iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** X **Month of publication:** October 2019

DOI: http://doi.org/10.22214/ijraset.2019.10005

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: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

An Improved Data Mining Model for Predicting Medical Issues

Shredha Parmar[1], Nikhil Chaturvedi[2]
[1]*Computer Science Engineering, RGPV*
[2]*Computer Science engineering, Shri Vaishnav Institute of Technology*

Abstract: ***The medical science is significantly usages the services of data mining and machine learning. In different domains of medical science the data mining techniques are helpful to research and planning. A number of applications are possible by involving the resources of another computing domain. In this context an association rule mining technique based prediction system is proposed. The association rules are developed on the basis of item sets frequencies. Therefore it is a slow process of rule generation. The proposed system works over the problem of speeding up the speed of association rule generation. Because, existing apriori algorithm consumes a significant amount of time and memory for generating candidate sets. Therefore to improve the speed of data processing we implemented divide and conquer strategy used with the existing apriori algorithm. Because the generation of most possible combinations of elements or item set is required. In the proposed strategy the small size of data input reduces the computation time. The presented work basically a data model for predicting medical disease according to the different kinds of datasets available, such as UCI repository based medical datasets such as Heart and Diabetes datasets. In this presented work both the datasets are used for experimentation. The obtained results show that the proposed apriori algorithm increases their accuracy and reduces the algorithm running time.***

Keywords: apriori algorithm, speeding up algorithm, implementation, results analysis, performance improvement, association rule mining.

I. INTRODUCTION

The data mining techniques enable us to analyze and recover the different patterns which can helpful for providing assistance in decision making, prediction, classification, categorization and many more. Therefore these techniques are widely accepted in various applications such as engineering, medical science and others. In this context the mathematical processes are developed in form of algorithms for processing the data. In this presented work the aim is to involve the machine learning and data mining techniques for analyzing the different medical datasets such as heart and diabetes for preparing a common platform or data model to recognize the attributes and can predict the class labels accurately. Therefore an essential data mining algorithm that is frequently used in different applications namely apriori algorithm is taken for further improvements and system design.

The apriori algorithm is an association rule mining algorithm which usages the items frequency and develop the association rules. But during experiments it is observed that is expensive for processing the large amount of data. Therefore some key improvements on the existing apriori algorithm is proposed and implemented. The improvements are focused on reducing the algorithms running time and space complexity. In addition of it is also tried to improve the predictive performance of the target algorithm. First the data encoding is involved which make enable the proposed technique to work with different kinds of datasets, secondly the partition based association rule generation that reduces the size of item set scanning and candidate set generation process. Therefore by involving the two additional process on the existing apriori algorithm the improved.

The core objective of the work is to apply the apriori based association rule mining technique for medical domain datasets. Thus some additional modifications are suggested for improving the existing approach of the apriori algorithm based rule mining. The following objectives are included for work.

1) *To Study And Explore The Domain Of Association Rule Mining:* in this phase the existing association rules mining algorithms are explored and understanding about the functional aspects are recognized. Additionally the recently contributed articles for improvements are also studied.
2) *To Design And Implement An Improved Association Rule Mining Algorithm:* in this phase a new data model by including existing approaches are prepared.
3) *To Evaluate And Compare The Performance Of The Proposed And Traditional Apriori Algorithm:* in this phase the proposed system's outcome is compared with the traditional apriori algorithm.

II. PROPOSED WORK

The proposed work is aimed to explore the data mining and machine learning techniques. These techniques are used for evaluation of health care datasets for predicting the possible diseases based on the target attributes and their patterns. This chapter provides the details about the proposed system and their functional aspects.

A. System Overview

The dramatic growth in the domain of heath care industry is found in recent years. Additionally new verticals in this domain are also appeared. The key motive of all the efforts is to understand the nature of diseases, and recovery of meaningful patterns that can help on finding solutions for these diseases. Therefore some techniques are required which can evaluate the large set of data and automatically produces the classification and predictive outcomes. In this context the data mining techniques are having the ability to analyze a large volume of data and prepare a mathematical model by which the similar patterns can be recognized or become predictable. Therefore the proposed work is intended to use the data mining tools and techniques for predicting the possible diseases based on their characteristics.

In this presented work an association rule mining technique is adopted for preparing solution. In literature we found a number of techniques by which we can recover the rules and perform classification and prediction. Additionally we found two popular algorithms i.e. FP-Tree and Apriori algorithm. But the association rule mining techniques works on the relativity of available attributes thus the pattern understanding becomes more transparent as compared to other techniques. Thus the apriori algorithms are taken into consideration and an improved version of apriori algorithm is crafted. The proposed improved apriori algorithm helps to improve the algorithms running time and also promising to improve the prediction accuracy. This section provides the overview of the proposed data mining system and next section provides the details about the formulated solution.

B. Methodology

The proposed data model for predicting health care issues in real world is demonstrated in figure 2.1 with the basic components which are used for recovering the target application patterns.

1) *Medical Dataset:* The machine learning and data mining techniques need some initial examples for preparing the data model and recognizing the similar target patterns which are learned using examples. In the similar manner our proposed data model also needs some sample datasets to train and test the prepared model. In this context the UCI repository is explored and two health care data sets are recovered namely heart diseases dataset and diabetes diseases dataset. Both the datasets are further used for experimentation and performance evaluation of the proposed system.

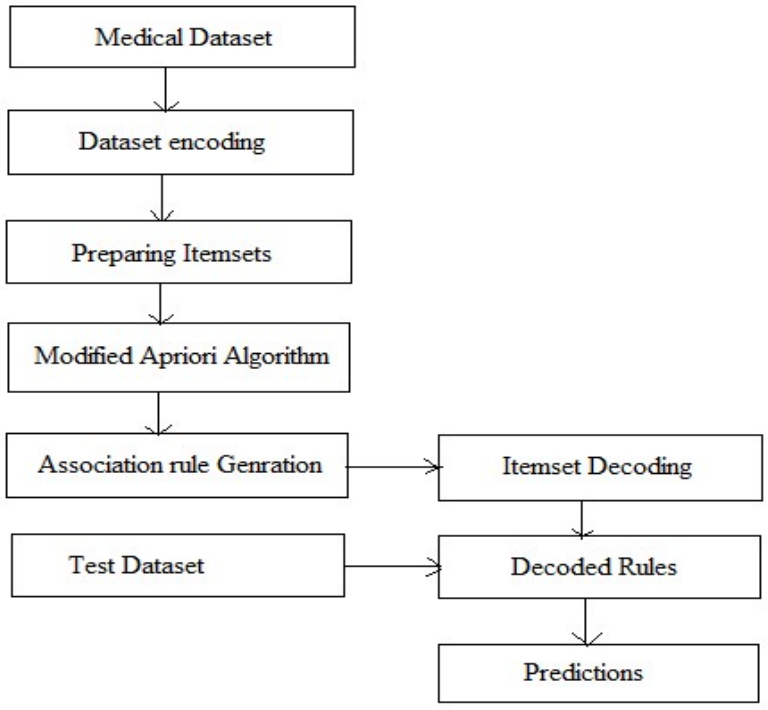


Figure 2.1 Proposed Data Model

2) *Dataset Encoding:* The aim of this module is to convert all the data into an intermediate format for easy data processing and rules recovery. The basic concept of this encoding is to converting the item sets into identifiable notations and reducing the efforts for evaluation of all the similar kinds of values. Therefore the following process is adopted as given in table I.

TABLE I
Data Encoding Process

Input: dataset D
Output: encoded dataset E

Process:

1. $[row, col] = readDataset(D)$
2. $max = getMaxAttribute(D)$
3. $for(i = 1; i \leq row; i++)$
    a. $for(j = 1; j \leq col; j++)$
        i. $if(D_{i,j}! = string)$
            1. $if(D_{i,j} \leq max * 0.2)$
                a. $E.Add(A)$
            2. $Else\ if(max * 20 > D_{i,j} \leq max * 40)$
                a. $E.Add(B)$
            3. $Else\ if(max * 40 > D_{i,j} \leq max * 60)$
                a. $E.Add(C)$
            4. $Else\ if(max * 60 > D_{i,j} \leq max * 80)$
                a. $E.Add(D)$
            5. $Else\ if(max * 80 > D_{i,j})$
                a. $E.Add(E)$
        ii. Else
            1. $E.Add(D_{i,j})$
        iii. End if
    b. End for
4. End for
5. Return E

3) *Preparing Item Sets:* After encoding of the dataset into symbolic format the most of the numerical attributes are transformed into a label or text. Therefore all the unique values from the attributes available is selected as the item set and by using these attributes the developed data instance is treated as individual transaction set.

4) *Modified Apriori Algorithm:* The traditional apriori algorithm is explained in previous chapter. According to the described algorithm the algorithm basically first scans the items and generates candidate set. Basically candidate set generation process generate all the possible combinations of item sets that are feasible according to the given transaction set. Therefore more number of items and transactions means large amount of computational resource usages. Therefore here we involved dataset splitting technique to create partition of the dataset and produce a number of subsets of given data. That partitioning is performed on the basis of data instance class labels. The less number of transactions multiply the speed of apriori algorithm. Therefore the process given in table II is used for categorizing the data according to class labels given in the dataset. The given algorithm usages all the attributes or instances and just check the available class labels. If the class label list is already created then assign the instance to the existing group otherwise we create and add a number class labels group list.

Table II. Data Partitioning

Input: encoded dataset E
Output: Clustered data according to class labels C

Process:

1. $E_n = readDataset(E)$
2. $for(i = 1; i \leq n; i++)$
   a. $if(E_i == NewClass)$
      i. $Create\ C_j\ where\ j = 1,2,\ldots,m$
      ii. $C_j.Add(E_i)$
   b. Else
      i. $C_{j=1,2,\ldots}.Add(E_i)$
   c. End if
3. End for
4. Return C = $C_{j=1,2,\ldots}$

5) *Association Rule Generation:* After the processing the data according to the process given in table II. The classical apriori algorithm is applied on the generated list of transactions. This list of data is used for generation of association rule by using the classical apriori algorithm.
6) *Items Decoding:* The generated rules from the previous phase are used with the recovery of actual values therefore a reverse mapping process is adopted to prepare relevant data over the rules. Finally these rules are used for further classification and prediction of the class labels.
7) *Decoded Rules:* When the data items are recovered for their actual values then the generated rules are also transformed into the suitable attribute values based rules are also recovered.
8) *Test Dataset:* The rule generation and the rule decoding is the final step of training process of the proposed system. Finally the test dataset is applied on the rules for generation of class labels. The test dataset is basically the part of existing dataset which prepared by selecting the random instance from the entire dataset. In this experimentation the 70% of randomly selected data is used for training and the 30% of dataset is used for testing of the prepared data model.
9) *Predictions:* The applied rules on the test dataset help to recognize the patterns similar which are learned during the training session of the algorithm.

C. Proposed Algorithm

This section summarizes the entire process steps in terms of algorithm steps, thus the input and output of the system is discussed in this section.

Table III: Proposed Algorithm

Input: Training Dataset D, Test Dataset T
Output: class label of test data instances C

Process:

1. $D_n = ReadDataset(D)$
2. $for(i = 1; i \leq n; i++)$
   a. $E = D_i.Encode$
3. End for
4. $R_m = ImproveApriori.GenrateRule(E)$
5. $for(j = 1; j \leq m; j++)$
   a. $R_j = R_j.Decode$
6. End for
7. $T_w = ReadDataset(T)$
8. $for(k = 1; k \leq w; k++)$
   a. $C = R_m.classify(T_k)$
9. $end\ for$
10. Return C

III.RESULT ANALYSIS

The performance of the implemented predictive data model is described in this chapter. Therefore the used parameters and their understanding is provided is chapter.

A. Accuracy

Accuracy is measurement of correctness of a data mining and machine learning algorithm. It is a ratio between total correct recognized samples and total samples produced for recognition. Therefore it is measured as:

$$accuracy\,(\%) = \frac{Correctly\ identified\ samples}{total\ samples\ to\ identify} X100$$

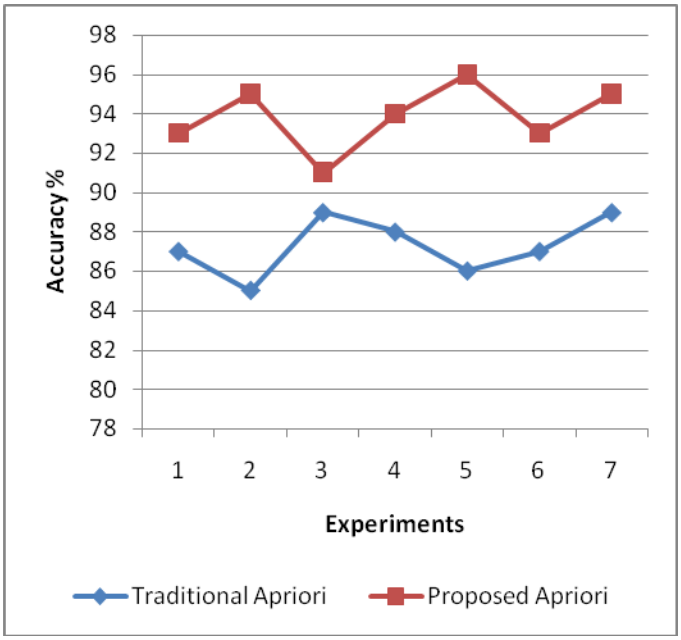


Figure 3.1 Comparative Accuracy

The accuracy of the rule based classification and prediction algorithm for both traditional and proposed algorithm is described in figure 3.1 as the line graph. That line graph is made using the observations collected in table IV. To represent the performance of both the algorithms X axis includes the number of experiments conducted and Y axis shows the percentage accuracy achieved in percentage. According to the calculated results the proposed technique of apriori algorithm improves the classification accuracy with respect to the classical apriori algorithm.

Table IV
Accuracy In Percentage

Experiments	Traditional Apriori	Proposed Apriori
1	87	93
2	85	95
3	89	91
4	88	94
5	86	96
6	87	93
7	89	95

B. Error Rate

The error rate of a system shows the rate of misclassification for an algorithm. That is also defined as a ratio between inaccurately classified samples and total samples to evaluated. The following way is used for computation of error rate:

$$error\ rate\ (\%) = \frac{inaccurate\ classification}{total\ samples} X100$$

Or

$$error\ rate\ (\%) = 100 - accuracy$$

TABLE V. Comparative Error Rate (%)

Experiments	Traditional Apriori	Proposed Apriori
1	13	7
2	15	5
3	11	9
4	12	6
5	14	4
6	13	7
7	11	5

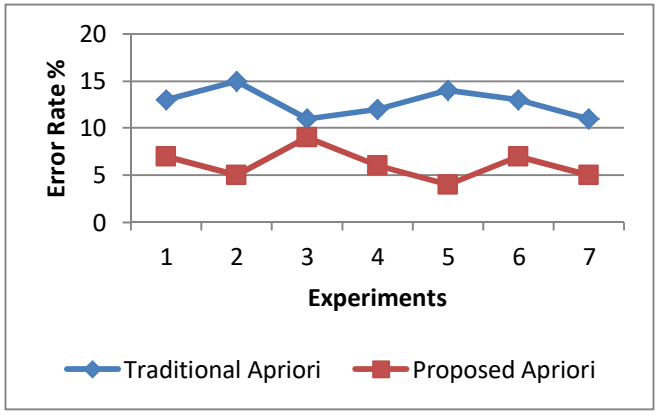


Figure 3.2 Comparative Error Rate (%)

Percentage comparative error rate for both the predictive algorithm is reported in figure 3.2. That is a line graph and prepared by using the values available in table V. To represent the error rate in different experiment the X axis shows the number of experiments performed with the system. Additionally the Y axis represents the percentage error rate of the system. The described results show the error rate of the proposed apriori algorithm is less than the traditional apriori algorithm for predicting accurate values.

C. Time Consumption

The processing of data and generation of outcomes required an amount of time. This time requirement is known as the time consumption of the proposed algorithm. In a java based implementation the following method is used for finding time difference.

$$time\ consumed = algorithm\ end\ time - start\ time$$

TABLE VI. Time Consumption In (MS)

Experiments	Traditional Apriori	Proposed Apriori
1	738	472
2	681	439
3	623	513
4	714	461
5	645	461
6	693	351
7	716	431

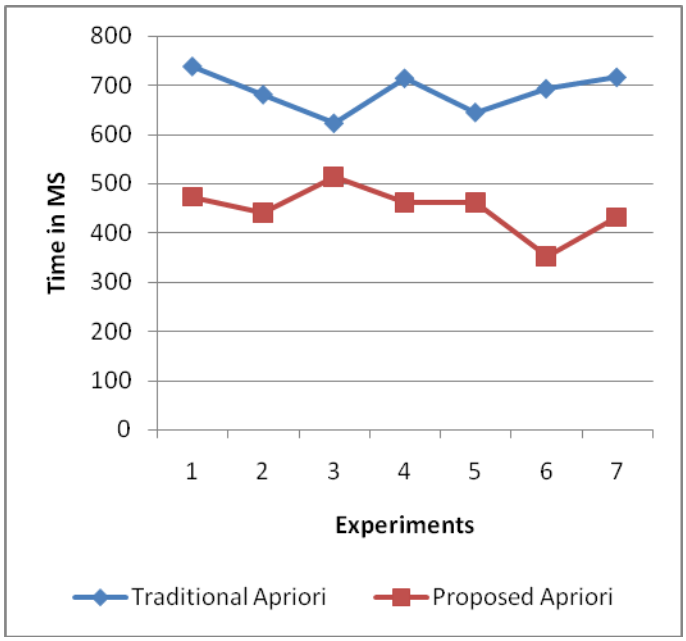


Figure 3.3 Time Consumption In (MS)

The comparative time requirements of both the algorithms are measured in milliseconds (MS). The improved apriori algorithms performance is denoted in red colors additionally the traditional apriori algorithm is defined in blue color. The reported values in table VI are the time consumption of both the methods which is represented using line graphs as given figure 3.3. The results clearly shows the performance of the proposed work is efficient than the traditional approach. Thus model is time preserving model by reducing the time during the scanning of the transactions.

D. Memory Usage

The processes required an fixed size of space in the main memory for hosting the data and instructions. These memory requirements are known as the space complexity of algorithm. To compute the java based process memory usages the following function can be used.

$$memory\ usages = total\ assigned\ memory - free\ memory$$

Table VII. Memory Usages IN KB

Experiments	Traditional Apriori	Proposed Apriori
1	27381	24726
2	26817	24391
3	28462	25135
4	27714	24611
5	27454	24615
6	26937	23519
7	27716	24317

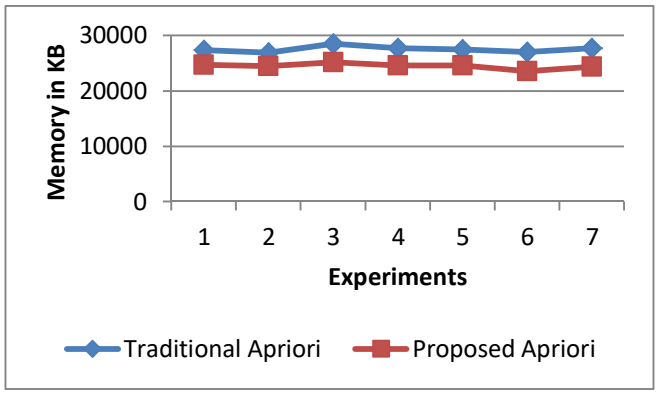


Figure 3.4 Memory Usages In KB

The memory usages of both the systems namely improved apriori algorithm and traditional apriori algorithm is explained here. The table VII contains the experimental observations of the implemented algorithms, these values of experiments are used with the figure 3.4 for representing the performance in terms of line graph. The figure contains the conduced experiments in X axis and the memory usages in Y axis. The memory is measured here in terms of milliseconds (MS). According to the results the proposed technique requires less amount of main memory as compared to traditional apriori algorithm.

IV. CONCLUSIONS

The chapter provides the conclusion of the work described in this thesis. Therefore the experimental as well as observations are used for providing the conclusion of the work. Additionally based on the obtained facts the future extension is also suggested.

A. Conclusion

The significant amount of development and research is conducted in the domain of medical and health care industries. Additionally the similar growth on technology and automatic data processing techniques are also observed. The health care industry generates a significant amount of data in structured and unstructured formats. Both the kinds of data is essential for research and supporting the human lives. In this presented work the aim is to employ the data mining model over the bench mark datasets and explore the possibility to predict early possible diseases based on the historical data analysis. Therefore a common data model is developed for including the datasets heart and diabetes. The proposed data model first transform the entire available attributes over a common data format and then apply the data mining algorithm for mining the rules. the data developed rules are used for classifying and predicting the possible diseases according to the input attributes.

The proposed data model is a data mining technique that accepts different medical dataset and develops rules for classify and predict the target disease. First the preprocessing technique is applied on data and then the technique involves an encoding process for minimizing the efforts of items scanning. Here based on the values range the encoding is performed. Additionally different number of subgroups is also created that optimizes the running time of the algorithm. Thus each sub group of data is processed individually and then combined to generate the rules. Finally for comparing the performance of the modified apriori algorithm the traditional apriori algorithm is also implemented and compared with the help of different performance parameters.

The implementation of the proposed system for predicting the health issues is given using JAVA technology. Additionally to store the computed performance the MySql server (database) is used. The performance of the proposed system is computed in terms of accuracy, error rate, memory usages and time complexity. The summary of performance is demonstrated in table VIII.

TABLE VIII. Performance Summary

S. No.	Parameters	Traditional apriori	Proposed apriori
1	Accuracy	85-89 %	91 – 96 %
2	Error rate	11 – 15 %	4 – 9 %
3	Memory usages	26817 – 28462 KB	23519 – 25135 KB
4	Time complexity	738 – 623 MS	352 – 472 MS

According to the obtained results the proposed method of apriori algorithm improves the performance of system. Additionally the approach is much promising for work with the different datasets.

B. Future Work

The main aim of the proposed work is to improve the existing apriori algorithm, additionally preparing a data model that can work with the different medical dataset for producing the effective outcomes. The following extension of the work is proposed for future:

1) The proposed technique is effective to work with the different dataset with the similar methodology in near future the real time data extraction and mining is proposed for work.

2) The existing technique is works on the basis of rule based classification technique that is sometimes expensive for computations in near future the opaque data model is suggested to work.

REFERENCES

[1] A. Mittal, A. Nagar, K. Gupta, R. Nahar, "Comparative Study of Various Frequent Pattern Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2015

[2] V. Rajalakshmi and G. S. Anandha Mala, "Anonymization by Data Relocation using Sub-clustering for Privacy Preserving Data Mining", Indian Journal of Science and Technology, Vol 7(7), 975-980, July 2014

[3] L. Li, R. Lu, K. K. R. Choo, A. Datta, and J. Shao, "Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases", DOI 10.1109/TIFS.2016.2561241, IEEE, Transactions on Information Forensics and Security

[4] K. M. S. N. Malli, S. Bezawada, V. Vijayan, "Analysis of Association Mining through Enhanced Apriori Algorithm", International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 56-60 (2013)

[5] A. Bhandari, A. Gupta, D. Das, "Improvised Apriori Algorithm Using Frequent Pattern Tree for Real Time Applications in Data Mining", Procedia Computer Science, Volume 46, 2015, Pages 644-651

[6] C. S. Dangare, and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques", International Journal of Computer Applications 47.10 (2012): pp. 44-48.

[7] S. Divya Meena and M. Revathi, "Predictive Analytics on Healthcare: A Survey", International Journal of Science and Research (IJSR), Volume 4 Issue 9, September 2015.

[8] C. C. Aggarwal, P. S. Yu, "Online Generation of Association Rules", ICDE Conference, 1998

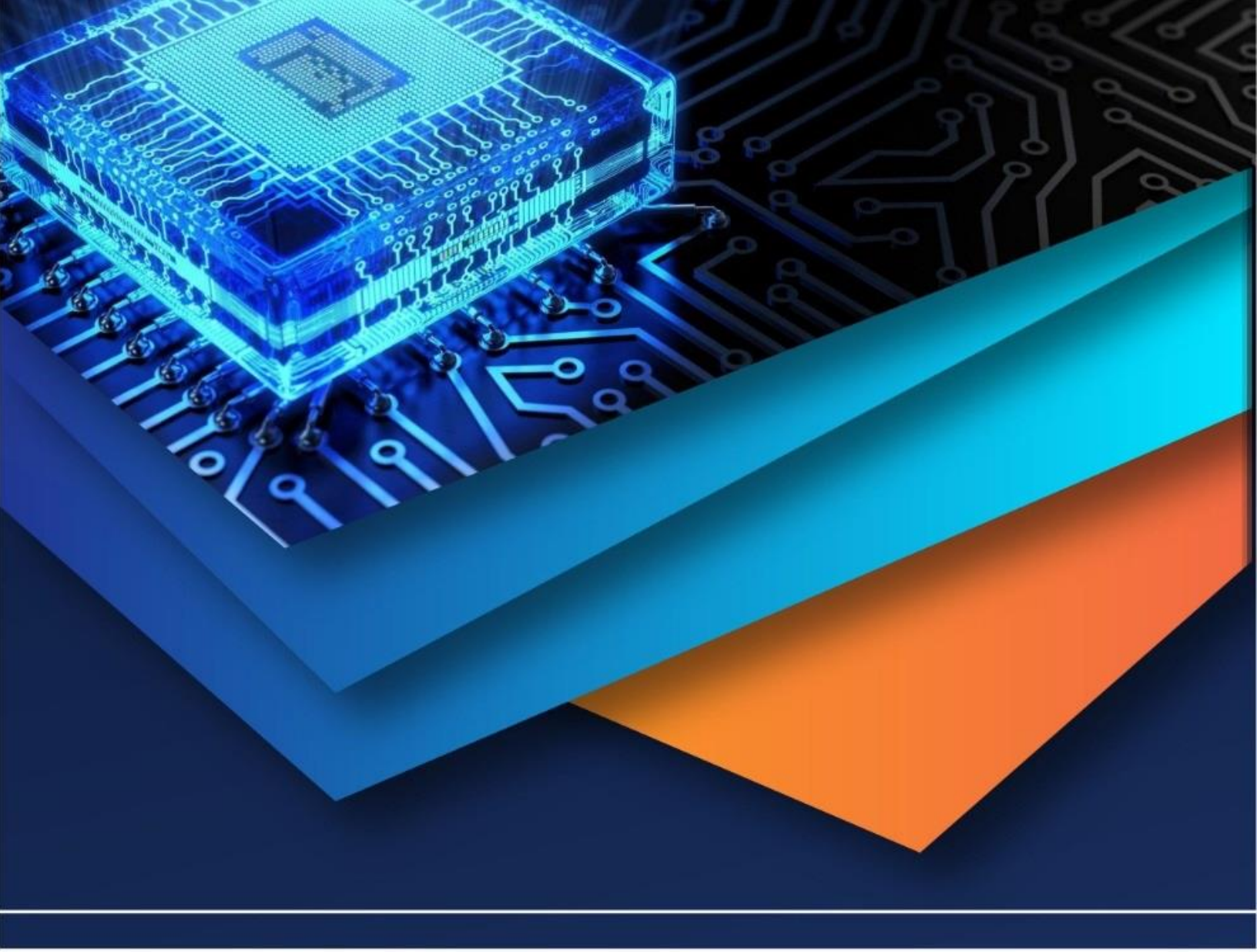
[9] N.G.B. Amma, "Cardio Vascular Disease Prediction System using Genetic Algorithm and Neural Network", IEEE International Conference on Computing Communication and Applications 2012.

[10] H. J. Li, A. Plank, H. Wang and G. Daggard, "A Comparative Study of Classification Methods for Microarray Data Analysis", Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, (2006), pp. 33-37.

[11] H. Haripriya, Prathibhamol Cp, Y. R. Pai, M. S. Sandeep, "Multi Label Prediction Using Association Rule Generation and Simple k-Means", 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 978-1-5090-0082-1/16/$31.00 ©2016 IEEE

[12] Z. A. Rana, M. M. Awais, S. Shamail, "Improving Recall of Software Defect Prediction Models using Association Mining", Knowledge-based Systems 00 (2015) 1

[13] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, Z. Abbas, "A model for early prediction of diabetes", Informatics in Medicine Unlocked 16 (2019) 100204



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY