.***
Keywords: Machine Learning, Supervised Learning, unsupervised Learning, Logistic Regression, Cardiovascular diseases

I. INTRODUCTION

The number of people suffering from cardiovascular disease is on the rise. Numerous factors carry the risk of developing this disease, such as age, high blood pressure, high cholesterol, diabetes, hypertension, genes, obesity, and unhealthy lifestyles. It is possible to identify a variety of symptoms by observing physical signs like chest pain, shortness of breath, dizziness, and wearing yourself out easily. Even though these diseases were found to be the leading cause of death, they have been classified as the most manageable and preventable illnesses. Identification of cardiovascular diseases is a difficult process. The early detection of cardiovascular disease is crucial since its complications can have an impact on a person's life as a whole.

The signs of a woman having a heart attack are much less noticeable than the signs of a male. In women, heart attacks may feel uncomfortable squeezing, pressure, fullness, or pain in the center of the chest. It may also cause pain in one or both arms, the back, neck, jaw or stomach, shortness of breath, nausea and other symptoms. Men experience typical symptoms of heart attack, such as chest pain, discomfort, and stress. They may also experience pain in other areas, such as arms, neck, back, and jaw, and shortness of breath, sweating, and discomfort that mimics heartburn.

Cardiovascular disease diagnosis and treatment are very complex. While invasive-based techniques are still employed through analysis of the patient's medical history, reports of physical examinations by the physician tend to be less accurate and take a long time to prepare. For this reason, a support system is implemented to predict cardiovascular disease through a machine learning model. A machine-learning approach may improve accuracy by leveraging the complex interactions between risk factors.

II. LOGISTIC REGRESSION

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model, like the Adaline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification. Scikit-learn has a highly optimized version of logistic regression implementation, which supports multiclass classification task.

©IJRASET: All Rights are Reserved

Logistic regression is another powerful supervised ML algorithm used for binary classification problems (when target is categorical). The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function defined below to model a binary output variable. The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio.

Formula of Logistic Regression Sigmoid function: $f(x) = \frac{1}{1+e-(x)}$

In the logistic function equation, x is the input variable. Let's feed in values −20 to 20 into the logistic function. the inputs have been transferred to between 0 and 1.

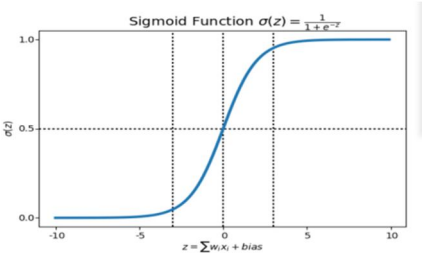


Figure 1: Sigmoid Function

III. METHODOLOGY

Several factors that affect the human cardiovascular system are examined in this study. The process begins with retrieved data, analysis of correlation between variables, splitting of the data, and prediction with the logistic regression algorithm, ending with data validation.

A. Data Retrieval

The first process is Data Retrieval. In this process, Heart Disease UCI Dataset will be used. It will be imported into the PyCharm software. The data obtained are categorical data and numerical data. The data in this study contain 14 variables with 76 attributes and 304 responses as the basis for analysis. First variable is age with units in years (age). Second, the gender with value one means male and value 0 means female (sex). Third, the variable type of chest pain (cp). Fourth, the variable trestbps-resting blood pressure in mm Hg at admission to hospital (tresbps). Fifth, chol-serum cholesterol variable in mg/dl (chol). Sixth, the fbs variable, which is blood sugar when fasting with a value of 1, means true, and 0 means false (fbs). The seventh variable is resting electrocardiographic outcome variables (restecg). Eighth, the maximum thalach heart rate variable is reached (thalac). Ninth, the exacting-exercise variable induced angina with value 1 means yes, and value 0 means no (exang). Tenth, oldpeak-ST variable depression caused by exercise relative to rest (oldpeak). Eleventh, the slope variables of the peak training segment ST (slope). Twelfth, ca-number of main vessels with values 0 to 3, colored by fluoroscopy (ca). Thirteenth, thal-3 variable means normal; 6 means permanent disability; 7 means reversible defects (thal). Fourteenth, the target variable with a value of 1 or 0 (target).

Name	Type	Description
Age	Continuous	Age Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang Continuous Maximum heart rate achieved	Discrete	Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Result	Discrete	0=no heart disease, 1=have heart disease

B. The Correlation between Variables Analysis

Besides, to facilitate data analysis, all variables in the imported dataset will be visualized in the form of a histogram to facilitate the reading of the data in general. In the process, Analyse the Correlation between Variables; the correlation between variables is examined to prove that the method to be used is the logistic regression model is the right model. Relationships between variables in the available dataset will be plotted in the form of a matrix. This is also done to check whether there is multicollinearity between variables in the dataset.

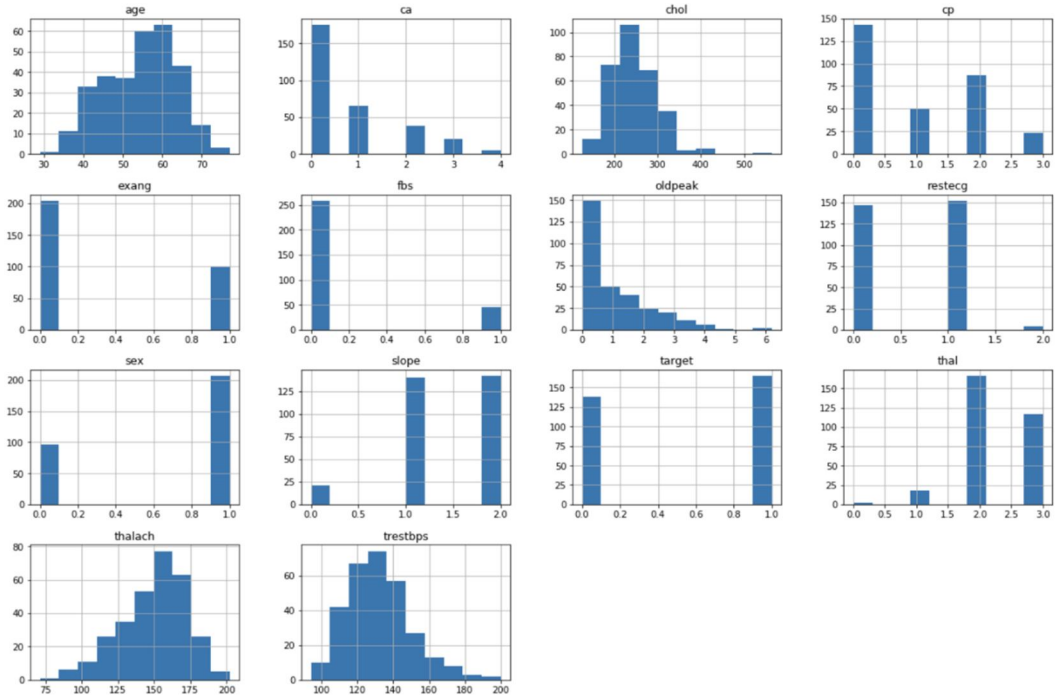


Figure 2: Variables in Data

C. Data Preparation

The dataset imported in PyCharm will be divided into two parts, namely training data and testing data. Training data is used as a basis for building models. Meanwhile, testing data is used as a basis for testing or validating the model. In this data preparation process, 304 data will be sampled. Then the data will be partitioned into train data and test data.

D. Prediction with Logistic Regression Algorithm

In this process, the data that has been partitioned in the previous process will be used. Prediction using the logistic regression method will produce several data that can be used as a basis for concluding to make predictions.

E. Data Validation

The technique used to validate the results is the method of the confusion matrix and K-fold cross validation with 10-fold. By using a confusion matrix, the accuracy of the use of the logistic regression model can be known. Besides, the use of the K-fold cross-validation method, produces values of errors that may occur when using a logistic regression model.



Figure 3: Confusion Matrix

IV. DATA ANALYSIS AND DISCUSSION

A. Data Analysis

The dataset obtained by the researcher as a basis for analysis is imported into PyCharm. The data retrieval process is also performed in the data visualization to see the value of each variable involved in the overall research analysis.

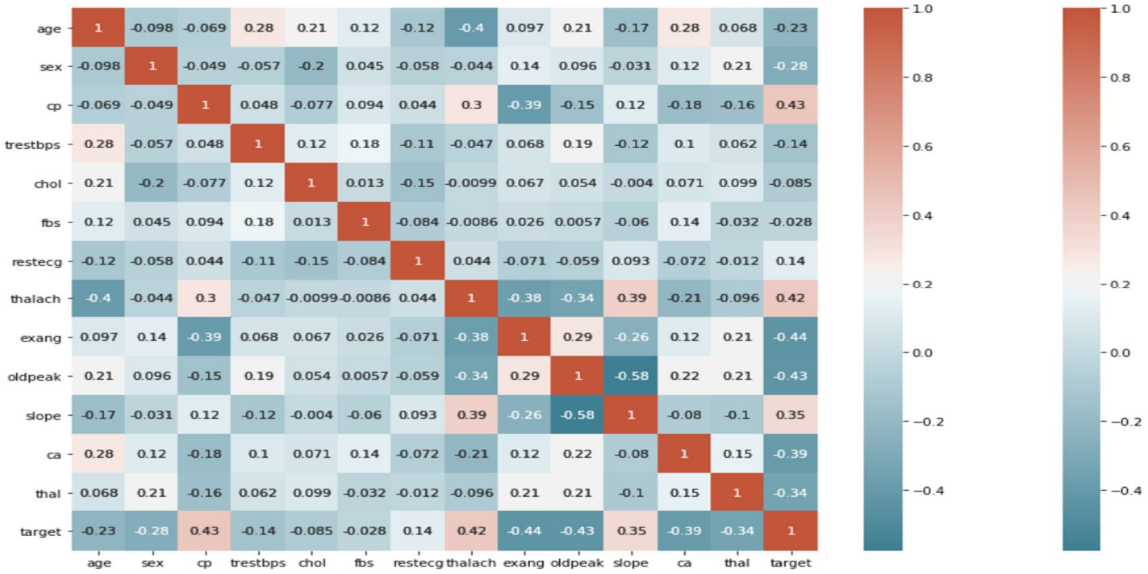


Figure 4: Correlation Matrix

In this process, the correlation between variables will be examined, which will be used as a basis for analysis to predict cardiovascular disease. Based on the matrix it was found that the variables induced angina (exang), chest pain type (cp), ST depression induced by exercise relative to rest (oldpeak), maximal heart rate (thalac) had a strong correlation with the target variable. Meanwhile, blood sugar (fbs) and cholesterol (chol) levels do not correlate with the target variable. Meanwhile, among the independent variables, there is a strong correlation between the slope and oldpeak variables. Besides, thalac, exhang, oldpeak, and slope variables are also strongly correlated. Strong correlation also applies to variables Exang, cp, and thalac. It proves that there is no multicollinearity in the relationship between variables where each independent variable does not correlate with each other. Data that has been imported will be taken as many as 293 random data as a basis for analysis. The data is divided into train data and test data. The data on the next page is data from train_data and test_data that will be used in this study. The training data is used to build a logistic regression model using the glm () function because logistic regression is included in the generalized linear model with binomial type families. Based on the results of using the logistic regression method, it is predicted that the sex, cp, trestbps, restecg, ca and that variables influence the target variable at an alpha value of 5% significantly. The selected variables are the variables that significantly affect the target variable. In logistic regression, the effect of each variable on the target variable can be seen from the odds ratio value. For example, for the sex variable having a coefficient value of -1.547601 with a reference category with a male value, the odds ratio value is 4.2655 which means that for male patients, the odds of getting heart disease are 4.2655 times the female odds or it can be said the tendency of men to heart disease is higher than women. For the trestbps variable with a coefficient value of -0.029713, it is found that the odds ratio value is 0.0822 which means that for the trestbps variable there will be a significant increase when trestbps enters the value 0.0822 mmHg. On the other hand, the thalach variable with a coefficient of 0.032028 will have an odd of 0.08856 which means that at that value there will be a significant change in the performance of the heart rate or cardiovascular rate. The exang1 variable is exercise-induced angina with an estimated coefficient of -1.05855 so that the exang variable with a reference value of 1 will have an odd of 2.92710 which means that if the value is achieved then cardiovascular performance will decrease. Next is the variable ca with reference ca values 1, 2, and 3. Ca1 with an estimated coefficient of - 1.430110 will have odds of 3.955, while ca2 with an estimated ratio of -3.329874 will have odds of 9.1777 and ca3 with an estimated factor of -0.553711 will have odds in the amount of 1.5261. It proves that when the number of fluoroscopy vessels reaches its value odds, this will have an impact on decreasing cardiac performance which will affect the increased potential for cardiovascular disease. Besides, the composition of value 0 and value 1 on variable target is 97:116, which is still fairly balance, so the result will be reliable and free from any imbalanced dataset problems.

B. Discussion

This study involved thirteen factors that affect cardiovascular performance as variables to build a logistic regression model. Among the variables, it is not found that there is no significant relationship between variables. Therefore, the potential for multicollinearity in this study tends to be smaller. This study uses a logistic regression algorithm as a solution to the problem. With the use of the algorithm, it was found that the logistic regression algorithm was classified as an effective and efficient algorithm in predicting the main factors causing cardiovascular disease as the problem raised in this study. With an accuracy of 85.45% and an error rate that tends to be small at 0.1406565, the logistic regression algorithm can be said to be successful in predicting factors that affect cardiovascular performance significantly. Especially with calculations using specific estimated values, it can be obtained the probability of the potential for cardiovascular disease in a person. By modelling data and predicting data using a logistic regression algorithm, it was found that not all factors had a significant influence on the performance of the cardiovascular system. The factors that affect cardiovascular performance are gender, trestbps - blood pressure level, thalach - heart rate, and canumber of vessels affected by fluorosophy. By obtaining an estimated value of these factors, probabilities can be obtained related to the potential for cardiovascular disease in a person.

V. CONCLUSION

This study utilized the Heart Disease UCI dataset which consisted of fourteen variables including age, sex, cp, fbs, restecg, thalac, exang, oldpeak, slope, ca, thal, and target to determine how well the logistic regression algorithm performs in predicting cardiovascular disease. Based on the results of data validation, the accuracy of the prediction results is 85% and the error rate tends to be small at 0.1406565. These results demonstrate that this algorithm can be utilized as a prediction algorithm in the current study. According to cardiovascular disease predictions, gender, trestbps - blood pressure level, thalach - heart rate, and the number of vessels affected by fluoroscopy have significant influence on possibility of heart disease. Increase in these variables value will have an impact on overall cardiovascular performance. The cardiovascular performance will decrease, whereas cardiovascular disease potential will increase. As predicted by the logistic regression algorithm, the main factors causing cardiovascular disease are gender factors, blood pressure level factors, heart rate level factors, and the color of the vessels (blood vessels).

REFERENCES

[1] Felman, Cardiovascular Disease: Types, Symptoms, Prevention, and Causes. The Technical Writer's Handbook. Mill Valley, CA: 1989.
[2] S. Palaniappan, and R. Awang," Intelligent Heart Disease Prediction System Using Data Mining Techniques", IEEE/AAACS International Conference on Computer Systems and Application, Doha, pp.108-115, 20008.
[3] M. Shouman, T. Turner, and R. Stocker," Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in The Diagnosis of Heart Disease Patients", Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldCom), 2012.
[4] Y. Boateng, and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research". Journal of Data Analysis and Information Processing, Vol.7. No.4, pp.190-207, 2019.
[5] J. Harrell and E. Frank, "Regression Modelling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis". Springer, 2015.
[6] Tyagi, Shivani, and S. Mittal, "Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning." Proceedings of ICRIC 2019. Springer, Cham, pp.209-221, 2019.
[7] Y. Khourdifi, and M. Bahaj, K-Nearest Neighbour Model Optimized by Particle Swarm Optimization and Ant Colony Optimization for Heart Disease Classification. In: Farhaoui Y., Moussaid L. (eds) Big Data and Smart Digital Environment. ICBDSDE 2018. Studies in Big Data, vol 53. Springer, Cham.
[8] Mr. ChalaBeyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique", International Journal of Pure and Applied Mathematics, 2018.
[9] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques" IEEE Access 7 (2019): 81542-81554.
[10] Ali, Liaqat, et al, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure" IEEE Access 7 (2019): 54007-54014.
[11] Singh Yeshvendra K., Nikhil Sinha, and Sanjay K. Singh, "Heart Disease Prediction System Using Random Forest", International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2016.
[12] Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning ,Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99
[13] B.L DeekshatuluaPriti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm" International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.
[14] Michael W.Berryet.al, Lecture notes in data mining, World Scientific(2006)
[15] S. Shilaskar and A.Ghatol, "Feature selection for medical diagnosis :Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[16] C.-L. Chang and C.-H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.

[17] T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013. [10] Y. C. T. Bo Jin, "Support vector machines with genetic fuzzy feature transformation for biomedical data classification.," Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.

[18] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," Expert Syst. Appl., vol. 41, no. 9, pp. 4434–4463, Jul. 2014.

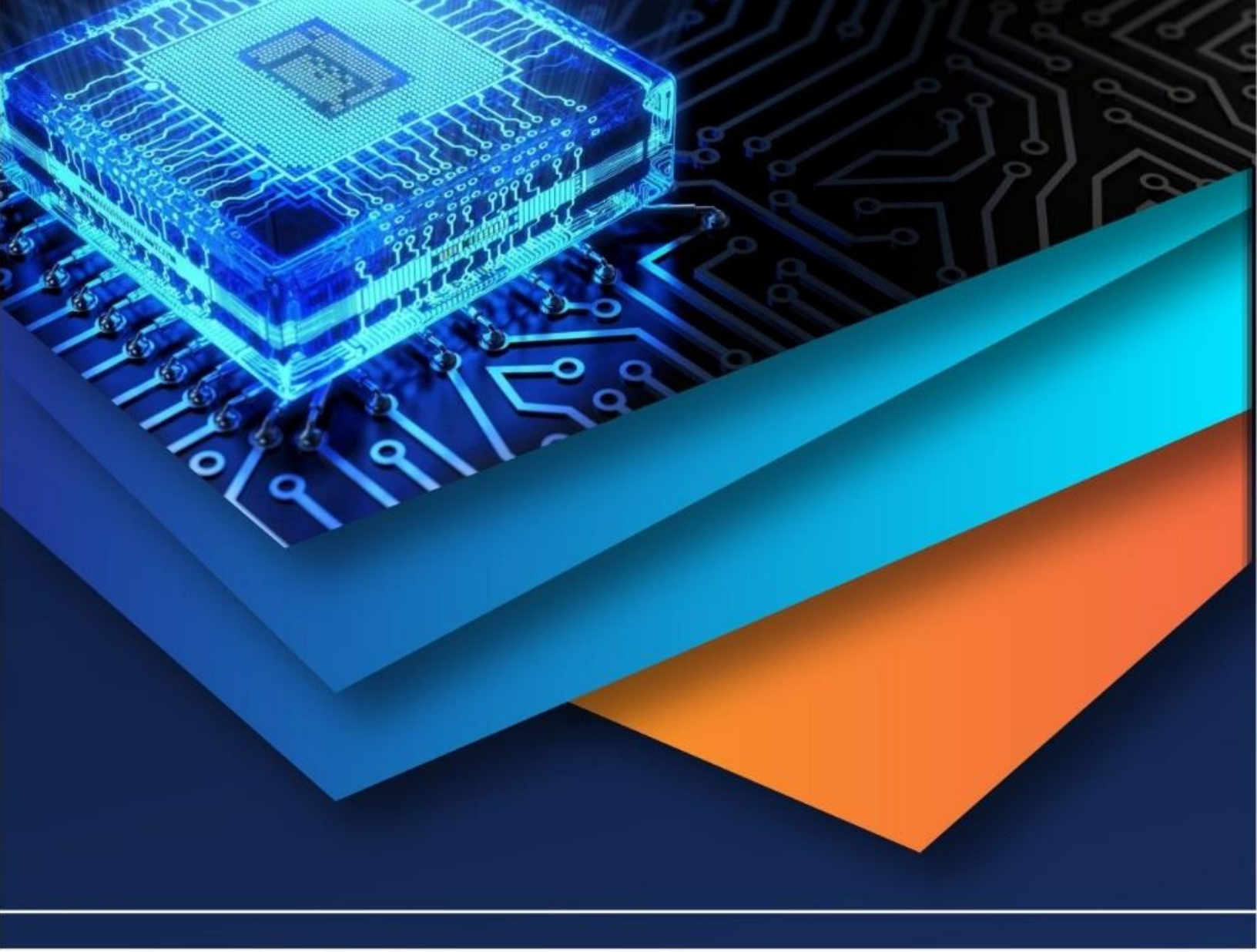
[19] Hassanien and T. Kim, "Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks," J. Appl. Log., vol. 10, no. 4, pp. 277–284, Dec. 2012.

[20] Sanjay Kumar Sen 1, Dr. Sujata Dash 21Asst. Professor, Orissa Engineering College, Bhubaneswar, Odisha – India. Domingos P and Pazzani M. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", in Proceedings of the 13th Conference on Machine Learning, Bari, Italy, pp 105-112,1996.

[21] Elkan C. "Naive Bayesian Learning, Technical Report CS97-557", Department of Computer Science and Engineering, University of California, San Diego, USA, 1997.

[22] B.L Deekshatulua Priti Chandra "Reader, PG Dept. Of Computer Application North Orissa University, Baripada, Odisha – India. Empirical Evaluation of Classifiers Performance Using Data Mining Algorithm"

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



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)