iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37368

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Student Campus Placement Prediction Analysis using ChiSquared Test on Machine Learning Algorithms

Ambika Rani Subhash
Department of Information Science Engineering, BMS Institute of Technology and Management, Bangalore, Karnataka, India

Abstract: ***Every higher education institute aims to provide the best career opportunities for their students as part of the outcome based education system. In India, campus placements for students while pursuing their 4th year of engineering is a predominant factor since the reputation of any institute largely depends on reputed recruiting companies visiting campus and the number of placement offers being given to eligible students. Hence, campuses offer personality development training to their students just before the commencement of the placement season while students try to maintain a minimum CGPA which would ensure their eligibility to apply for companies of their choice. The purpose of this paper is to predict a student's chances of obtaining a pre-placement offer while still in campus on the basis of various academic and non-academic factors. The dataset used for the prediction analysis consists of student related aspects such as their university seat numbers, academic grades and personality training parameters. The training models have been designed using the WEKA tool and in addition to supervised machine learning classification algorithms, Chi-squared tests has been implemented on the dataset to only obtain those attributes that might be the highest requirement for campus placements of students.***

Keywords: WEKA, Chi-squared test, Machine Learning Algorithms, Student campus placements.

I. INTRODUCTION

Every organization of repute when recruiting, looks for talented and accomplished professionals who will add value to the company they are being recruited for. Hence, campus placements of students while in their final year of undergraduate or postgraduate program plays a big part in a student's career. Top recruiting companies visit campuses every year and provide pre-placement offers to students who are still pursuing their final year course and this process helps all the stake holders involved. The stake holders, namely, the students are offered employment even before they step out of campus which saves them the time and energy of looking for jobs after graduation, while, companies are able to handpick fresh candidates who are best suited to their respective industries without much effort and without wasting too many resources. The profession oriented educational institute is also profitted by this because various accrediting committees and ranking surveys conducted by the media help the institutions strive for a healthier competitions among each other. They make an effort to increase the various factors that influence their ranking consistently, and providing placement opportunities on campus for their students , while they are still pursuing their studies is a major factor . Campus placements by reputed multinational companies and the salary packages offered plays a very important role and is a parameter that decides the branding and the demand for the educational institution [1].

India has a large youth population with about 65% of the population below the age of 35, and even though the levels of education being offered in the recent years has improved, enhanced skill development is still a critical problem [2]. In order to bridge this gap, many institutes across India impart an additional personality development placement training in the form of aptitude and soft skills, since some of the characteristics looked for in potential hires are usually, communication skills, teamwork, analytical skills, interpersonal skills, etc., [3]. It is also noticed that most recruiters set an eligibility cut-off criterion to avoid the tedious process of going through the profiles of all the available candidates in a batch.

Given the above scenario, this paper aims at developing a placement prediction model to predict the probability of students getting placed in a company hiring on-campus, by applying supervised machine learning algorithms using the WEKA tool. The main objective of this model is to try to effectively predict the possibility students getting a pre-placement job offer during campus recruitment. To achieve this, the data considered is the academic history of student like marks obtained in school, pre-university, cumulative grade points, arrears, attendance of placement training etc,.[4]. The dataset considered for analysis is of 1,200 students from BMS Institute of Technology, a reputed engineering college in Bangalore, India.

To ensure correct analysis of data, the chi-squared test methodology has been applied to gather only the attributes that have the highest possibility of getting a student placed. Supervised machine learning classification algorithms namely, J48, Neural Networks/Multilayer Perception, SVM and Naïve Bayes are then used only on the attribute set obtained after implementing the chi-squared test and results obtained for each of the algorithms are compared based on the accuracy measures.

Since employment of youth after college education is a major concern across the globe, it is necessary to design an educative system, by involving data mining techniques to discover useful details which can be applied to evaluations for learning to enable educators device pedagogical methods that would help them make important decisions about improvements in the education system. Data Mining techniques applied in educational eco-system is termed Educational Data Mining (EDM) and is deals with formulating innovative methods to mine knowledge from educational databases which can be utilized for decision making in the educational environment.

II. PROPOSED METHODOLOGY

A. Working Methodology

The end goal of this paper is specifically aimed at identifying classification algorithms which will faultlessly predict only those attributes that will help a student gain a pre-placement offer while still studying in campus. The following fig.1, gives a diagrammatic model representation for the proposed study.

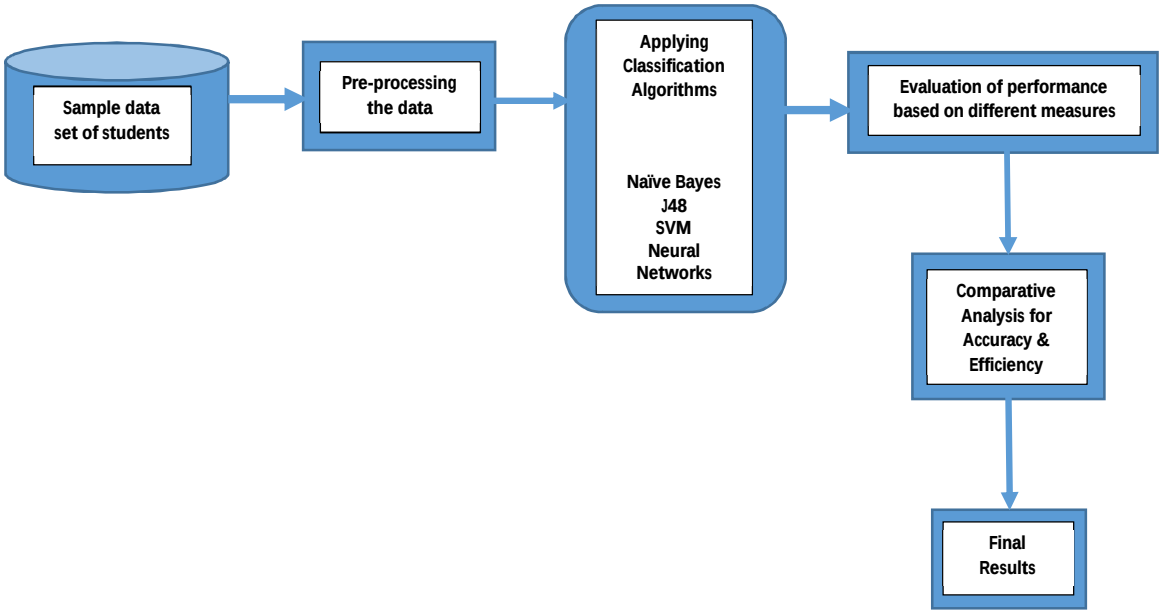


Fig 1: Block Diagram

From the above figure, it can be noted that the following steps have been considered in this study.

1) *Step 1*: The dataset that is used to train the model is the database of about a thousand students from a well-known technical institute in Bengaluru, India. This training set is loaded on to the WEKA tool. Since the study is on campus placements of final year students, the academic history of the students from their grade 10, 12 and their B.E CGPAs till their final semesters were considered along with their eligibility to attend placements and current backlogs. In addition to the academic history of the students, certain non-academic criteria, such as the department of engineering, their placement training attendance, gender and date of birth was also considered. Since this study is aimed at identifying the academic / non-academic parameters that are necessary for students to obtain pre-placement offers from top recruiters, certain limitations, such as missing CGPAs / gender etc., was placed on the large dataset while choosing the instances. The dataset also considers the numbers of pre-placement offers obtained by an individual student based on their credentials. The considered dataset has instances of 1002 students out of which 368 are female students and the remaining 634 students are male. The below Table 1, depicts the 9 characteristics along with the considered abbreviations and types of attributes.

Attributes	Abbrv	Type
Department of Engineering	branch	nom
Gender	gender	nom
Grade ten marks	tenth	num
Marks obtained in pre-university	puc	num
CGPA obtained in engineering	CGPA	num
Eligibility to apply for placements	eligibility	nom
Placement Training Attendance	attend	nom
Backlogs	backlogs	nom
The last column of the training set states if the student has a pre-placement offer or not, this is represented by (1) or. (0) respectively.	offers	disc

Table 1: Attributes of Student Dataset

2) *Step 2:* As the next step, the WEKA tool is used for pre-handling the information. The data is utilized in .arff design. The Waikato Environment for Knowledge Analysis, abbreviated to WEKA is an open source programming accessible suite of AI programming written in Java and created by the University of Waikato, New Zealand. Now, as the dataset that is used comprises of additional information that might not be appropriate for use for this work, data mining is considered for examining the information. As the initial step, the dataset is pre-processed since the existing data could be insufficient and with many errors. The existing information might also lack attributes and may have outliers. Thus, to solve the problem of inconsistency, the information is pre-processed to obtain quality and consistent data. The below Fig 2, shows the WEKA GUI and the visualization of the pre-processed and cleaned data.

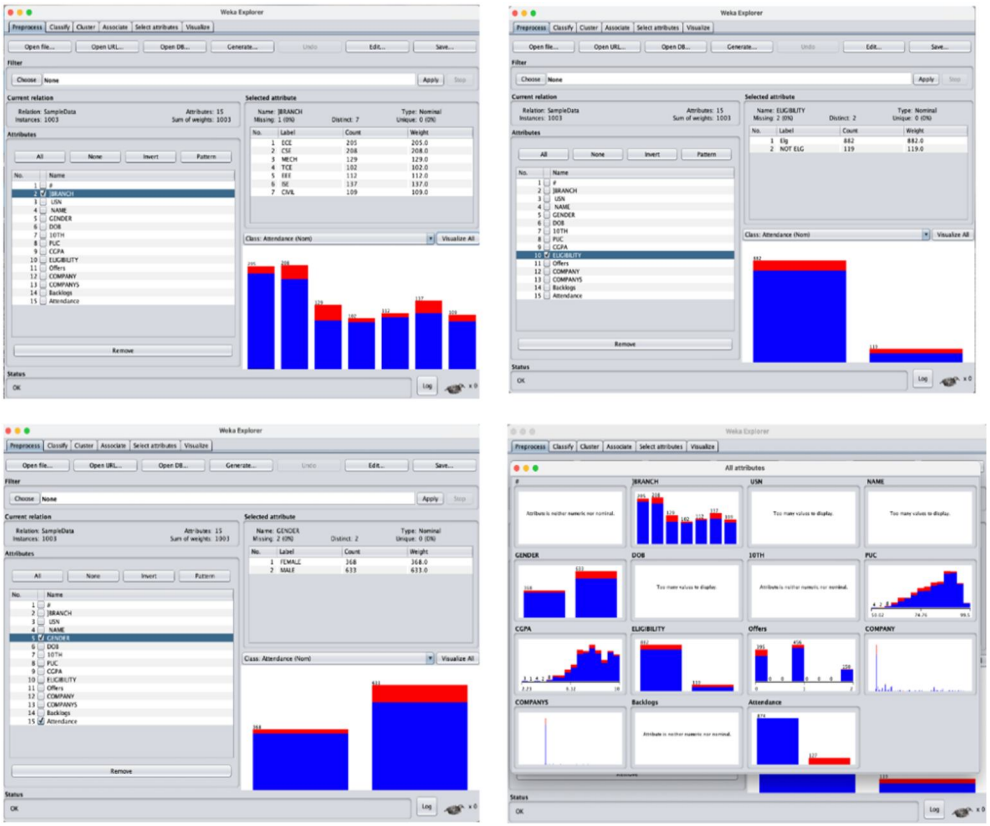


Fig 2: Sample visualization of pre-processed data on WEKA GUI

3) *Step 3* -- As part of the next process, information mining is to be done and this process depends on the attributes that has been chosen to perform the process. In order to do this, a Chi-square factual test can be utilized to choose the most notable attributes, and this will in-turn choose the correct objective classes so that the attributes that are least significant can be eliminated and only the most necessary ones can be retained. The idea behind using the Chi-Square method is to only select those features or attributes that plays an utmost important role in predicting the output that we are interested in. This method is used to narrow down the potential list of attributes to attain only those attributes, that might have most importance in predicting the final outcome in the model.

Formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

The WEKA tool's machine learning capabilities are made use of to try and gather only those patterns that are usually not always evident in visualizations. Using the attribute evaluator, 'chisquareattributeeval' in WEKA, we can see the most important attributes ranked, along with merit weightage given to each attribute. This rank and merit analysis helps us understand about the attributes that are most powerful in helping us predict the pre-placement of students while they are still pursuing their final year of engineering. The below figures show the visualization and top ranked attribute selection.

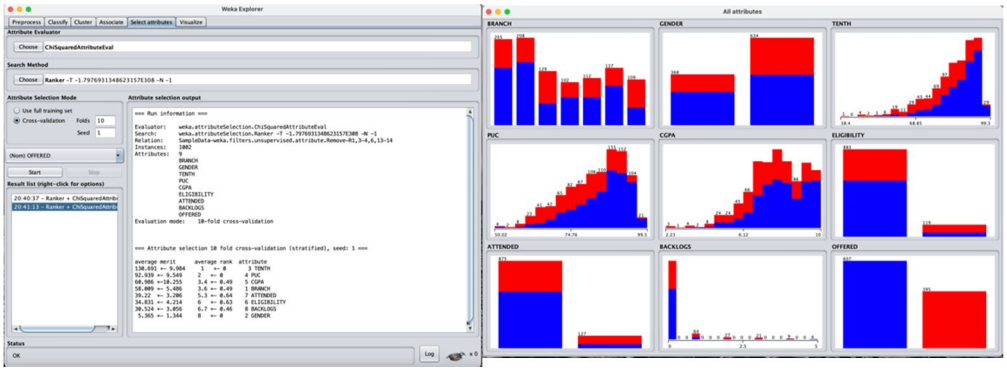


Fig 3: Visualization and selection of top-ranked attributes using ChiSquare Test

From the above figure 3, we see that on running the ChiSquare test on the student dataset, with 'offered' as the selected attribute, we have the top ranked attributes that could be a major indicator about the important aspects that could get a student a pre-placement offer. We see that the following factors are really important in determining if a student will get placed in companies that are visiting on-campus and are ranked based on their importance level,

- Marks obtained in grade 10 - tenth
- Marks obtained in pre-university - puc
- CGPA obtained in the engineering course - CGPA
- Department of engineering a student belongs to - branch

We can also see that the eligibility of a candidate ranks after the above listed details and this shows that a student can be eligible to apply for visiting companies only if he satisfies the marks obtained criteria and belongs to the most sought after departments of engineering. Another observation made is that backlogs and gender of a student plays very less importance when a student is applying for campus drives as recruiting companies do not differentiate between male or female students.
Another interesting observation is that the attribute, placement training attendance - 'attended', of a student also plays an interesting role in placements of a student. We see that this attribute is ranked at number 5 and shows that students who have attended atleast 50% of placement training sessions have been given pre-placement offers. From the below figure 4, we can see that on running the

ChiSquared test on the 'attended' attribute we see that the top ranked attributes are as follows, and shows us that students who have attended the placement training are more likely to be eligible to apply for on-campus placements and get the pre-placement offers.

- Eligibility to apply for placements - eligibility
- Marks obtained in pre-university - puc
- Students getting pre-placement offers - offered

Also seen in below figure 4, is that when 'eligibility' is selected as the attribute on which ChiSquare test is run, the academic marks of a student, CGPA, Tenth marks and Pre-university marks are the top ranked attributes that will get a student placed on-campus.

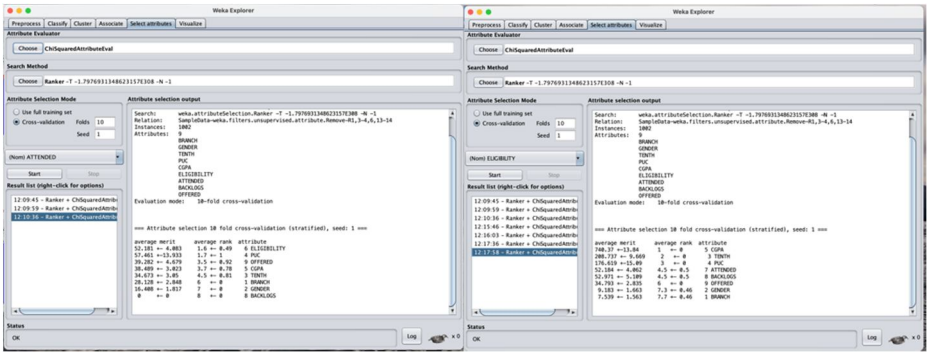


Fig 4: ChiSquare Test on 'attended' and 'eligibility' attributes

4) *Step 4:* In this step we apply different supervised classification algorithms as follows, on the training student dataset which is obtained after the initial first three steps. Though there are many supervised classification algorithms, for this experiment the algorithms used are J48, Neural Networks/Multilayer Perception and Naïve Bayes,.

a) *J48:* This is a machine learning algorithm that is used to examine the datasets continuously and categorically. It is an open source Java implementation of the C4.5 decision tree algorithm.

b) *Multilayer Perception:* This is a deep artificial neural network which consists of an input layer, a hidden layer and an output layer. The input layer receives the signal, and the output layer makes a decision or prediction about the input. The hidden layer is the computational engine of the algorithm.

c) *Naïve Bayes:* In this algorithm, each input variable is assumed to be independent. The assumptions are quite unrealistic for real data but, this method is quite reliable on a large range of complex problems. This classification algorithm helps to build fast machine learning models which help make fast predictions.

The below table 2 shows the Confusion Matrix created for each of the above listed classification algorithms. A confusion matrix is a technique for summarizing the performance of classification algorithms. Confusion matrices are used to visualize important predictive analytics like recall, specificity, accuracy, and precision. Confusion matrices are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives [12].

Supervised Classification Algorithms	Tested Negative (a)	Tested Positive (b)
J48	471	136
	156	239
Multilayer Perception	454	153
	151	244
Naïve Bayes	505	102
	193	202

Table 2: Confusion Matrix for the classification algorithms

5) *Step 5:* This is the last and final step of this experiment which involves comparison and analysis of the accuracy measures and performance on the training dataset by the various supervised machine learning algorithms as defined above. All the experiments were carried out by the internal 10-folds cross-validation method. Cross-validation is a method to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. With the 10-fold cross-validation, there is one dataset that is divided randomly into 10 parts, out of which 9 pf those parts are used for training and one tenth is reserved for testing. This process is repeated 10 times and on each of those times, a different tenth is reserved for testing. From the below Table 3, we see the various accuracy measures along with their representations and the formulae that is applied to obtain the final results.

From the table, it is seen that the formulae for calculating the accuracy measures the confusion matrix has two classes, the positive class and the negative class which give the positive and negative predictions respectively and are classed as True Positives and False Positives and Ture Negatives and False Negatives.

Accuracy Measures	Denotations	Formulae
Accuracy (A)	Accuracy with which the algorithm predicts instances	A = (True Positives + True Negatives) / (Total Samples)
Precision (P)	Gives the exactness of classifier algorithms	P = True Positives / (True Positives + False Positives)
Recall (R)	Measures the sensitivity of classifier algorithms	R = True Positives / (True Positives + False Negatives)
F-Measures (F)	Gives the average of Precision & Recall	F = 2*(Precision*Recall) / (Precision + Recall)
ROC	Expanded as Receiver Operating Characteristic curves, and it compares the tests.	

Table 3: Accuracy Measures

III. EXPERIMENT AND RESULT

To discuss the results of this study, we first run experiments on the training set that we have obtained after running the Chi-Squared test. Each of the training models were compared based on the performance of each classifier algorithm. The comparison of each algorithm was done based on the different accuracy measures. From the below table, Table 4 we can see the accuracies with which each of the considered machine learning algorithms predicts the placements of students while they are still pursuing their studies. We can also see the number of instances that have been classified correctly or incorrectly from the Table 5, for each of the algorithms.

Supervised Classification Algorithms	Accuracy % (A)	Precision (P)	Recall (R)	F-Measure (F)
J48	70.85 %	0.706	0.709	0.707
Multilayer Perception	69.66 %	0.697	0.697	0.697
Naïve Bayes	70.55 %	0.700	0.706	0.697

Table 4: Accuracy Measures Comparison

Supervised Classification Algorithms	Correct Instances	% of correctly classified instances	Incorrect Instances	% of incorrectly classified instances
J48	710	70.85 %	292	29.14 %
Multilayer Perception	698	69.66 %	304	30.33 %
Naïve Bayes	707	70.55 %	295	29.44 %

Table 5: Classification of Instances

Hence, from the above tables 4 and 5, it can be seen that the J48 algorithm has performed better that the other algorithms in predicting if a student will be placed on campus based on the most important attributes derived after applying the ChiSquared test. We also see that this algorithm has classified 710 instances correctly out of the total 1002 instances. From this result we can also see that the J48 algorithm is one of the best machine learning algorithms to examine data continuously and categorically. We also notice that Multilayer

Perception is the least performing algorithm and hence can say that with real world data analysis, the J48 algorithm performs better.

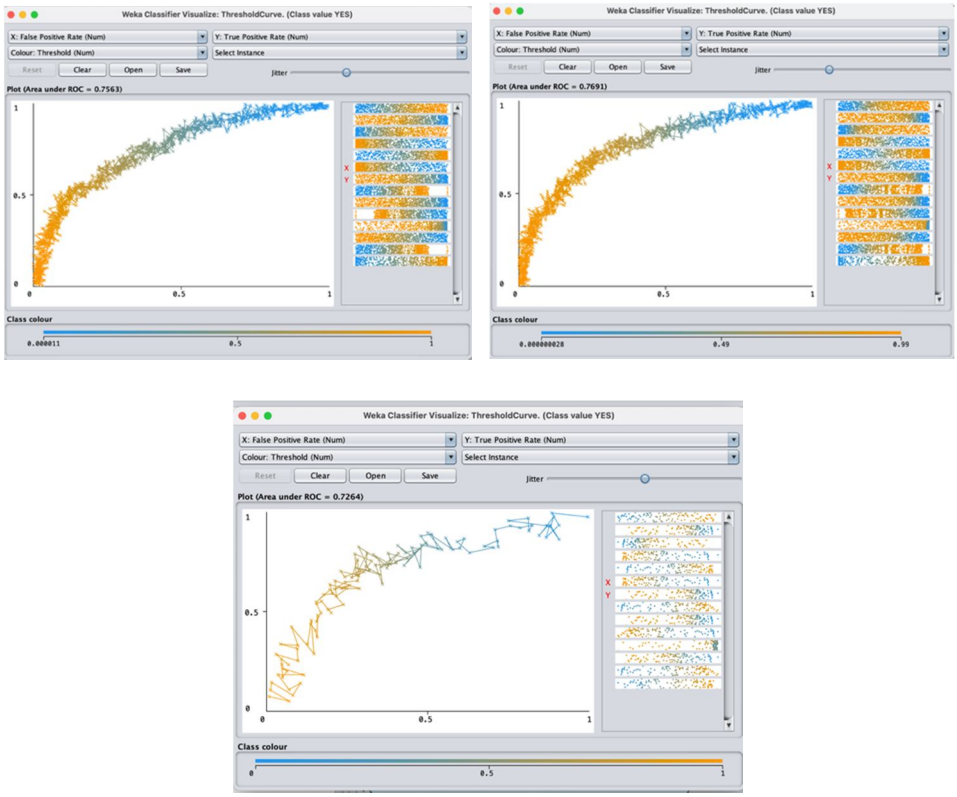


Fig 5: Threshold curves of Multilayer Perception, Naïve Bayes and J48 respectively

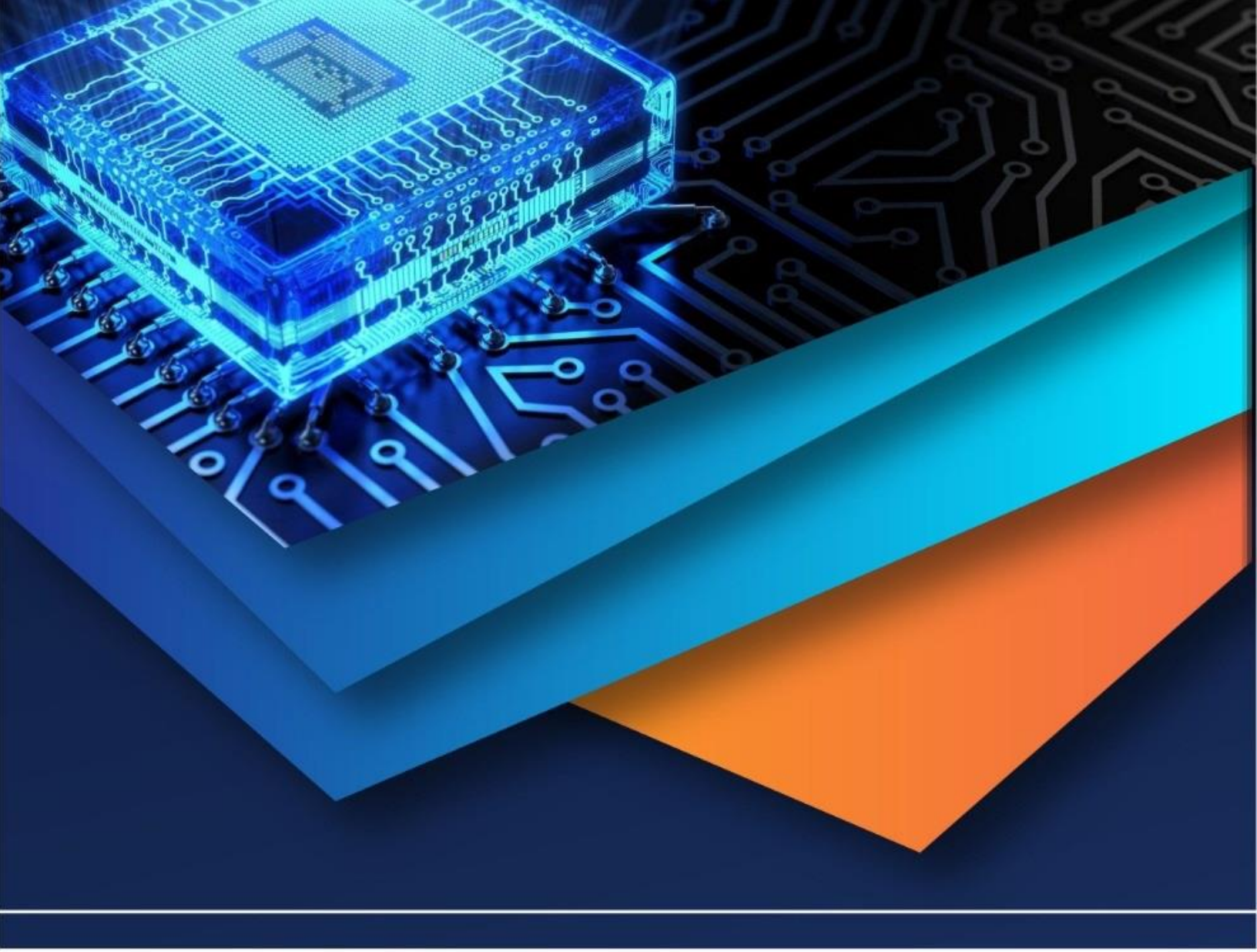
The above figure 5, shows the threshold curves or the Receiver Operating Curve (ROC) area for each of the compared algorithms. The ROC curve benefits in interpreting the performance measures of the classification algorithms. The curves are usually used to choose the most appropriate threshold for any test on the dataset. With the help of these threshold values we can predict how many students may get pre-placement offers based on the exactly specified negative and positive values.

IV. CONCLUSION

This paper discusses the various aspects of campus placements of engineering students and the main factors based on which students can expect to be placed while studying their professional courses. The training dataset was subjected to supervised classification algorithms to understand which one of the algorithms could precisely predict job opportunities for students based on subjective criteria. In order to correctly predict this, the dataset was initially subjected to the ChiSquared test to pull out only those attributes that had the most impact on the students' placement records. Hence, for this paper it can be concluded that out of the compared machine learning algorithms the J48 algorithm out performs the others to predict student placements on campus while they are still studying. It was seen that academic criteria such as their grades obtained in their class 10 and pre-university and in their professional courses played a big part in bagging a good job in a reputed company.

REFERENCES

[1] Mohamed Tajudeen S, Aravindh Kumaran L, "Campus Recruitment Process - A perspective of the Stakeholders", International Journal of Engineering Technology, Management and Applied Sciences Vol. 5, Issue 2, 2017, ISSN 2349-4476.
[2] K. Shashikanth, G.Pranay, "A Study Report On Importance of Campus Placement - A Boon to Students' Career", International Journal & Magazine of Engineering Technology, Management and Research Vol. 3, Issue 2, 2016, ISSN 2348-4845.
[3] Suresh Kumar N, Prashanth MK, Ajith Sundaram, "Campus Placements in Kerala-An empirical study at the selected Engineering Colleges in Kerala", International Journal of Scientific and Research Publications, Vol. 3, Issue 1, 2013, ISSN 2250-3153.
[4] Pallavi Devendra Tawde, Dr. Sarika Chouhan, "A Survey of Machine Learning Techniques for Student Performance and Placement" Journal of Xi'an University of Architecture & Technology, Vol. XI, Issue XII, 2019, ISSN 1006-7930.
[5] Pothuganti Manvitha, Neelam Swaroopa "Campus Placement Prediction Using Supervised Machine Learning Techniques", International Journal of Applied Engineering Research, Vol. 4, Number 9, 2019, ISSN 0973-4562.
[6] Shreyas Harinath, Aksha Prasad, Suma HS, Suraksha A, Tojo Mathew "Student Placement Prediction Using Machine Learning", International Journal of Scientific and Research Publications, Vol. 6, Issue 4, 2019, ISSN 2395-0056.
[7] K Sreenivasa Rao, N Swapna, P Praveen Kumar, "Educational Datamining for student placement prediction using machine learning algorithms", International Journal of Engineering & Technology, 7 (1.2), 2018, 43-46.
[8] Joshitha Goyal, Shilpa Sharma, "Placement Predictions Decision Support System using Data Mining", International Journal of Engineering and Techniques, Vol. 4, Issue 6, 2018, ISSN 2395-1303.
[9] K Manikandan, S Sivakumar, M Ashokvel, "A Classification Model for Predicting Campus Placement performance Class using Data Mining Technique", International Journal of Advance Research in Science and Engineering, Vol. 7, Special Issue 6, 2018, ISSN 2319-8354.
[10] P. N. Shejwal, Nageshwar Patil, Akash Bobade, Akshay Kothawade, Sadashiv Sangale, "A Survey on Student Placement Prediction using Supervised Learning Algorithms", International Journal of Research in Engineering, Science and Management, Vol.2, Issue 11, 2019, ISSN 2581-5792.
[11] Eibe Frank, Mark A. Hall, and Ian H. Witten, "The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"", Morgan Kaufmann, Fourth Edition, 2016.
[12] https://kambria.io/blog/confused-about-the-confusion-matrix-learn-all-about-it/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY