iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44018

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 VI June 2022- Available at www.ijraset.com

Prediction of Cardiovascular Disease Using PySpark Techniques

Ms. Dayana[1], K Keerthika[2], E Bibilin Manuela[3], J Julie Christina[4]
[1]*Associate professor - Department of CSE - Jeppiaar Institute of Technology*
[2, 3, 4]*UG-student- Department of CSE, Jeppiaar Institute of Technology, Sriperumbudur, India'*

Abstract: ***On a day after day, human life is affected by differing kinds of diseases that is why their life is in distress. cardiovascular disease may be a generic class of disease that's effective in spreading infections and notably, it affects the heart and veins. it's determined that vessel diseases have become modest in old individuals besides in children too. it's terribly requisite to portend this sort of illness within the starting phases; many varieties of tests square measure used for diagnosticating these ailments. This implementation has been done by employing a big data tool that's Apache Spark and victimization spark's MLlib and PySpark libraries that square measure integrated with it. Apache Spark is among the foremost wide used big data technologies, and it's a stack of some libraries that are Spark SQL, Spark MLlib, Spark Streaming, etc. This analysis work aims to create a prediction model to predict whether or not people have cardiovascular disease or not, using machine learning classification techniques that embrace logistic regression, decision tree, random forest to enhance the performance of models. They compared the analysis of all applied machine learning models. The results obtained are compared with the results of existing models within the same domain and located to be improved.***
Keywords: heart, blood vessels, Xampp server, data analytics, cardiovascular diseases.

I. INTRODUCTION

Health care means that the maintenance or advancement of health through interference and diagnosing of individuals. these days health care is increasing day by day because of life-style and hereditary reasons. cardiovascular disease has become the deadliest enemy caused by fat suppression. This sickness happens because of overpressure within the physical body. someone affected by cardiovascular disease can't be cured simply. Therefore, diagnosis patients at the proper time is that the toughest task within the medical trade and it has to be diagnosed within the early stages to reduce the danger on the patient in future. every physical body has totally different numbers for pressure, cholesterol, and heart rate. however traditional values would be, pressure is 120/80, cholesterol is 200 mg/dL and heart rate are 72. In ancient methodology doctors might build some mistakes in found a unwellness, however currently days Machine learning play an excellent appear prediction. And within the existing system, they use some formula of machine learning to predict cardiopathy. we will predict internal organ unwellness employing a type of parameters within the dataset. The obtained results are compared with the results of existing models among an equivalent domain and located to be improved. the information of cardiovascular disease patients collected from the UCI laboratory is employed to find patterns with Random Forest and call Tree. to create this method user friendly. it's so helpful to mix these machine learning algorithms with medical knowledge sources. This paper suggests PySpark MLlib that is helpful for predicting the uncertainty level of heart condition for a personal supported collective characteristic.

II. LITERATURE REVIEW

In 2018, Dewan, Ankita Sharma, Meghna proposed a hybrid technology with the potential to solve complex Scepticism That Is Inevitable to a Diagnosis of Heart disease that can help doctors diagnose the condition. Proposed hybrid technology based on a dataset Features that were taken from the UCI repository. The evaluation metrics used are accuracy and sensitivity. Comparison performance between decision trees, Naïve Bayes, SVM and ANN. Results show that ANN outperforms all others Classification for non-linear data.

In 2019, Shamsullah, M. Badi, A. Ghazanfari built a model using a mix of descriptive and predictive analysis of KDD (knowledge discovery in databases). The authors determine the number of clusters using clustering index. After that, the authors planted some decision trees Methods and artificial neural networks for all groups. They Used the original dataset to build the aggregated model From the Heart Clinic Database. The results showed that the cart decision tree model achieved the best of all methods.

©IJRASET: All Rights are Reserved

In 2020, Komal Kumar developed a proposed method for heart disease. he used heart disease Dataset that was collected from UC Irvine (UCI) Repository, which had 10 attributes.

Classification Techniques used for random forest, decision tree, logistic regression, k nearest neighbour and support vector machine Creating a predictive model of cardiovascular disease. Later Comparing the model's performance, the researchers have Shown is that the random forest machine learning model best accuracy achieved with area under 85.71%. Therefore, he used a random forest machine. Learning model that helps in building a system for classifying Patients affected by cardiovascular disease.

In 2020, Ahmed created a system using the Spark ml library that is among the massive information platform Apache Spark. to create the prediction model for diabetes patients, they need used classification ways that included are support vector machine algorithmic rule, decision tree algorithm, logistic regression rule, naive Bayes algorithm, and random forest algorithmic rule. Following that, evaluated all models exploitation some matrices like accuracy, recall, and exactness then they found that the logistic regression model accomplished the most effective proportion score of Accuracy (82%), Recall (92%), and preciseness (82%).

III. EXISTING SYSTEM

In the past decades, heart disease is a common and dangerous disease caused by the suppression of fat. This disease occurs due to high pressure in the human body. In traditional method doctors may make some mistakes in detecting disease, but nowadays machine learning plays a big role in prediction. And in the current system, they use some of machine learning's algorithms to predict heart disease. We can predict heart disease using different parameters in the dataset. The results obtained are compared with the results of existing models in the same domain and found to be improved. Data from heart disease patients collected from the UCI laboratory are used to find patterns with random forest and decision trees. To make this system user friendly.

IV. PROPOSED SYSTEM

We studied about using different machine learning to predict key features predictive of heart disease. Our goal is to boost the performance of the model by removing unneeded and insignificant attributes from the dataset and solely grouping those who are most informative and helpful for the classification task. We proposed PySpark Machine Learning Library (MLIB) technology for heart disease prediction of critical features. The ML process starts with the pre-processing data phase, which is followed by classification modelling, based on the results with improved accuracy. It looks like Decision tree performed the best here, with an accuracy of 84%. Logistic Regression was third at 63% and Random forest algorithm came in second at 75%. we get live data from user with the help of Xampp server. That live data will be added to the existing dataset. We have an admin login, here we can add live data to the dataset. Firstly, the customer user should register himself on the registration page in the web application. Once the user logs into the system he gets all the access, and the user gives inputs to predict heart disease. Thus, the most focus of the system is to create use information analytics to predict the presence of the sickness and level of illness among patients.

V. ARCHITECTURE DIAGRAM

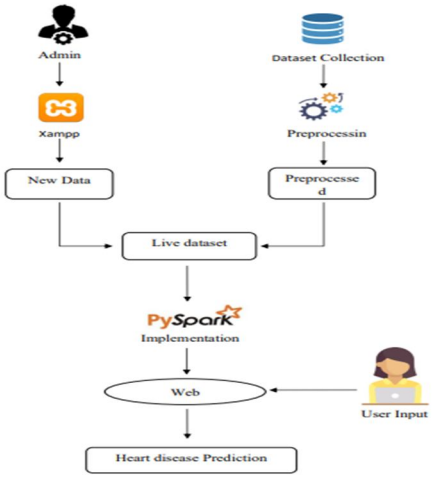


Figure 1: Architecture Diagram

A. Machine Learning Algorithms

Machine learning (ML) is a collection of algorithms that are capable of learning to execute tasks that is forecast and classify with help of a dataset. In this phase, for predicting the model of cardiovascular ailment, machine learning algorithms are implemented on (training & testing) dataset. We have used classification algorithms that included logistic regression classifier, decision tree classifier, random forest classifier has been used to apply machine learning models.

B. Logistic Regression Algorithm

Logistic regression is a classification algorithm that uses for binary classification. We assume all predictors are independent of each other and the outcome obtained by the dependent variable that comes with two target values which are 0 or 1, or yes or no.

```
In [36]: from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(
    labelCol='target',
    predictionCol='prediction',
    metricName='accuracy')
accuracy = evaluator.evaluate(lr_predictions)
print('Train Accuracy = ', accuracy)
LR_SC=accuracy*100

Train Accuracy =  0.6730769230769231
```

Figure 2. The execution of the Logistic Regression Model

C. Decision Tree Algorithm

The decision tree is an algorithm that helps to build a classification and regression model. It generates a set of rules which used to categorize the data. It has the shape of a tree and offers a high level of precision and reliability. And it can deal with both numerical and categorical data.

```
In [31]:
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(
    labelCol='target',
    predictionCol='prediction',
    metricName='accuracy')
accuracy = evaluator.evaluate(dt_predictions)
print('Train Accuracy = ', accuracy)
DT_SC=accuracy*100

Train Accuracy =  0.8461538461538461
```

Figure 3. The execution of the Decision Tree Model

D. Random Forest Algorithm

The Random Forest (RF) classifier is another popular machine learning technique, which contains a large number of distinct decision trees. Each tree produces a class prediction in the RF and the class with the greatest quantity of poll takes place in the prediction of our model. It is very flexible and is capable of resolving any type of problem that can be classification or regression

```
In [38]: from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(
    labelCol='target',
    predictionCol='prediction',
    metricName='accuracy')
accuracy = evaluator.evaluate(rf_predictions)
model=accuracy
print('Train Accuracy = ', accuracy)
RF_SC=accuracy*100

Train Accuracy =  0.75
```

Figure 4. The execution of the Random Forest Model

VI. MODULE DESCRPTION

- Live Data collection
- MLlib
- Classification

A. Live Data Collection

Our Heart Disease Project datasets are collected from kaggle.com. Then we create an admin login. Admin can add live data to dataset. That data will be collected with the help of Xampp server. This is how we can collect live data.

Feature Information

1. age: The person's age in years
2. sex: The person's sex (1 = male, 0 = female)
3. cp: The pain intimate with (0 = typical angina, 1= atypical angina, 2= non-anginal pain, 3 = asymptomatic)
4. trestbps: The person's resting pressure level (mm Hg on admission to the hospital)
5. chol: The person's cholesterin measure in mg/dl
6. fbs: The person's fast blood glucose (> 120 mg/dl, 1 = true; 0 = false).
7. restecg: Resting medical instrument measure (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. thalach: The person's most pulse rate achieved
9. exang: Exercise evoked angina (1 = yes; 0 = no)
10. oldpeak: ST depression evoked by exercise relative to rest.
11. slope: the slope of the height exercise ST phase (0 = upsloping, 1 = flat, 2 = downsloping)
12. ca: the number of major vessels (0–4)
13. thal: A blood disease known as hypochromic anaemia (3 = normal; 6 = mounted defect; 7 = reversable defect)
14. target: heart condition (0 = no, 1 = yes)

```
In [19]: df
Out[19]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3.0	145.0	233.0	1.0	0.0	150.0	0.0	2.3	0.0	0.0	1	1
1	37	1	2.0	130.0	250.0	0.0	1.0	187.0	0.0	3.5	0.0	0.0	2	1
2	41	0	1.0	130.0	204.0	0.0	0.0	172.0	0.0	1.4	2.0	0.0	2	1
3	56	1	1.0	120.0	236.0	0.0	1.0	178.0	0.0	0.8	2.0	0.0	2	1
4	57	0	0.0	120.0	244.0	0.0	1.0	163.0	1.0	0.6	2.0	0.0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5	61	1	2.0	150.0	243.0	1.0	1.0	137.0	1.0	1.0	1.0	0.0	2	1
6	62	0	0.0	160.0	164.0	0.0	0.0	145.0	0.0	6.2	0.0	3.0	3	0
7	64	0	2.0	140.0	313.0	0.0	1.0	133.0	0.0	0.2	2.0	0.0	3	1
8	62	1	1.0	120.0	281.0	0.0	0.0	103.0	0.0	1.4	1.0	1.0	3	0
9	64	1	3.0	171.0	227.0	0.0	0.0	155.0	0.0	0.6	1.0	0.0	3	1

313 rows × 14 columns

Figure 5: Attributes name and values



Figure 6: Getting the input from user

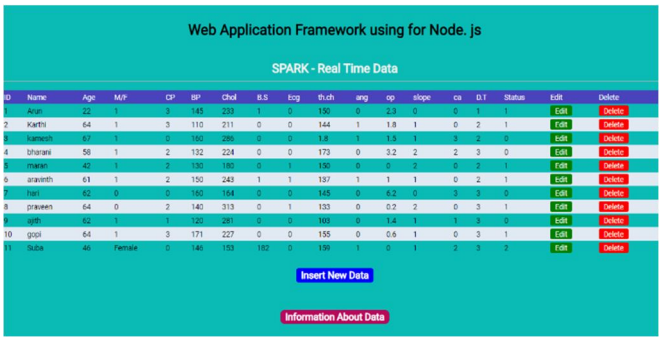


Figure 7: Real Time Data

B. MLLIB

MLlib is Spark's machine learning (ML) library. Its Aim is to make practical machine learning scalable and easy. Here we use Random Forest algorithm for prediction. Random Forest supports machine learning classification algorithm. Random forests generate a collection of decision trees. The Random Forest algorithm supports both binary and multiclass labels, as well as both continuous and categorical features.

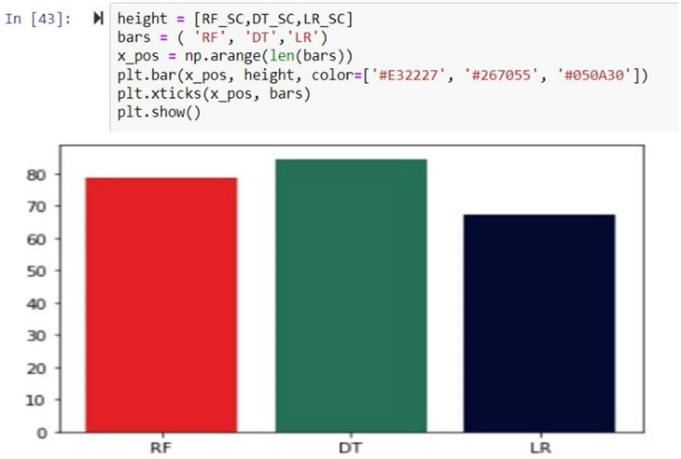


Figure 8: Comparison with accuracy values of all implemented models.

Red indicates random forest algorithm; DT stands for decision tree and LR denotes logistic regression. Among all three algorithms Decision tree is best.

C. Classification

Several standard performance metrics such as accuracy were considered to calculate the performance efficacy of this model. New data is trained, and user-supplied input goes into the trained dataset. After prediction estimate the given expected value as output on web application using flask.

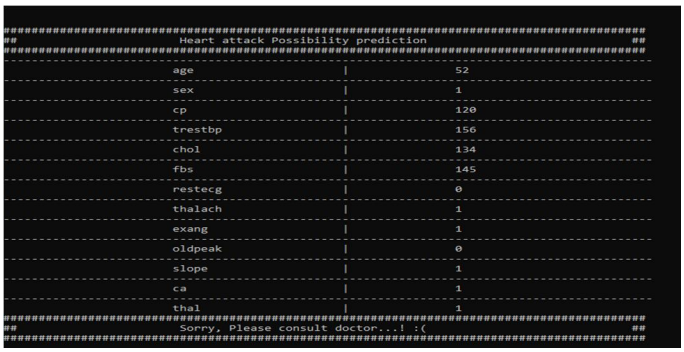


Figure 9: Output

XI. CONCLUSION

From this research work, we have built a predictive model for predicting Cardiovascular Disease, using PySpark and Spark MLlib libraries which are integrated with the Apache Spark framework. We tried to apply supervised algorithms that included decision tree classifier, logistic regression classifier, random forest classifier, and for evaluating the performance of the models. Included some stages for making a proposed framework such as loading cardiovascular diseases dataset, and pre-processing of the dataset, classifiers, and at the end evaluating classifiers. There will be a thorough study of huge datasets with more attributes to attain the highest accuracy in future work. Another future work is; researchers can use deep learning techniques to improve the performance of the model.

REFERENCERS

[1] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the World Congress on Engineering and computer
[2] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.
[3] Rajmohan, K., Paramasivam, I., & SathyaNarayan, S. (2014, February). Prediction and Diagnosis of Cardiovascular Disease--A Critical Survey. In 2014 World Congress on Computing and Communication Technologies (pp. 246-251). IEEE.
[4] Using machine learning algorithms in cardiovascular disease risk evaluation.
[5] Study of Intelligence Techniques for Cardiovascular Disease.
[6] Maini, E., Venkateswarlu, B., & Gupta, A. (2018, August). Applying machine learning algorithms to develop a universal cardiovascular disease prediction system. In International Conference on Intelligent Data Communication Technologies and Internet of Things Springer, Cham.
[7] Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access 2019
[8] Sultana, M.; Haider, A.; Uddin, M.S. Analysis of data mining techniques for heart disease prediction. In Proceedings of the 2016 3rd International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2016, Dhaka, Bangladesh, 22–24 September 2016
[9] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019
[10] Kodati, S.; Vivekanandam, R. Analysis of Heart Disease using in Data Mining Tools Orange and Weka Sri Satya Sai University Analy
[11] Heart Disease Dataset. Available online: https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed on 24 May 2021).
[12] Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms.
[13] Maini, E.; Venkateswarlu, B.; Maini, B.; Marwaha, D. Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. Med. J. Armed Forces India 2021
[14] Kang, K.; Michalak, J. Enhanced Version of AdaBoostM1 with J48 Tree Learning Method. [1802.03522] Enhanced Version of AdaBoostM1 with J48 Tree Learning Method. Available online: arxiv.org (accessed on 27 June 2021).

AUTHORS PROFILE



Mrs. R Dayana is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Jeppiaar Institute of Technology, Chennai. She has 7.10 years of Teaching experience. She was awarded Gold medal in her PG degree. She has handling various Computer science subjects like Compiler Design, Computer Networks, Programming in Data Structures, Object Oriented Programming, C Programming, Mobile Computing, Adhoc Sensor Networks, Database Management Systems, Adhoc and sensor networks. She has published 6 papers in Journals and also presented 12 papers in various National and International Conferences. She is doing her research in the field of Wireless Sensor Networks and Machine learning Techniques.



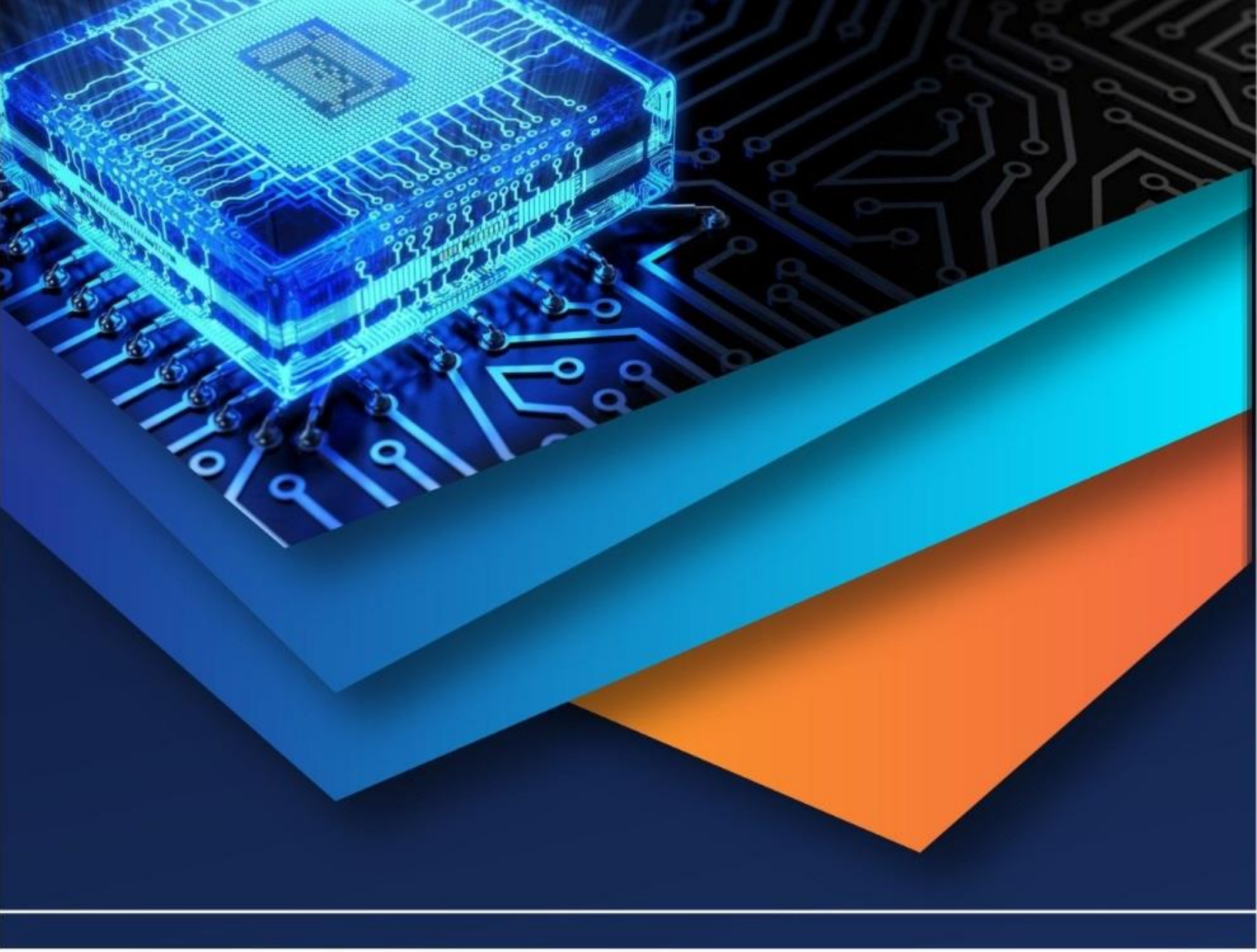
.**E.Bibilin Manuela** is currently pursuing her bachelor's degree in the field of Computer Science and Engineering at Jeppiaar Institute of Technology, Kanchipuram, Tamil Nadu, India. She did his schooling in Kanyakumari. She is particularly interested in Python and pyspark



J. Julie Christina is currently pursuing her bachelor's degree in the field of Computer Science and Engineering at Jeppiaar Institute of Technology, Kanchipuram, Tamil Nadu, India. She did his schooling in Chennai. She is particularly interested in Python and Machine Learning.



K. Keerthika is currently pursuing her bachelor's degree in the field of Computer Science and Engineering at Jeppiaar Institute of Technology, Kanchipuram, Tamil Nadu, India.She did his schooling in Chennai. She is particularly interested in Python and Machine Learning.



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)