iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4125

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: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Machine Learning Approach to Predict Student Academic Performance

Amandeep Kaur[1], Nitin Umesh[2], Barjinder Singh[3]
[1, 2, 3]*Department of Computer Science and Engineering, LPU, Phagwara*

Abstract: ***In today's world, Data mining in the field of education is utmost important to do predict the performance of student. Education is the Power and by predicting performance in education by considering relevant parameters we would be able to work on the weaknesses of student at right time by using right pedagogies and approaches. Different aspects are evaluated like social, economic, personal, cultural, geographical, institute environment and other in education research study. Such aspects may either help a student in shining during academic period or halt academic program. Such failure is known as drop-out. Data mining algorithm helps in finding those factors; that are mostly contributing the student's performance. In our research we are going to use the machine learning approach or machine learning classification models to predict student academic performance and that can fit in Educational data mining. In our research study, we would like to predict performance and compare the results of these prediction models . In this research, we would also discuss how a these machine learning models can help to improve an education system by considering the different factors in terms of accuracy, Specificity, Precision, Prevalence, Recall, F-Measure etc in results. In addition to that, this research is carried on by applying Standard and Hybrid approaches as well as fuzzy logic and compared outcomes. This research would also help in predicting the performance of other state student by identifying the appropriate factors. The main aim of this research is to predict the performance of student and to identify the most appropriate model for that. So that on the basis of that we would be able to take the appropriate steps to improve their performance and to cut down the dropout rate.***

Keywords: Classification, Prediction, EDM, KDD, Machine Learning, SVM, Decision Tree, Naive Bayes, Logistic Regression, LMT, Fuzzy Logic

I. INTRODUCTION

Data Mining is the process to find the hidden information as well as pattern from a bulk amount of data i.e. the data should be coming from different sources such as data warehouse, data mart etc.[1] Data is hidden taken out through techniques of data mining. It gives imperative information that is important to take appropriate decisions. Pre-processing of information contain information cleaning to decrease noise, pertinence investigation to eliminate unessential qualities, forecast accuracy, scalability and interpretability. [2] Data mining is process of extracting knowledge from large amount of data. The main reason for what data mining algorithms are used is that it gather relevant information which provides us better outcomes. Information mining apparatus is utilized to discover questions and relations between them.[3] This technique incorporate measurable and additionally numerical model. Data mining process is performed on gathered data which is represented in different forms like text form, web, image processing and visuals. The very important step is to find knowledge from data by using Knowledge discovery process. It includes various steps for extracting meaningful data. Data mining is concerned with more than one areas such as database management system, statistics, visualization etc. It merges techniques from so many fields such as image processing, ecommerce, retail, pattern mining etc. Data mining system consist of operational units for tasks such as association, classification, prediction and clustering data analysis. [4] Data Mining is extensively useful in Educational Data Mining. EDM is a rising field for knowledge learning finding from vast measure of Educational data. The purpose of EDM is to find the pattern of educational data so that qualification of education can be improved. EDM is the educational research study of Variety of methods in which different aspects are evaluated like social, economic, personal, cultural, geographical, institute environment and other. Such angles may either help an understudy in exceeding expectations during scholarly period or end scholastic program of an understudy. Such disappointment is known as drop-out. [5]

II. DATA MINING ALGORITHMS

Data mining algorithm helps in finding those factors, that mostly contributing the student's performance. If we work on most contributing attribute better results can be achieved.

1) *Association rule algorithm:* It mainly deals with search statistical relations between objects in dataset. It finds how events aggregate together.
2) *Classification algorithm:* It can describe or classify objects related to dataset into predefined set of classes. It is supervised learning approach. It includes objects in dataset used to understand existing objects and predict behaviour of new objects. For instance Naive Bayes, SVM, Decision Tree, KNN etc. [6]
3) *Clustering algorithm:* It is collection of objects of similar type in one group. The cluster provides us better results. Clustering analysis has been a developing exploration issue in information mining due its assortment of uses. For instance K-means clustering, DBSCAN etc. [7]
4) *Machine Learning:* Both data mining and machine learning used same methods. But there is difference, machine learning focused on prediction, based on known properties, whereas data mining focuses on identification of unknown properties. SVM is machine learning technique to build a linear binary classifier. It defines the decision boundary between two classes. [8]
5) *Fuzzy logic:* It is a method to determine the "degree of facts" instead of the general "true or false" (1 or 0). Data mining uses different methods and assumption from a broad areas or fields for the knowledge extraction from huge amount of data. But uncertainty is a general phenomenon in data mining problems. Therefore, it is applied to manage with the uncertainty in actual world.[16]

III. KNOWLEDGE DISCOVERY PROCESS

Data mining is a procedure of eliciting or mining knowledge from enormous amount of data. It means knowledge extraction, knowledge mining of data, pattern analysis and data knowledge discover from data. It is the process of discovering required knowledge from database. It includes various operations such as selection, processing, transformation, interpretation and evaluation. Knowledge Discovery Process is abbreviated as KDD. [6] [7]

There are various steps to discover knowledge. It selects a dataset or its subset. It removes noise from data.

1) *Data cleaning:* It is the process of removing noise and inconsistent data. It can fill missing values. It is a first step in which dirty data and inconsistent facts or data are eliminated or discarded.
2) Data Integration: It can combine multiple sources in data warehouse. It includes multiple database, data cubes and files. Redundancy is duplication of data. It is removed by correlation analysis.
3) *Data selection:* It can retrieve data from database which is required for analysis. It can describe how to select various attributes. [9]
4) *Data Transformation:* In data transformation, information is change into forms fitting for mining. It can include different advances:-

a) *Smoothing:* It helps us to remove noise from data.

b) *Aggregation:* Data aggregation is a process of gathering information and expressed in a summary form such as statistical analysis. A common purpose of aggregation is to get more information about particular groups based on specific variables such as age, income.

c) *Generalization of data:* In generalization there is replacing of low level data to high level concepts through use of concept hierarchies. [10] [11]

5) *Pattern evaluation:* It can identify those patterns which represent knowledge based on some measures. Data mining is a procedure of taking out knowledge from big data repositories or databases. It can evaluate results in form of patterns. The large amount of knowledge is collected from different knowledge engineers. [12]

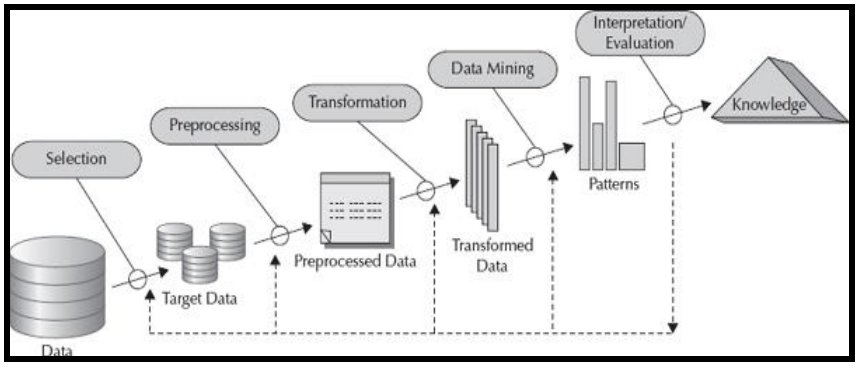


Figure1: Knowledge Discovery Process

IV. LITERATURE REVIEW

During my literature review study, to define the research problem as well as objectives, I would like to mention that I went through a lot of research papers and web application statistics. Few of paper's summarized details are as follows.

In this first research study, author used feature selection technique to reduce the number of feature form the large attribute set. In this paper author use ASSISTments platform dataset which is a web based teaching system developed at Worcester Polytechnic institute and used with 4^{th} to 10^{th} grade math students. In this paper author used technique to remove irrelevant, redundant or noisy data. In this paper author used various classification algorithm and ranker algorithm to find top most contributed attribute and removed the less appropriate attribute. This helps to speeds up the process of data mining and improves its performance parameters such as predictive accuracy. [1] In this research paper author used three different approaches. Cross tabulation analysis, Feature selection and balancing imbalance data. Features selection method is used to select those attribute which are highly affected dependent variables. Classification tree is built considering all available attributes. This method finds out all possible splits that can occur for each indicator variable at each node. The search stops when the split with the largest imprudent in goodness of fit is found. A few element choice calculations are connected and includes positioning higher in numerous calculations are chosen. In this way 15 vital parameter are chosen from unique 77 attributes. Misbalancing issue is resolved by using data balancing and rebalancing algorithm specifically SMOTE(Synthetic Minority Over sampling technique). Ten fold cross validation is used for establishing training and testing data from original data. This data set is prepared in three categories. First category contains data with all 77 attributes. Next category contains data with 15 important attributes. Last category contains balanced data after applying rebalancing technique in weak. [2] This paper provides the hybrid approach for outlier detection. They used two algorithms: K-mean and Neural Network. The proposed method use Integrating Semantic Knowledge (SOF- Semantic outlier factor) for outlier detection. This method detects the semantic outlier. This technique identifies the semantic anomaly. Semantic exception is an information point that acts uniquely in contrast to other information focuses in a similar class or same bunch or cluster. The main motive of this research was to reduce the number of outliers in clusters as well as data by improving the cluster formulation methods so that outlier rate reduces. It also decreases the error and improves the accuracy. The result showed that the hybrid algorithm performs better than that of genetic k-means. This proposed strategy manages content and date dataset that has not been executed before using genetic k-means. [3] This research study describes the various approaches such as Neural network, K- Nearest Neighbour, Bayesian Classifier, Fuzzy Logic and decision tree classification Algorithms for implementation of intrusion Detection system. With the help of this paper, it is clear that the data mining methods are used to perform the intrusion detection system But this paper don't describe which technique is best for all of these. [4]

V. PROBLEM FORMULATION

This study will analyse and predict student's future performance on the basis of their academic records as well as other major factors effecting it, which is extremely important for effectively carrying out necessary pedagogical approaches as well as to emphasize on student's weaker zones. This would also help in curriculum design, education policy design as well as in placements strategies. To sum up, this research is about analysing and predicting the performance of student by applying various Standard machine learning classification models on collected dataset and compared it with Hybrid machine learning approach using data mining and machine learning tools.

VI. RESEARCH METHODOLOGY

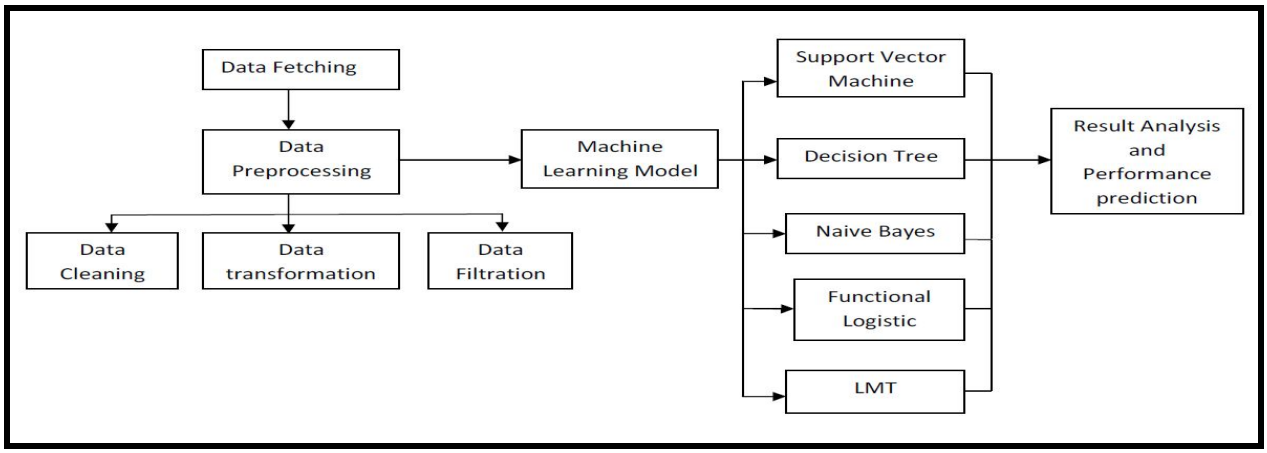


Figure2: Research methodology

Our input file consists of 1735 instances and 37 attributes. These are the results of student of BTech Second Year. We can remove the unique attribute from dataset, as they does not affect our classification. During First phase we need to do preprocessing of data. In filtration we can do perform data cleaning, transformation and data Filtration. Eliminate the NULL values by using filtration.
Next we have to choose J48 tree, LibSVM, FL and NaiveBayes Standard Classification algorithms as well as Hybrid approach LMT for better classification. Further, We have to choose training set and do focus on Result attribute(nominal attribute) and start classification. Visualize J48 tree and here we can see all the rules.
Now test J48 rules on another file, say test file. All the attributes in that file are same as of training set but Result attribute is empty here, which we need to predict. Now do classification after inserting test file.
Furthermore, visualize classifier errors and save the model. Now either you can do execute Java code to do obtain the predicted results or we can extract the predicted data by importing it in Microsoft Excel.
Moreover, Fuzzy Logic is used to predict students performance in terms of result and expected Dropout rate. A fuzzy system is an information-based rule system. The core of this system is a database which is configured with if-then rules. Fuzzy inference is the procedure of expressing input/output mapping using by applying the fuzzy logic. It tries to conclude answers from a knowledgebase by utilizing a fuzzy inference engine. The inference engine which is examined to be the brain of the expert systems gives the methodologies to reasoning around the data in the knowledgebase and explains the results. [16]
Following parameters are supposed to affect the academic performance of student. It would be very interesting to carry on our research by considering most of them. [13] [14]

Table I. Factors affecting performance

Variable	Description	Possible Values
CA	Continue assessment	{First Second Third Fail}
MTE	Mid-term marks	{First Second Third Fail}
ETE	End term marks	{First Second Third Fail}
ATT	Attendance	{Poor , Average, Good}
HW	Assignment/Home Work	{Yes, No}
LW	Lab work	{Yes, No}
FHS	Feel home sick	{Yes, No}
CS	Communication skills	{Poor , Average, Good}
CG	communication gap	{Yes, No}
LC	low confidence	{Yes, No}
MDP	more dependence on ppts	{Yes, No}
LTR	lack of text book reading	{Yes, No}
PLS	poor listening skills	{Yes, No}
PRS	poor reading skills	{Yes, No}
PWS	Poor writing skills	{Yes, No}
G	Gender	M or F
FE	father's education	{Poor , Average, Good}
ME	mother's education	{Poor , Average, Good}
MJ	mother's job	{Yes, No}
FJ	father's job	{Yes, No}
ACC	Accompany(Friend circle)	{Low, Medium, High}
LM	lack of maturity	{Yes, No}
LP	lack of patience	{Yes, No}
UC	uncertainties	{Yes, No}
LC	lack of concentration (focus)	{Yes, No}
BHA	bad habits	{Yes, No}
BHE	bad health	{Yes, No}
LOC	lack of consciousness, alert	{Yes, No}
BEH	bad eating habits	{Yes, No}
ID	Indiscipline	{Yes, No}
LCR	lack of creativity (innovation)	{Yes, No}
CISYS	complexities in system (ums etc)	{Yes, No}
LCG	lack of counselling	{Yes, No}
IF	infrastructural facilities	{Yes, No}
QE	quality of education	{Poor , Average, Good}
EO	employment opportunities	{Low, Medium, High}
MISC	participation in science fairs, quiz's, competitive exams, MOOC's	{Yes, No}

VII. RESULTS AND DISCUSSION

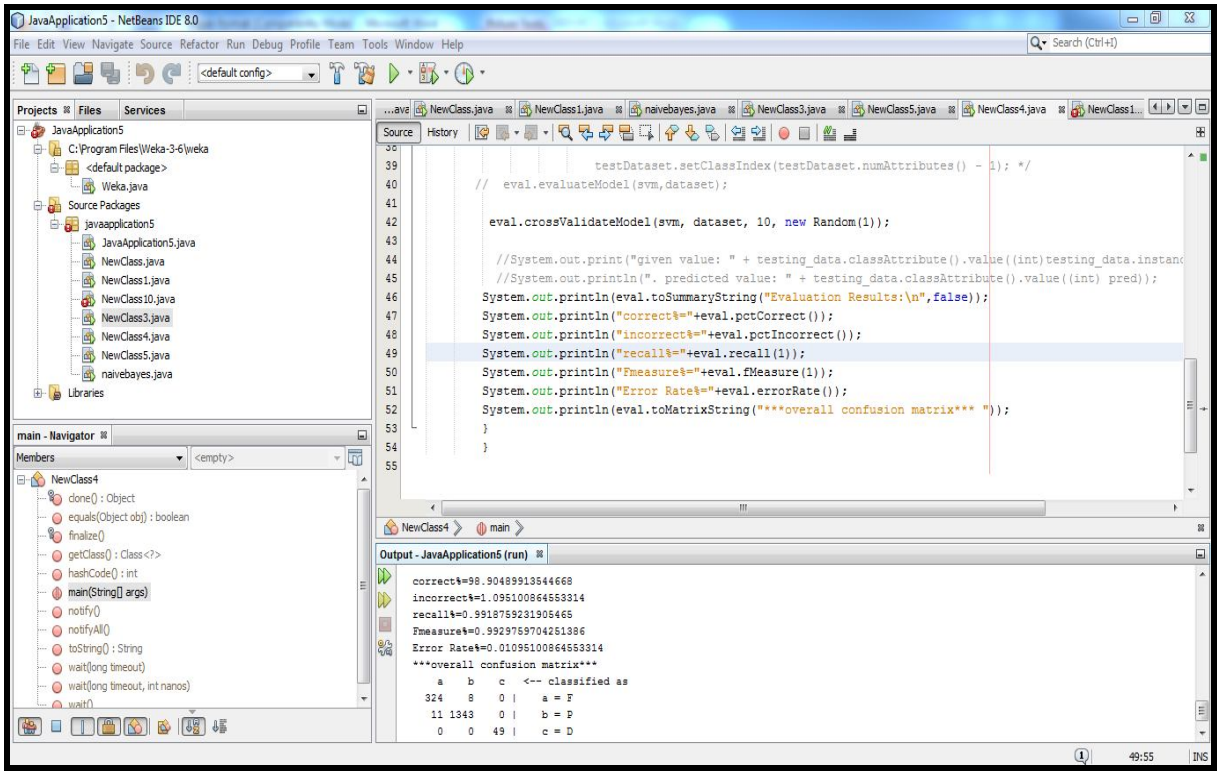


Figure3: SVM approach implementation in Java

The above Figure depicts the implementation of Standard and Hybrid approaches of machine learning to predict the performance of student on collected dataset.

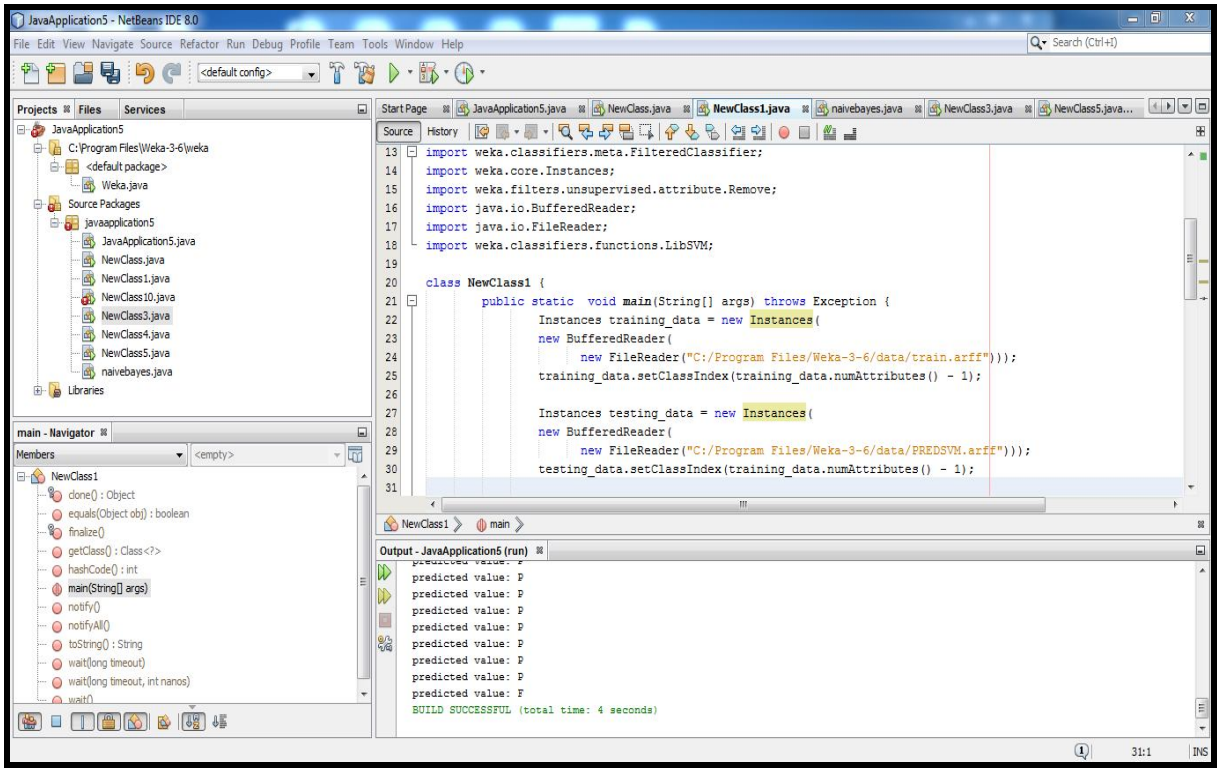


Figure4: Machine Learning approach for prediction in Java

The above Figure shows the predicted results in Net Beans Weka Java tool. To predict the performance and to obtain the accuracy of results, its compared with trained dataset.

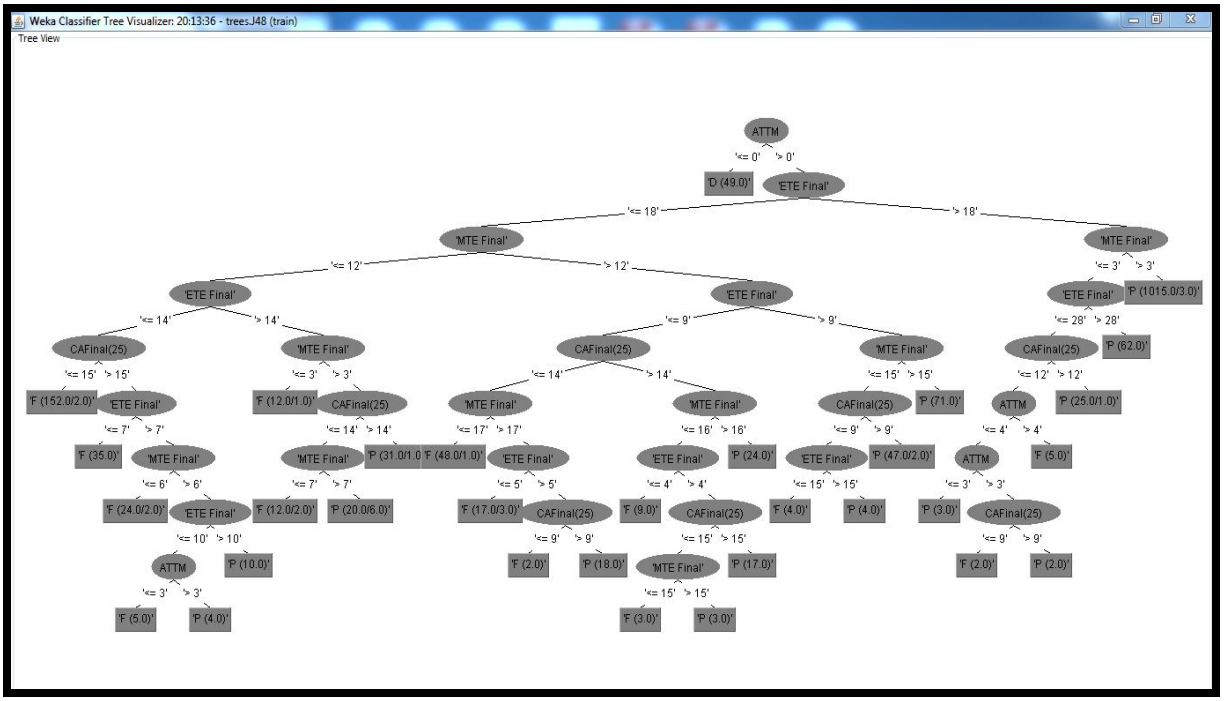


Figure5: C4.5 decision tree visualize all the rules

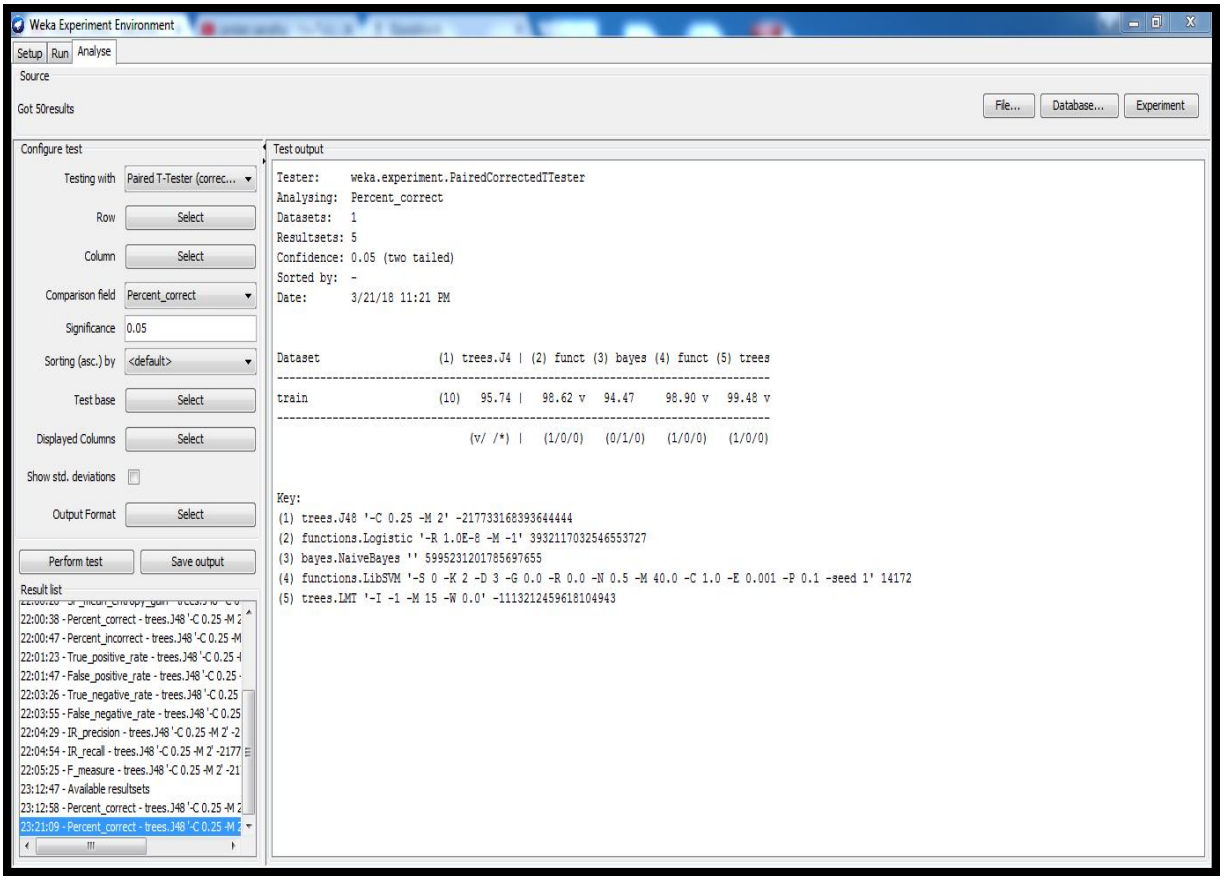


Figure6: Machine Learning approaches comparison in Weka Experimenter

The above Figure depicts the result comparison of various standard and Hybrid approaches in Weka Experimenter.

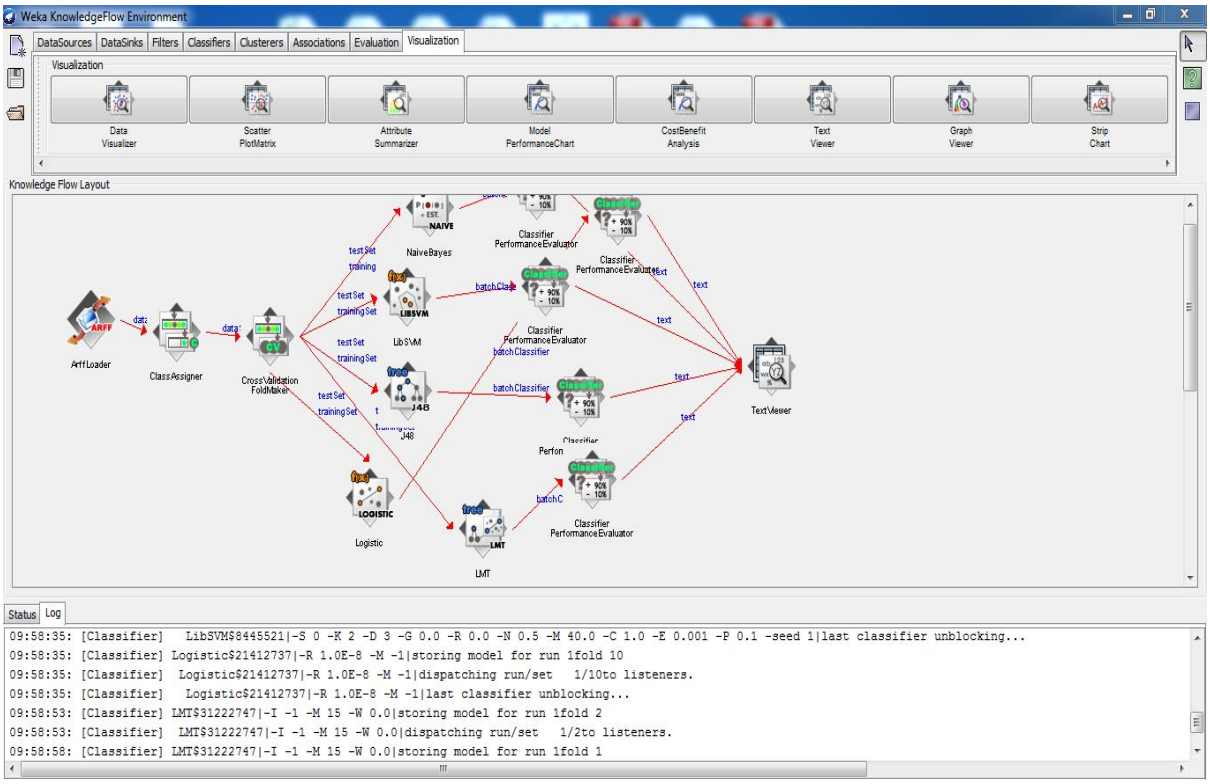


Figure7: Machine Learning approaches implementation through Knowledge Flow

The above figure depicts implementation by applying standard machine learning approaches C4.5, Naive Bayes, LibSVM, Logistic Regression and Hybrid approach LMT.

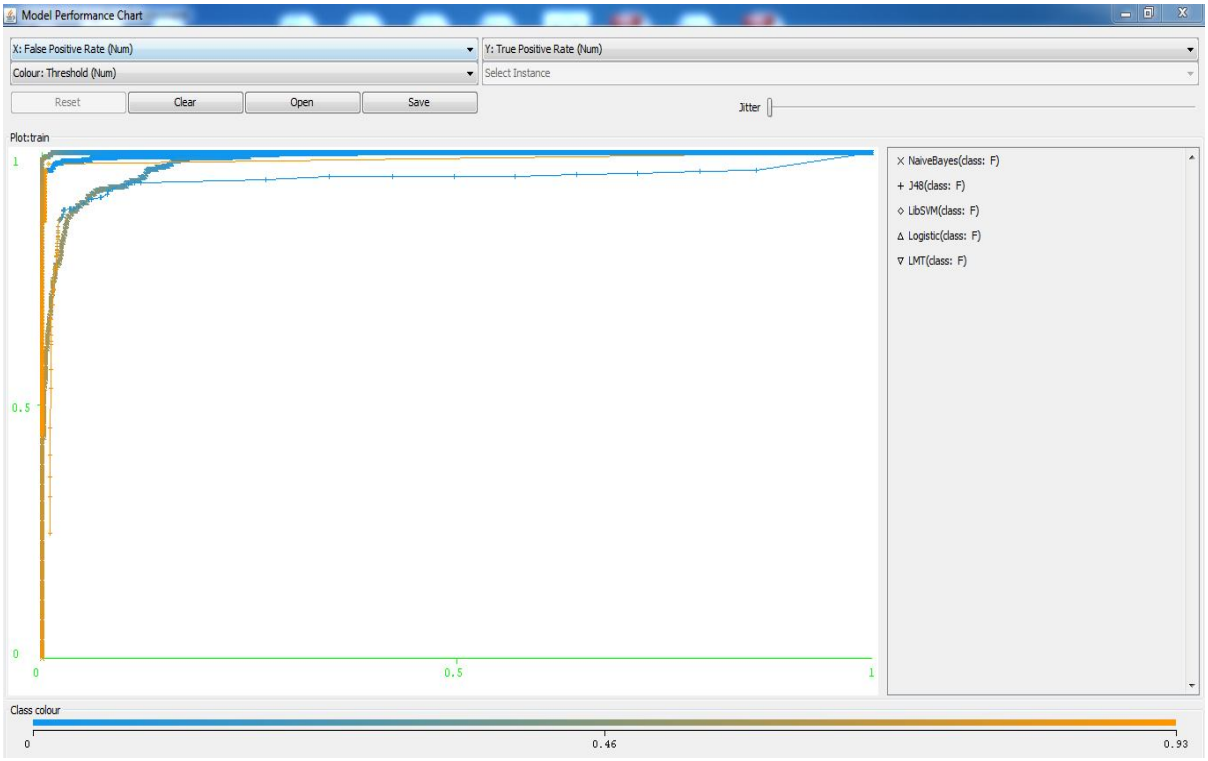


Figure8: Machine Learning approaches implementation through Knowledge Flow

The above figure depicts accuracy in results by comparing different machine learning approaches in terms of Model Performance Chart. Its conspicuous that prediction results in terms of TP rate is more significant for hybrid approach LMT as compare to C4.5, SVM, Naive Bayes and Logistic Regression.

Table II. Accuracy and Performance Measures of Machine Learning Standard and Hybrid approaches

	C4.5	**LibSVM**	**Naive Bayes**	**Logistic Regression**	**LMT**
ACCURACY	95.20	98.32	94.88	98.61	99.48
True Positive Rate	0.86	0.95	0.89	0.96	0.99
False Positive Rate	0.03	0.01	0.04	0.01	0.00
True Negative Rate	0.97	0.99	0.96	0.99	1.00
False Negative Rate	0.14	0.05	0.11	0.04	0.01
Precision	0.89	0.96	0.85	0.98	0.99
RECALL	0.86	0.95	0.89	0.96	0.99
F-MEASURE	0.87	0.87	0.95	0.97	0.99
KAPPA STAT	0.86	0.95	0.86	0.96	0.99
RMSE	0.16	0.09	0.18	0.09	0.07

The above table shows the comparative results of standard and hybrid machine learning approaches and from this its concluded that prediction results provided by Hybrid approach are much significant as compare to the standard approaches of machine leaning in terms of accuracy, correct and incorrect classification as well as other mentioned factors.

LMT exhibits higher accuracy i.e. 99.48 as compare to accuracy of LibSVM, FL J48 and Naive Bayes approaches. LMT, LibSVM and FL provides significantly better accuracy as compare to J48 and Naive Bayes approaches. In terms of precision, recall, F-measure and Kappa stats also LMT outperforms as compare to all other standard algorithmic approaches.[15]

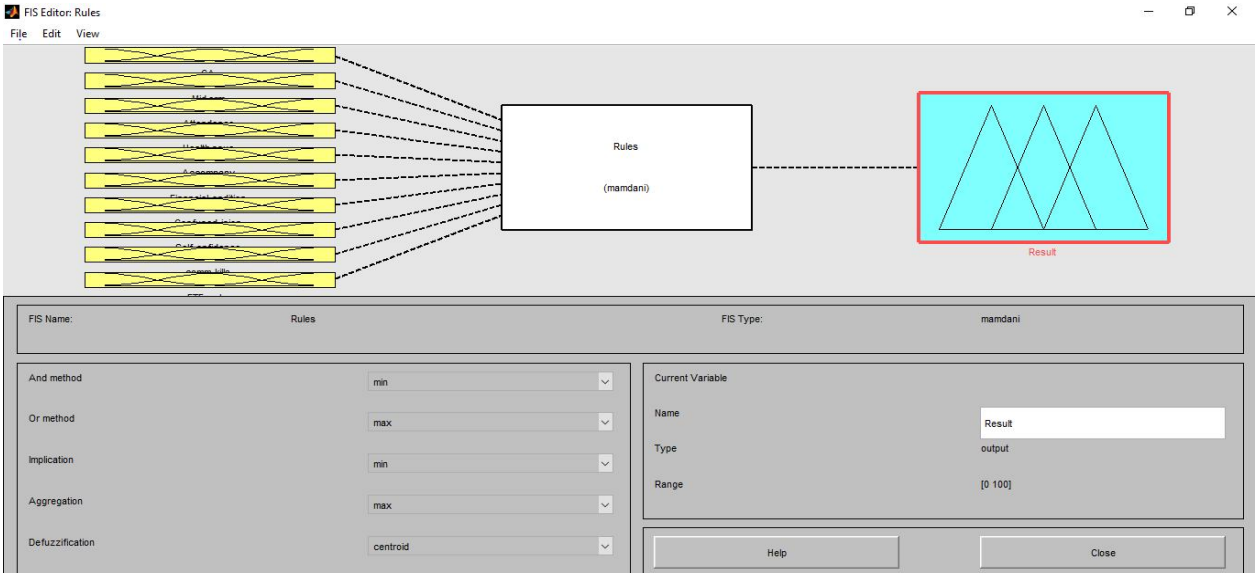


Figure9: Fuzzy input and output variable for predicting performance

The above Figure shows Membership function for each Input and Output variables. Fuzzy logic tool box has a list of prominent membership functions, you just have to choose the one you want to use. We have used triangular membership function in this research work. In Short, we would be able to predict student performance by applying fuzzy logic as well and by obtaining results in terms of accuracy.

VIII. CONCLUSION

Education System data mining is very relevant to do analyse and predict the performance of students in academics by considering different performance factors. This study would play a major role in constructive development of student and by taking care of them at right time in right direction, after predicting and analysing the instances and hopefully, would leads to decline in dropouts. Moreover, Education Mining will help in the analysis of various data related to education in terms of how various factors affecting overall performance of student. In addition of that, here the Hybrid machine learning classification approach gives better results in terms of accuracy and other parameterization and hence it would be prefer in achieving prediction results. Overall, these outcomes can be used by various educational organizations to work on the weak zones of student at right time by using right pedagogies and hopefully, would be able to achieve their respective objectives.

REFERENCES

[1] Hany M. Harb and Malaka A. Moustafa, “Selecting Optimal Subset of Features of Student Performance Model”, IJCSI International Journal of Computer Science Issue, Vol. 9, Issue 5, No, September 2012, pp. 253-262.

[2] Carlos Marquez, Cristobal Romero Morales and Sebastian Ventura Soto “Predicting School Failure and Dropout by Using Data Mining Techniques” IEEE Journal Of Latin-American Learning Technologies, Vol. 8, No. 1, February, 2013, pp. 7-14.

[3] Kiri Wagstaff and Claire Cardie “Constrained K-means Clustering with Background Knowledge” Proceedings of eighteenth international conference on machine learning, 2001, pp. 577-584.

[4] Grigorios F. Tzortzis and Aristidis C. Likas, Senior Member, IEEE “The Global Kernel K-Means Algorithm for Clustering in Feature Space” IEEE transactions on neural networks, VOL. 20, NO. 7, JULY 2009, pp. 1181-1194.

[5] K.A Abdul Nazeer and M.P Singh “Improving the accuracy and efficiency of k means, kohenon self organizing map and hierarchical agglomerative clustering”. Proceedings of world congress on engineering. Volume 1, London u.k, (2002).

[6] Saadat Naziova “Survey on Spam Filtering Techniques”, Communication and Network, August 2011, pp. 153-160.

[7] P. Moniza and P. Asha “An Assortment of Spam Detection System”, International Conference on Computing, Electronics and Electrical Technologies [ICCEET] 2012, pp.77-83.

[8] Patricia Bellin Ribeiro, Luis Alexandre da Silva and Kelton Augusto Pontara da Costa “Spam Intrusion Detection in Computer Networks Using Intelligent Techniques”, IFIP IEEE IM Workshop: 1st International Workshop on security for Emerging Distributed Network Technologies (DISSECT), 2015, pp. 304-311.

[9] Yen-Liang Chen, Hsiao-Wei Hu and Kwei Tang, “A Novel Decision-Tree Method for Structured Continuous-Label Classification” IEEE Transactions on Cybernetics, 2013, pp. 1734 – 1746.

[10] Qiang Yang, Senior Member, IEEE, Jie Yin, Charles Ling, and Rong Pan, “Extracting Actionable Knowledge from Decision Trees” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 1, JANUARY 2007.

[11] Jasna Soldic-Aleksic , Journal of Economics and Engineering, ISSN.: 2078-0346, Vol. 3. No.1, April 2012, pp. 241-248.

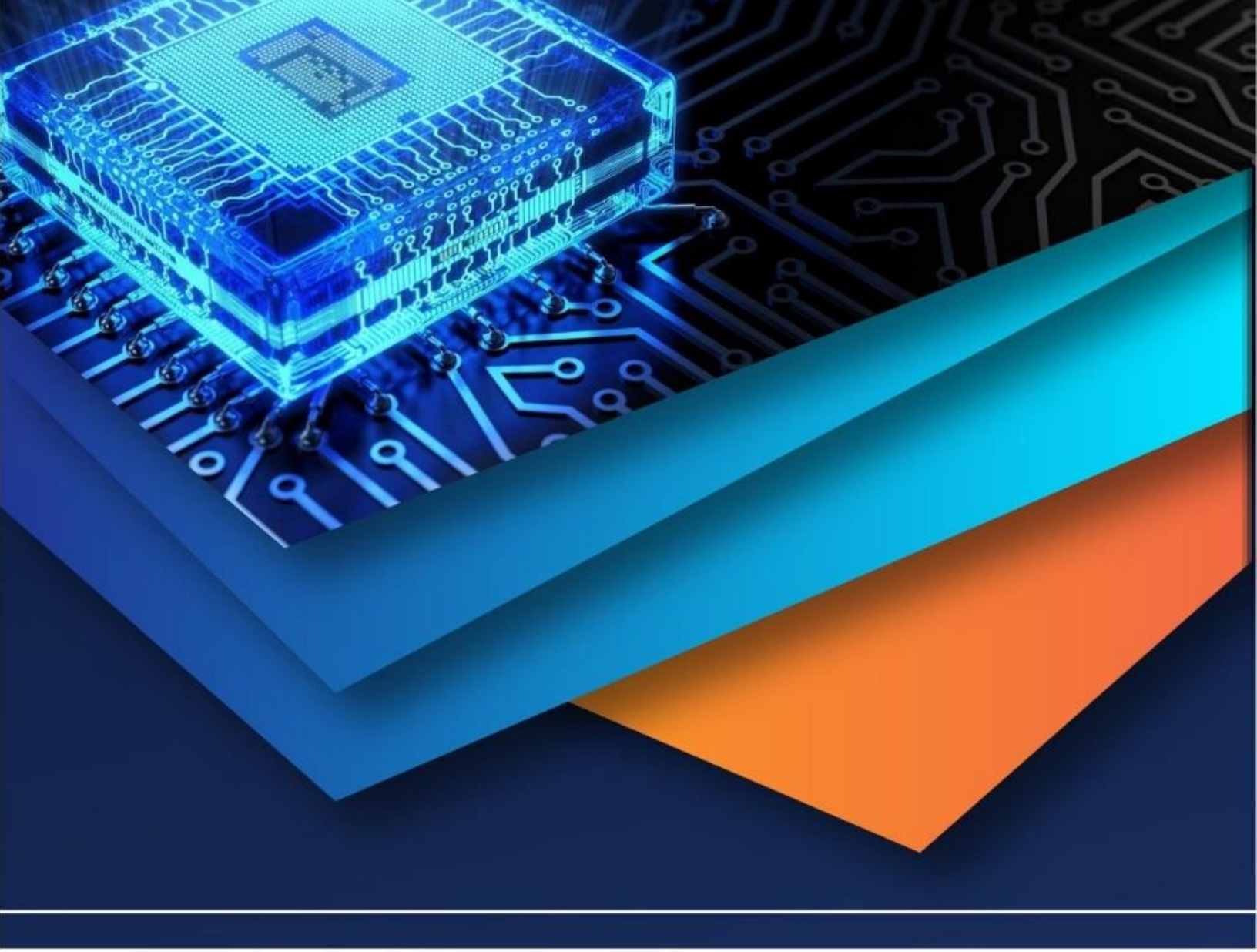
[12] Olaiya Folorunsho, “Comparative study of different data mining techniques performance in knowledge discovery from medical database”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013, pp. 11-15.

[13] Nancy Lekhi and Manish Mahajan, “ Outlier Reduction using Hybrid Approach in Data Mining” I.J. Modern Education and Computer Science, 2015, pp. 43-49.

[14] John Jacob, Kavya Jha, Paarth Kotak and Shubha Puthran ‘Educational Data Mining Techniques and their Applications’ International Conference on Green Computing and Internet of Things ,2015, pp. 1344-1348

[15] Motaz M. H. Khorshid , Tarek H. M. Abou-El-Enien ,Ghada M. A. Soliman, "Hybrid Classification algorithms for terrorism prediction in Middle East and North Africa" International Journal of Emerging Trends & Technology in Computer Science,2015, pp. 23-29

[16] Khan A. R., Amin, H. U., Rehman, Z. U., "Application of Expert System with Fuzzy Logic in Teacher's Performance Evaluation," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 2, February 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY