INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 05 ISSUE: 06 | JUNE - 2021 ISSN: 2582-3930

Novel Campus Recruitment Prediction using Machine Learning Techniques

Shruti Sodhe
Department of computer engineering
MES College of Engineering
Pune,India

Purva Suryavanshi
Department of computer engineering
MES College of Engineering
Pune,India

Swarangi Kokate
Department of computer engineering
MES College of Engineering
Pune,India

Santosh Kharpude
Department of computer engineering
MES College of Engineering
Pune,India

Prof. Sharmila Wagh
Department of computer engineering
MES College of Engineering
Pune,India

Prof. Amol Kamble
Department of computer engineering
MES College of Engineering
Pune,India

Abstract— **One of the important activities in the academic curriculum of final year students is the placement activity. The placement activity also plays an important role in establishment of name and future admissions of various institutes. Hence all institutions strive to strengthen their recruitment practices. This project is aimed to analyze student's academic data and predict their placement possibilities and provide aptitude links so that the students can study accordingly and thus increase the placement percentage of the institutions. Monitoring the student's progress for his or her campus recruitment helps in monitoring the student's development within the academic backgrounds. The objective of institutions is to supply superior opportunities to their students. This proposed campus prediction system is utmost vital method which can be used to distinguish the student data/information on the basis of the student performance. Handling placement and tutoring records in any larger institutions is quite hard because of the large number of students. This system can categorize the student knowledge easily and can be useful to many educational institutions. There are numerous classification algorithms and mathematics-based techniques which are good assets for classifying the students' information set in the education field. In our system Logistic Regression algorithm is applied to predict student performance which can facilitate to identify performance of students and also provides suggestion to improve performance, for students we are going to classify the student's skills set for placement and nonplacement classes based on that result, education institutions can give more training to their students accordingly. Based on data received by system, student's performance is analyzed in numerous views to check the achievements of the students through their activities and suggests improvement for better placement. To predict whether the student will be placed or not is the main purpose of the proposed system. It will also provide students with links which they can refer for aptitude practice. This will help them to improve their skills so as to increase their possibility of getting placed. This model helps the placement cell and the companies to spot the students having potential and focus to and improve their technical and social skills.**

Keywords— **Machine Learning, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Naïve Bayes, SVM, Sigmoid Function, ROC curve, Confusion matrix, Database.**

I. INTRODUCTION

There is a rapid growth of educational institutes and the aim of every educational institute is to get their students a well-paid job in a well-established organization through their placement cell and also to help the students to achieve their dream job. One of the main challenges that institutes face nowadays is to enhance the performance of students with respect to the expectation of the hiring company. Therefore, it is very important to make a process hassle free so that, students would be able to get required information as and when they require. Also, with the help of the good system it would be easy for staff of the Training and Placement cell to update students easily and the work would be less. The "Campus recruitment Prediction using Machine Learning" is determined to eradicate and, in some cases, lessen the challenges faced by the existing system. Moreover, this system is designed for a need of company to carry out operations in smooth and effective manner. Most of the companies prefer campus recruitment to fill up their positions instead of using the traditional method to provide employment. The companies select talented and qualified professionals before they have completed their education. Due to this method the students get best companies at the beginning of their career itself.

Every Institution whether big or small is aimed to get their students in a well-established organization. Therefore, the goal of "Campus recruitment Prediction using machine learning " is to reduce the hardships faced by the institutions and the students in final year by predicting the students possibility of getting placed. So, the final year's students can focus on getting employed in reputed companies. It will also

help Faculty of placement cell to keep a record of the students and help them accordingly.

II. LITERATURE REVIEW

[1] The author has compared algorithms like Decision Tree, Naïve Bayes, Naïve Bayes Tree, K-Nearest Neighbor and Bayesian Network algorithms for estimating students' performance. Based on the grades of the students four categories are used to classify the students. To improve the accuracy and efficiency of each classifier Bootstrap method is used. Algorithms like IBK, Decision Tree and Bayes Net algorithm have delivered an excellent performance. The outcome of the algorithms compared by the author are good but the individual outcome is not satisfactory. In the paper author has also compared the results of past study to overcome the drawbacks of the algorithms, the individual class accuracy has also improved significantly. Therefore, according to the study we can conclude that the best algorithm for prediction of student's future grades is J48 classifier.

[2] This paper is focused on student's results. These days at the end of every year or semester in the schools or colleges, the teachers have to sort the students in their classes who are eligible for exams and who are not. They have to examine a lot of data and find who is eligible and who not every time is. To overcome it, the author in this paper has proposed a student result prediction model which is a web service, in which some machine learning algorithms are used to find whether the students are eligible for writing the exams or not based on their overall performance. Initially a csv file is created with the fields Student full name, Illness, Attendance, SSC result, HSC result all of which are in percentage format it also consists of some additional fields like Fathers education, Mothers education, Hostel staying, Study Hours, Sports, Disability and Medium studied which are also equally important in the prediction process. The given data is split for data transformation. The data is divided in 70-30 manner, in which 70% of data is used for training the model and the remaining 30% is used for testing purpose. Train the model based on the data set. The train model has two inputs ML algorithm, 70% of the data split provided by the user. The score model has two inputs of data train model, 30% of the data split provided by the user. This model evaluates the score results and calculates the machine learning parameters to predict whether the student is eligible for exam or not. The drawback of this system is that it can be used for small dataset thus the system can be made more effective by training the system for large dataset.

[3] The aim of the Employment process is to get a right person with effective process, in minimal time and cost. The dataset includes attributes such as experience of work, Present income, if already placed, list of past companies, highest degree, if the candidate belongs from a well reputed institute, the number of Research papers published and whether they have done any internship. Based on training set entropy and information gain is calculated for creating a decision tree. The pair of parameters will determine whether they will be placed or not. The attributes like experience of work, Present income, highest degree and whether they have done any internship are the main features for placement study. The ID3 algorithm makes the recruitment process easy and fast as it helps the recruiter to come into conclusion whether to hire the applicant or not according to the job profile.

[4] The author has proposed a system that uses data mining techniques to predict the possibility of student getting placed in this paper. To predict the possibility of student's placement, author has categorized the data into the two segments, first segment is the training segment which is historic data of passed out students. Another segment consists of current data of students, based on the historic data author has designed the algorithm for calculating the placement chances. Author has used the various data mining algorithms such as decision tree, Naive Bayes, neural network and the prosed algorithm were applied, and decision are made with the help of confusion matrix.

[5] In this paper the authors have proposed a system to ease the efforts required for prediction of student's placement. They have developed a system in which students will register into the system and enter their complete biodata. Based on the student's academic data the system will predict whether the student has the potential to get select by the company or not. Along with it, it will also recommend courses to the students. The admin will create a course and register students accordingly. Only the admin can access the student's data and which course they are registered to. If the student is eligible a mail will be sent to the student and that students name will be displayed on the dashboard of the respective college. Various machine learning algorithms like SVM, KNN and naïve bayes are used. Along with these algorithms' ensemble methods are also used to improve the accuracy and stability of learning models.

[6] In this paper a study was conducted based on the performance of students from 5 different degree college. A new technique was used to produce a dataset. To fill up the missing, transforming values a relevant dataset was used. Thus, for the construction of Byes classification prediction model they had 300 student records. It was found that the attributes such as students' personal information including grades, native place, medium of teaching, qualification of their parents, students' behavior, and annual income of the family and status of family were equally important to boost the students' overall performance after applying Bayesian classification method on 17 attributes. According to the recent study academic performances of the students are not always depending on their own effort. The aim of this system is to overcome the drawbacks of existing methods.

[7] This paper to predict placement status of the student provided whose information is provided through text input using machine learning technique. The objective of this paper is to study ex-student's qualification and predict whether the current student will be recruited or not and also help the institutes for strengthening their TPO. Companies can also be predicted based on previous

year's student's data. Here two different algorithms are used, namely Naive Bayes Classifier and K Nearest Neighbors [KNN] algorithm. The results of these algorithms are compared to achieve effectiveness in prediction. This algorithm uses attributes such as USN, Tenth and PUC/Diploma results, CGPA, Technical and Aptitude Skills. The required data is retrieved by using data mining techniques. KNN algorithm is applied to the input dataset then Euclidean distance is calculated by comparing past and new parameters and the output is predicted on the resemblance events. The efficiency of the system can be improved by adding more attributes.

[8] Now-a-days institutions are facing many challenges regarding student placements. For educational institutions it is much difficult task to keep record of every single student and predict the placement of student manually. To overcome these challenges, concept of machine learning and various algorithms are explored to predict the result of class students. For this purpose, training data set is historical data of past students and this is used to train the model. This system predicts placement status in 5 categories- dream organization, core organization, mass recruiter, not suitable and not concerned in placements. This system is also helpful to weaker students. Institutions can provide extra care towards weaker students so that they can improve their performance. By use Naïve Bayes algorithm all the data will be monitor and appropriate decision will be provided.

III. EXISTING SYSTEM

This paper focusses on a placement prediction system to predict whether the student will be placed in a company or not. The k-nearest neighbor's classification is used for the same. The output obtained by this technique are compared with the results of other machine learning algorithms like logistic regression and SVM to ensure efficiency. They have also considered the past qualification of the students and their skill set like, programming, communication, analytical and team work, which are equally important. Recruiter also examine skill set of students along with the academic history. The KNN algorithm is applied to the dataset to predict the possibility of placement. The idea of comparing the result of KNN algorithm with models like Logistic Regression and SVM makes the system more effective. But the system would be more efficient if there were some aids so as to improve the student's skills. The advantage of this system is that it can be used for predicting whether the student will be placed or not.

Variable	Variable Range	Variable Type	IV/DV
Gender	0,1	Numeric, Discrete	IV
10^{th} Percentage	0-100	Numeric, Continuous	IV
12^{th} Percentage	0-100	Numeric, Continuous	IV
B.E. Aggregate	0-100	Numeric, Continuous	IV
Backlogs	0,1	Numeric, Discrete	IV
Prediction	0-100	Numeric, Continuous	DV

Table 1 – Data variables considered for classification

Variable	Variable Range	Variable Type
Technical Skills	0-10	Numeric, Discrete
Communication Skills	0-10	Numeric, Discrete
Analytical Skills	0-10	Numeric, Discrete
Teamwork	0-10	Numeric, Discrete

Table 2 – Skill set used for post processing

The result that we obtained from the k-nearest neighbors was compared against the results obtained from logistic regression that gave an accuracy of around 75% and SVM with an accuracy of around 77.38%.

IV. PROPOSED SYSTEM

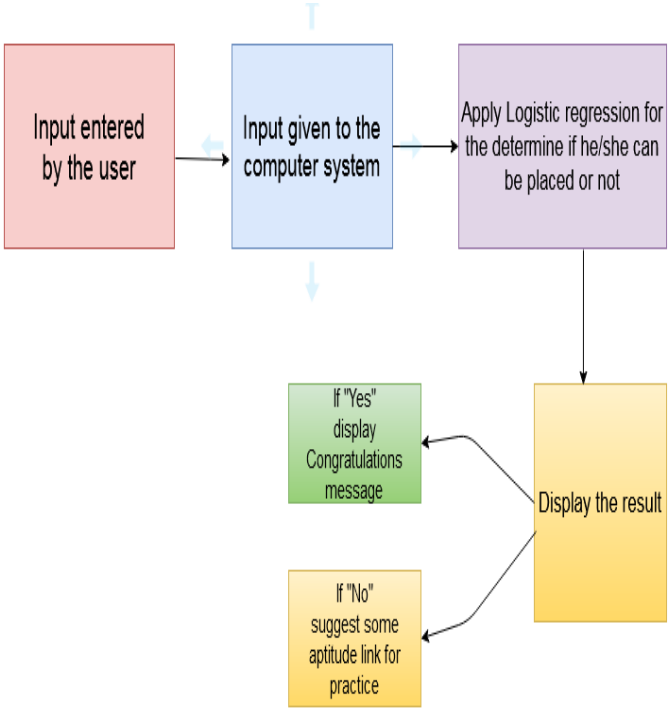

Fig 1 – System Architecture

Campus placement plays a very important role in the academics of final year students. Placement is an activity in which various companies visit the institutes to select potential students before they complete their graduation. The main Aim of this project is to help the TPO and the students to get an idea about which students have the potential to get placed and which students need help to improve their skills. It will also provide students with links which they can refer for aptitude practice. This will help them to improve their skills so as to increase their possibility of getting placed.

Project Modules

Input module: -Data Analysis is all about finding some interesting insights into the data and we can find more insight by asking more questions. In this module, the user is asked to insert his/her information which will be used to predict whether he/she will get placed or not. Python language and libraries which we are going to use are seaborn, NumPy, pandas, matplotlib, and scikit, and for the front-end, we will use Html, CSS, and JavaScript.

Prediction module: - This module is used in classifying data based on the data received from the input modules. The output data which is extracted will be used to train our logistic regression to find the possibilities of he/she getting placed. We are going to use different datasets that are available online to train our module in order to improve the performance of logistic regression Classification.

Suggestion module: -This is the final module in which the system will suggest some links to the student which will help them for their further development. Python is the most compatible language henceforth we are going to use python for this module.

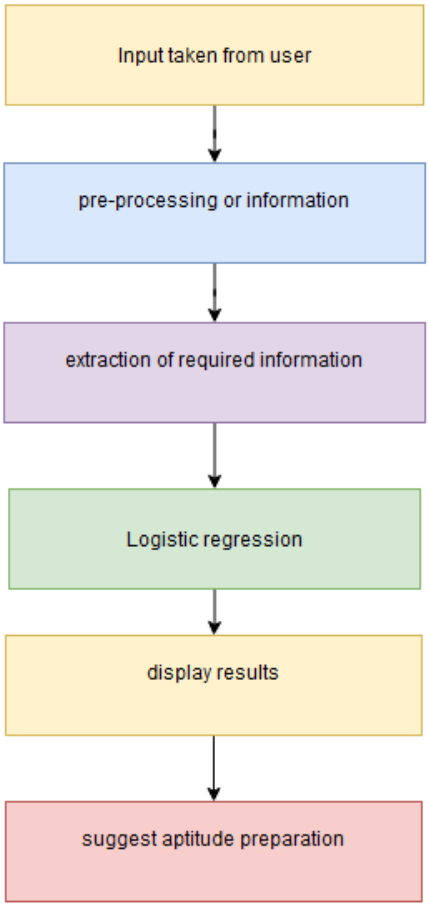

Fig 2- System Flow

ADVANTAGES OF PROPOSED SYSTEM

- Provides study aids to students.
- Less manual efforts required.
- More accurate as compared to other models.
- Efficient.
- Reliable.

V. METHODS AND MATHEMATICAL MODEL

Placements hold great importance for students and educational institutions. It helps a student to build a strong foundation for the professional career ahead as well as a good placement record gives a competitive edge to a college/university in the education market.

This study focuses on a system that predicts if a student would be placed or not based on the student's qualifications, historical data, and experience. This predictor uses a machine-learning algorithm to give the result.

The algorithm used is logistic regression. Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X. Talking about the dataset, it contains the secondary school percentage, higher secondary school percentage, degree percentage, degree, and work experience of students. After predicting the result its efficiency is also calculated based on the dataset.

A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

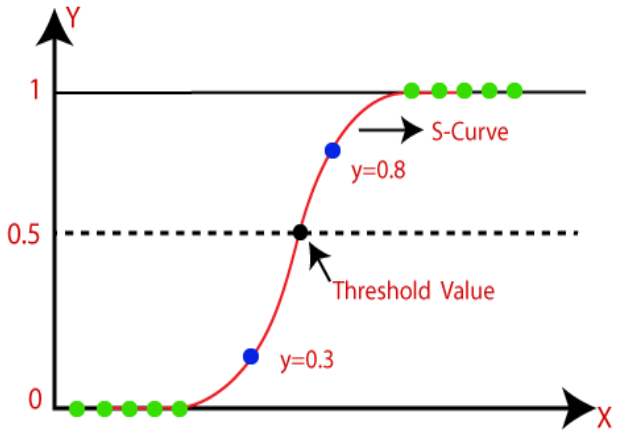

Fig 3- Logistic function

A. *Logistic Function (Sigmoid Function):*

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

B. *Logistic Regression Equation:*

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Algorithm:

1. Start.
2. Import libraries.
3. Split the dataset as training data and testing data.
4. Feature scaling.
5. Apply logistic regression algorithm on the dataset.
6. Predict the test set result.
7. Display the confusion matrix.

8. End.

Pseudo Code:

Step 1. Feature Scaling

scaler = StandardScaler ()

scaler.fit(X_train)

X_train = scaler. transform(X_train)

X_test = scaler. transform(X_test)

Step 2. Fitting logistic regression to dataset

lr = LogisticRegression ()

lr.fit (X_train, y_train)

Step 3. Predicting the test set result

pred = lr. predict(X_test)

Step 4. Marking the confusion matrix

confusion_matrix (y_test, pred)

Mathematical Model

S = {I, P, O, DP, FS}

I. I(Input): Dataset.

II. P(Parameters):'ssc_p','hsc_p','degree_p','worke,'mba_p','etest_p','gender','degree_t','specialisation' and 'statuses.

III. Output: CM (Confusion Matrix)

In the proposed system for campus recruitment prediction is derived using Accuracy, Precision and Recall.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP +FP)

Recall = TP / (TP +FN)

IV. DP (Data Processing): We have used Standard scaling for data processing.

The Standard Scaler assumes your data is normally distributed within each feature and will scale them such that the distribution is now centered around 0, with a standard deviation of 1.

The mean and standard deviation are calculated for the feature and then the feature is scaled based on:

Xi – mean(x)/stdev(x)

VI. DATA AND RESULTS

	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000.0
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000.0
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000.0
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	NaN
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000.0

Fig 4- Dataset

The following set of data is considered as the base set for the proposed system. The dataset comprises of different qualitative and quantitative measures of 215 students. The dataset consists of the following attributes- Gender of student, SSC percent, SSC board, HSC percent, HSC board, HSC specialization, Degree percent, Degree type, work experience, Specialization, MBA percent, Status and Salary. From the above specified attributes, the attributes such as 'ssc_p','hsc_p','degree_p','workex','mba_p','etest_p','gender','degree_t','specialisation' and 'status' are taken into consideration. The combination of various attributes determines whether the candidate is recruited or not.

The quantitative aspects of 'ssc_p','hsc_p','degree_p', 'mba_p','etest', 'degree_t' and 'specialisation' form the major aspects for recruitment analysis. The qualitative skills like the previous work experience s helps the managers to know their sound knowledge in the field.

EXPERIMENTAL ANALYSIS

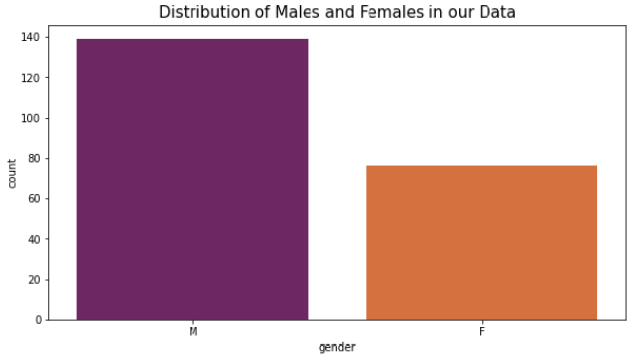

Fig 5- Distribution of males and females in our data

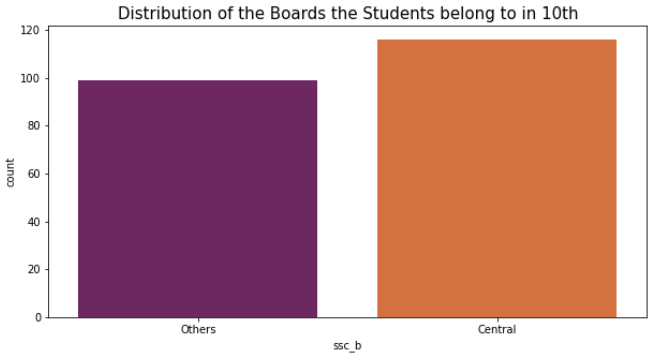

Fig 6 - Distribution of the Boards the Students belong to in 10th

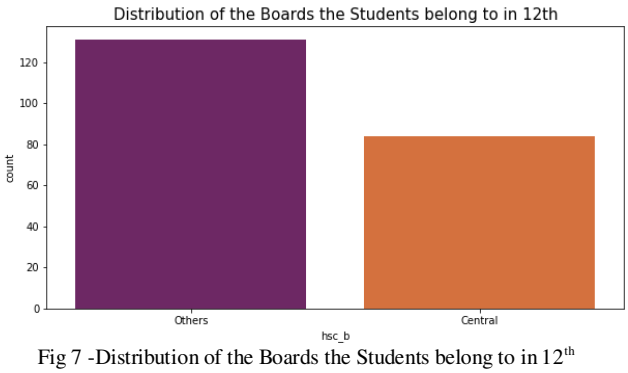

Fig 7 -Distribution of the Boards the Students belong to in 12th

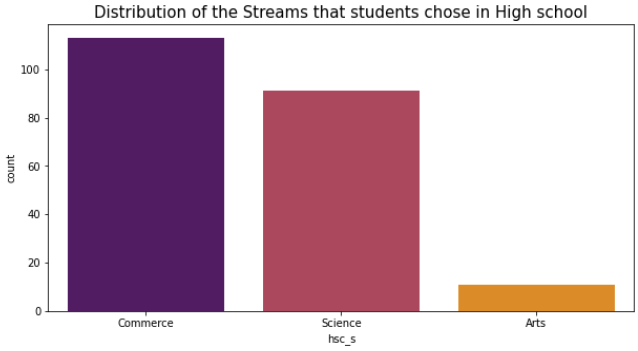

Fig 8- Distribution of the Streams that students chose in High school

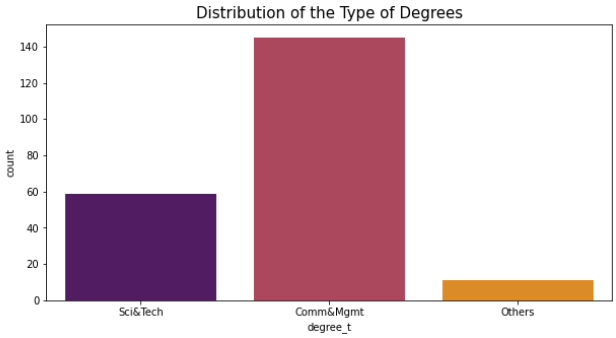

Fig 9 - Distribution of the Type of Degrees

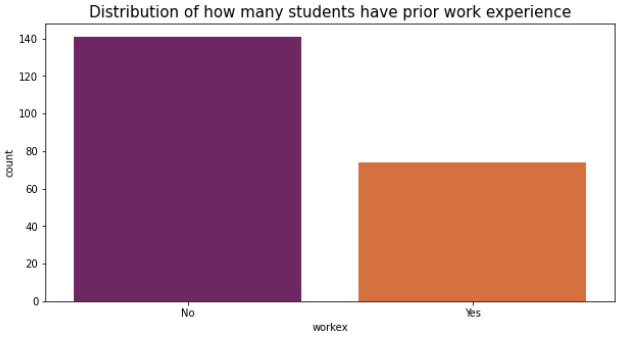

Fig 10 - Distribution of how many students have prior work experience

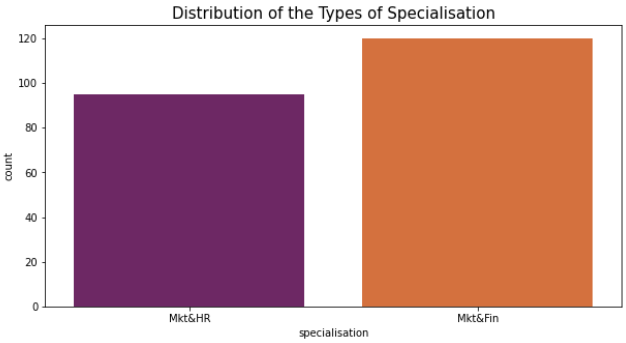

Fig 11 - Distribution of the Types of Specialisation

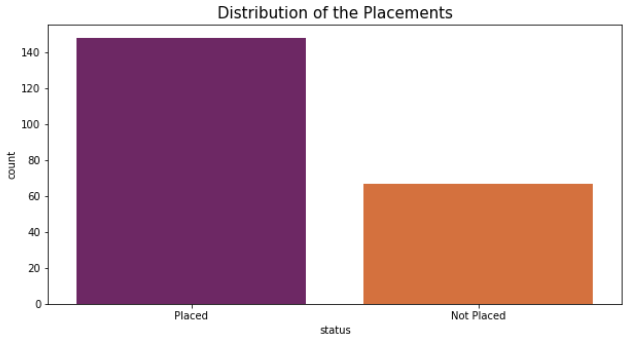

Fig 12 - Distribution of the Placements

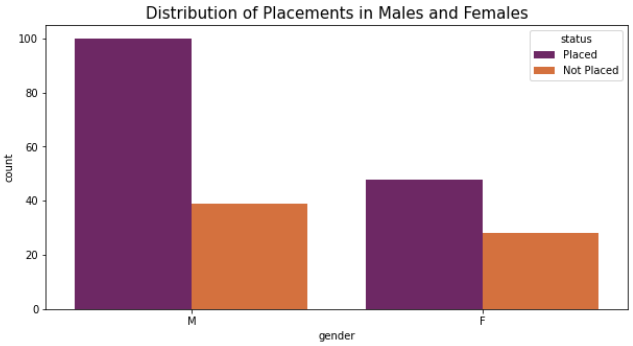

Fig 13 - Distribution of Placements in Males and Females

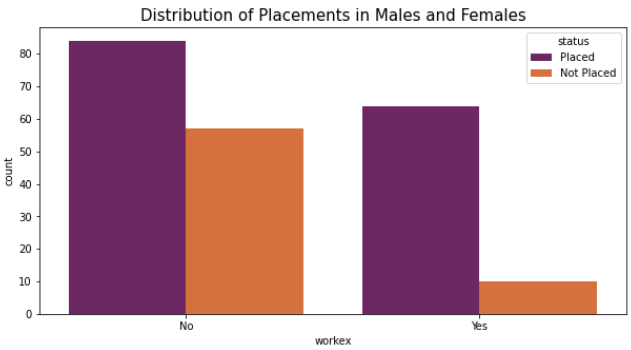

Fig 14 - Distribution of Placements in Males and Females with respect to w ork experience

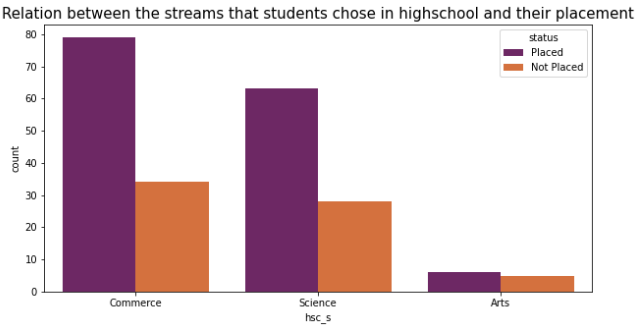

Fig 15 - Distribution of Placements in Males and Females with respect to HSC percentage

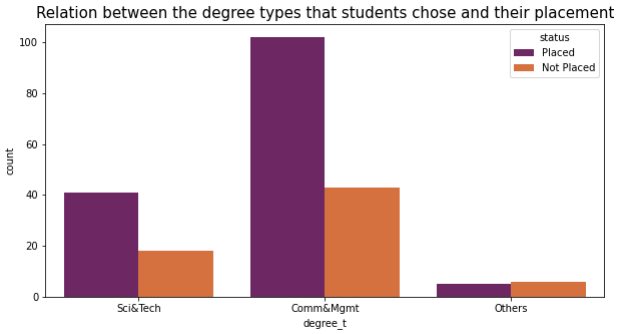

Fig 16- Relation between the streams that students chose in high school and their placement

VII. DISCUSSION

The tests were conducted with the help of standard placement dataset. The dataset was divided into training dataset and testing dataset. The dataset consists of the following attributes- Gender of student, SSC percent, SSC board, HSC percent, HSC board, HSC specialization, Degree percent, Degree type, work experience, Specialization, MBA percent, Status and Salary. From the above specified attributes, the attributes such as 'ssc_p','hsc_p','degree_p','workex','mba_p','etest_p','gender','degree_t','specialisation' and 'status' were applied to logistic regression algorithm to predict the possibility of the student getting placed. The obtained accuracy, precision, recall of logistic regression algorithm is as follows:

Accuracy: 88.37209302325581
Precision 89.28571428571429
Recall: 92.5925925925926

The result that we obtained from the Logistic regression was compared against the results obtained from KNN that gave an accuracy of around 74% and SVM with an accuracy of around 79%.

Table give the Confusion matrix for Logistic regression

		Predicted	
Actual		Placed	Not Placed
	Placed	13	3
	Not Placed	2	25

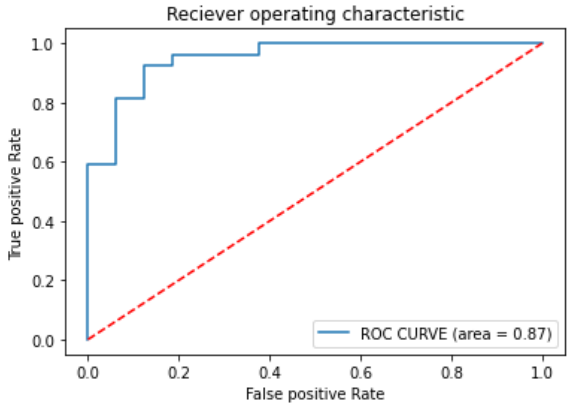

Fig 17 – ROC curve

VIII. FUTURE SCOPE

The above work was carried out with respect to the academics of the students like the SSC and HSC percentage, their specialization and the higher education stream. Further work can be carried out by using other skill sets of students like their programming knowledge, communication skills, aptitude skills etc. and by applying other algorithms that could lead to improvement in results. The key to this would be to identify the student's knowledge in the additional skills which also are important for getting placed in a core company for their respective branches.

CONCLUSION

Placement of students in final year is one of the main activities in the curriculum of final year students. Hence all institutions aim to improve their placement department. Based on the study of various papers we have concluded that each existing system have their own pros and cons. But the main thing the systems are lagging in are solutions or say suggestions to the students who have less probability of getting placed. Thus, to overcome this disadvantage of the existing systems we are aimed to develop a campus recruitment prediction system which will not only predict the possibility whether the student will get placed or not but also some solutions and aptitude tests to overcome their weaknesses and increase their chances of getting placed. This system is aimed to help the TPO of various institutions to get an idea about the students who have the possibility of getting placed and which students need help to improve their skills, and thus increase the placement percentage.

REFERENCES

[1] S.Taruna , Mrinal Pandey, "An Empirical Analysis of Classification Techniques for Predicting Academic Performance," IEEE International Advance Computing Conference (IACC) At: ITM ,Gurgoan. February 2014

[2] Vivek Anand, Saurav Kumar, Neelam Madheshwari,"Students result prediction using machine learning techniques" International journal of advance science and engineering, 2015.

[3] Souvik Hazra ,Satyaki Sanyal, "Recruitment Prediction Using ID3 Decision Tree," International Journal of Advance Engineering and Research Development Volume 3, Issue 10, October -2016.

[4] Professor. Ashok M Assistant Professor Apoorva A, "Data Mining Approach for Predicting Student and Institution's Placement Percentage", International Conference on Computational Systems and Information Systems for Sustainable Solutions,2016.

[5] Mohana Bangale, Shubham Bavane, Akshay Gunjal, Rohit Dandhare, Sudhir D. Salunkhe," A Survey on Placement prediction system using machine learning", International journal of advance scientific research and development (IJSART), Feburary 2019.

[6] B.K. Bharadwaj and S. Pal," Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security, Vol. 9, No. 4

[7] Shreyas Harinath , Aksha Prasad, Suma H S, Suraksha A, Tojo Mathew," Student placement prediction using machine learning", International Research Journal of Engineering and Technology (IRJET) Apr 2019 .

[8] Apoorva Rao R, Deeksha K C, Vishal Prajwal R, Vrushak K, Nandini, "Student Placement Analyzer: A Recommendation System Using Machine Learning", JARIIE-ISSN(O)-2395-4396

[9] Animesh Giri, M Vignesh V Bhagavath, Bysani Pruthvi, Naini Dubey," A Placement Prediction System Using K-Nearest Neighbors Classifier" IEEE Xplore, DOI: 10.1109/CCIP.2016.7802883, January 2017.

[10] Jay Torasakar, Rakesh Prabhu, Pranay Rambade, Manoj Kumar Shukla – "Logistic Regression Analysis as a Future Predictor" International Journal of Technical Research and Applications.

[11] Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor, Keshav Kumar,"PPS - Placement Prediction System using Logistic Regression" in 2014 IEEE International Conference on MOOC, Innovation and Technology in Education.

[12] S.Taruna , Mrinal Pandey ,"An Empirical Analysis of Classification Techniques for Predicting Academic Performance" in 2014 IEEE International Advance Computing Conference (IACC).

[13] Kotsiantis, Sotiris B., and Panayiotis E. Pintelas, "Predicting students marks in hellenic open university", in Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on, pp. 664-668. IEEE, 2005.

[14] Saha, Goutam, "Applying logistic regression model to the examination results data.,in Journal of Reliability and Statistical Studies 4, no.2(2011):1-13.

[15] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque and Rashedur M Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," in Decision Analytics (2015) 2:1 DOI 10.1186/s40165-014- 0010-2(Springer Journel).

[16] Chen, Joy Iong Zong, and S. Smys. "Social Multimedia Security and Suspicious Activity Detection in SDN using Hybrid

Deep Learning Technique." Journal of Information Technology 2, no. 02 (2020): 108- 115.

[17] Smys, S., Joy Iong Zong Chen, and Subarna Shakya. "Survey on Neural Network Architectures with Deep Learning." Journal of Soft Computing Paradigm (JSCP) 2, no. 03(2020): 186-194.