CAMPUS PLACEMENT PREDICTION USING ML ALGORITHMS IN R

Abha Marathe[1] Siddhesh Karande[2] Sohail Kasmani[3] Navnath Kate[4] Trupti Kale[5] Bhakti Karangale[6]
Undergraduate students, Department of computer engineering, Bansilal Ramnath Agarwal Charitable Trust's
Vishwakarma Institute of Technology
(An Autonomous Institute affiliated to Savitribai Phule Pune University)

Abstract :

A placement predictor will be created to determine the likelihood of a student getting placed in a firm based on the company's criteria. The placement predictor takes into account a number of factors that may be used to determine a student's ability level. Some criteria are derived from university assessments, while others are derived from tests completed within the placement management system. The predictor uses these data points to properly estimate whether or not the student will be placed in a firm. The predictor is trained using data from previous pupils.

However, the challenge was to create a classification algorithm that could accomplish the job with the highest level of accuracy for our data set. Depending on the sort of problem to answer and the data set to work with, different algorithms have varying degrees of accuracy. As a result, we chose four algorithms: KNN, SVM, Logistic Regression, Decision tree and Random Forest, and compared the accuracy levels of each method with regard to our problem and data set. The results of this test will aid us in deciding which method to employ when putting our predictor into the placement management system.

Keywords— Prediction, SVM, KNN, Logistic Regression, Decision tree, Random Forest

I. INTRODUCTION

As part of developing a placement management system at the college level, we want to create a placement predictor that forecasts the likelihood of students being placed and assists them in improving their abilities before the recruiting process begins. For the placement prediction, we are employing machine learning. To categorise students into relevant clusters, we use K-nearest neighbour (KNN), Support Vector Machine (SVM), Logistic Regression, Decision tree and Random Forest. With the comparison of various machine learning approaches, both recruiters and students will benefit from placements and associated activities.

We apply machine learning approaches to predict students' placement status based on a dataset in this study. HSC scores, SSC scores, graduation percentage, etest percentage and MBA percentage are the parameters in the dataset that are taken into account for the forecast. Machine learning is used to forecast placement using Logical Regression, Decision tree, Random Forest, KNN, and SVM.

Based on the dataset, we built dataframe. AnyNA(dataframe) is responsible for handling null data fields. We utilise dplyr, caTools, caret, randomForest which are more powerful libraries. We also use sample.split to create training and test sets from the dataset. It forecasts each student's placement using the appropriate method, and the accuracy may be shown in the confusion matrix.

II. LITERATURE REVIEW

In the paper "Campus Placement Predictive Analysis using Machine Learning" [1],the author focused on the planning of understudies' exhibition design using an information mining technique and a contextual research from five engineering departments in this article. To find patterns in the data that are connected with the prediction model, machine learning methods such as KNN (K-Nearest Neighbours), SVM (Support Vector Machine), and ANN (Artificial Neural Network) were used. Technical events attended, internships done, certification courses finished, and CGPA gained were among the qualities chosen. The kids' performance was divided into three categories, allowing them to be put in Dream, Super Dream, or MRC.

In the paper ""Predictive Analysis Tool for Predicting Student Performance and Placement

Performance using ML algorithms",[2] highlights on the mapping of students' performance pattern using a data mining approach with the help of a case study from five engineering departments. The data set used for is further splitted into two sets consisting of two third as training set and one third as testing set. The dataset collected consist of over 1000 instances of students. In this regard to improve the student's performance, a work has been analyzed and predicted using the classification algorithms Decision Tree and the Random forest algorithm to validate the approaches. The accuracy obtained after analysis for Decision tree is 84% and for the Random Forest is 86%. H [2]

In the paper "STUDENT PLACEMENT CHANCE PREDICTION"[3] Logical Regression, Random Forest, KNN, and SVM are used to complete the scenario expectation. SVM has a precision of one hundred percent. On the given dataset, Logistic Regression also performs well, with an accuracy of 97.59. SVM, Logistic Regression, Random Forest, and KNN are all good for binary classification tasks, as shown by the results, since they all yield accuracy of over 95 percent.

In "Placement Prediction using Various Machine Learning Models and their Efficiency Comparison"[4] the software method used to forecasts placement status in five categories: ideal company, core company, mass recruiter, ineligible, and not interested in placements. All data is monitored using the Nave Bayes technique. This study describes a method for developing a placement prediction model for pre-final year engineering graduate students using a deep learning algorithm. The student dataset combining academic and placement records is a viable source for forecasting future placement prospects, according to the study..

The suggested framework in "Campus Placement Prediction Using Supervised Machine Learning Techniques"[5]research includes a few machine learning computations such as KNN (k-Nearest Neighbor), Linear Regression (LR1), Logistical Regression (LR2), Lasso Regression (LR3), and Decision Tree (DT) calculations. Linear Regression is the highest performing method for forecasting student performance, according to the data. It exceeded all other approaches, with a 94 percent accuracy rate. Only seven of the twenty-four characteristics in our sample had an effect on the CGPA.

In "Student Placement Prediction Model: A Data Mining Perspective for Outcome-Based Education System"[6],paper, the proposed system helps in predicting whether a student will get placement or not. This system can also be helpful for identifying the areas where student needs to work on for placement. This system uses student's details like academic marks, coding skills, etc. This paper describes the creation of a indicator for placement utilizing the concept of predictive modeling, called Logistic Regression.

III. METHODOLOGY

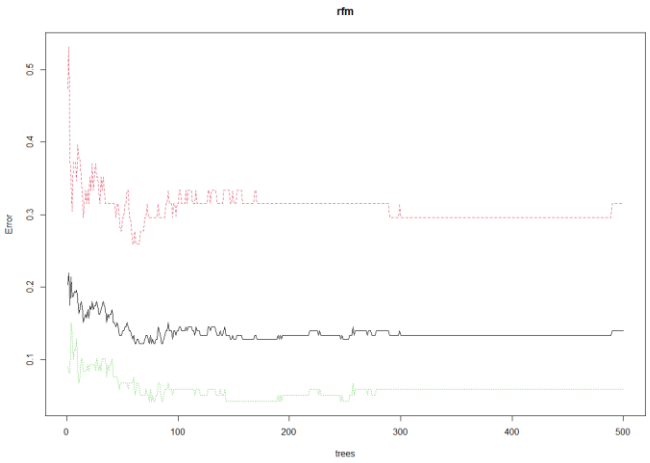
Random Forest :

SVM, Logistic regression, decision trees, and the Naive Bayes classifier, to mention a few, are among the many classification techniques at our disposal. The Random Forest Classifier, on the other hand, is near the top of the hierarchy of classifiers. We'll look into how decision trees function because the random forest classifier is made up of them.

It is mainly a flowchart-like structure in which each node (except the leaf node) represents a feature test (i.e., what will happen if some activity, such as flipping a coin, is performed), leaf nodes represent the class label (the decision made after all features have been computed), and branches represent the feature conjunctions that lead to those class labels. The pathways from the root node to the leaf node are the categorization rules of a decision tree.

So, let's have a look at random forest classifiers now. It is, as previously said, a collection of decision trees."The wisdom of the multitude" is the core concept behind random forest. It's a strong notion in which a huge number of unrelated models, in this instance trees, working together provide a considerably more reliable result than any of the individual models. So, by training we get the accuracy of 93.02326 % for that we set the seed value as 123 with split ratio of 0.80.

So, in a random forest, each individual tree would try to identify a suitable class label for the problem using



multiple attributes and classification criteria. Each tree would respond in its own way. Within the random forest, voting is conducted to determine which class label earned the most votes. This results in a more accurate model for predicting class labels.

Decision tree:

Data preprocessing for decision tree:

We remove all the unnecessary feature from our dataset and convert our categorical variables into factors.

Decision tree working:

Decision tree Algorithm comes under supervised machine learning .It is used mainly for classification and regression purposes .It uses a flowchart like tree structure to make prediction from a series of feature based split. It starts with the root node and ends on the leaf node(i.e predicted output variable).in decision tree the selection of features for the node depends on the information gain from that particular feature which depends on the entropy which calculates the impurity of the node .So feature selection and their order is a crucial part in decision tree .So after pre-processing we randomly divide our dataset into training and testing data to avoid class imbalance problem .After that we select our root node as a feature from which we get the highest information gain by using the impurity of the node . For improving the accuracy we can change the splitting ration and also changing the order in which we split the data set

KNN

Data Preprocessing for KNN :-

We have normalized the discrete variables in our dataset in order to reduce the noise.

KNN Working :-

KNN is a supervised machine learning algorithm , which is useful to solve classification as well as regression problems. Basically, it works by calculating the distance between the two data points. we calculate the distance of the data point from its k neighbors. Then we classify it to the class to which it's most number of neighbors belongs to from k. We calculate the accuracy by changing the value of K. k is the number of neighbors.

After preprocessing the data, we have divided the data into training and testing. We have randomly divided the data into training and testing to avoid the problem of class inbalancing .At first, we have considered only HSC and SSC percentage features to train the data.

SVM

Support Vector Machine or SVM is one of the most famous Supervised Learning calculations, which is utilized for Classification just as Regression issues. Be that as it may, fundamentally, it is utilized for Classification issues in Machine Learning.

The objective of the SVM calculation is to make the best line or choice limit that can isolate n-layered space into classes so we can undoubtedly put the new item in the right classification later on. This best choice limit is known as a hyperplane.

SVM picks the extreme points/vectors that assist in making the hyperplane. These outrageous cases are called as support vectors, and subsequently calculation is named as Support Vector Machine.

Logistic Regression-

Logistic regression is a measurable strategy for investigating a dataset in which there are at least one independent variables that decide a result. The result is estimated with a dichotomous variable (where there are just two potential results). Logistic Regression works by utilizing the Sigmoid function

to plan the forecasts to the result probabilities. This function is a S-formed bend that plots the anticipated qualities somewhere in the range of 0 and 1. The qualities are then plotted towards the edges at the top and the lower part of the Y-axis, involving 0 and 1 as the labels. Then, at that point, depending upon these qualities, the independent variables can be ordered. This is the way the Sigmoid function resembles:
The Sigmoid function depends on the accompanying condition:

y=1/(1+e^x)

Where e^x= the exponential consistent with a worth of 2.718.The Sigmoid function's condition above offers the anticipated value(y) as zero assuming x is viewed as negative. In the event that x is a huge positive number, the value anticipated is near one.

IV. RESULT:

Random forest:

	precision	recall	f1-score	Support
Class0	0.77	0.77	0.87	13
Class1	1.00	1.00	0.95	30
Accuracy			0.93	43
Macro avg	0.88	0.88	0.91	43
Weighted avg	0.93	0.93	0.93	43

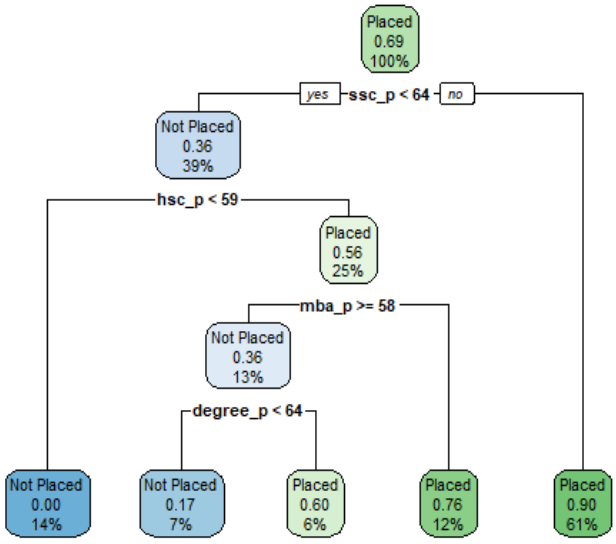
Observation table of Random forest

Decision Tree:

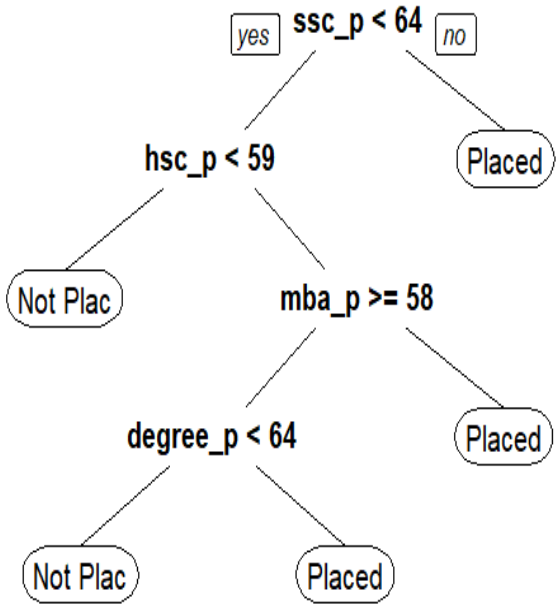
Splitting ratio	Accuracy(%)	Sensitivity(%)	Specificity(%)
0.8	86.05	61.54	96.67
0.75	81.48	76.47	83.78
0.85	78.12	70.00	81.82
0.88	73.08	87.50	66.67

Observation table of decision tree

We can also plot our decision tree using rpart.plot() or by prp



Decision tree plot



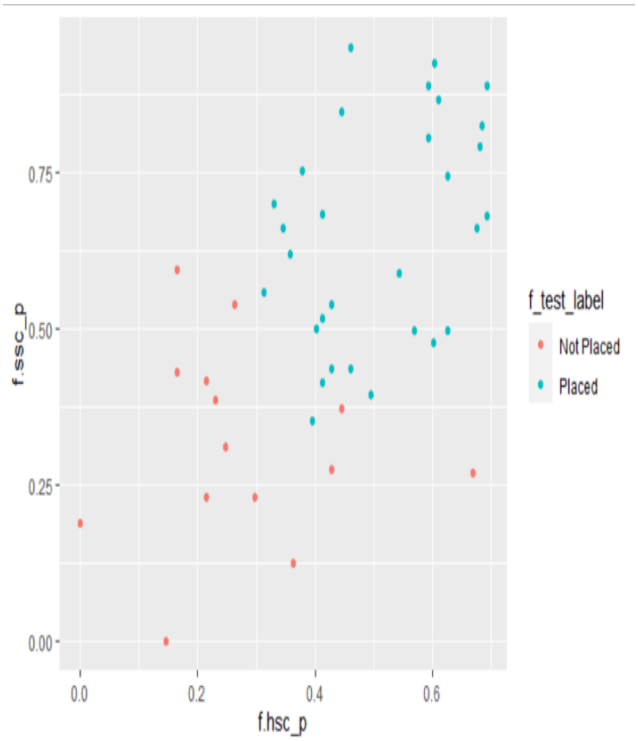
Decision tree plot

So as you can see in the above image root node is ssc_p with leaf node as our target variable that is status of student whether he/she is placed or not .As you can see in the above decision tree not all other variable are considered to make our decision tree because decision tree is a non-parametric which means parameters which are to be considered depends on the splitting ratio as well the parameters considered changes with the splitting ratio and also with randomness with which we select our data for training

KNN:

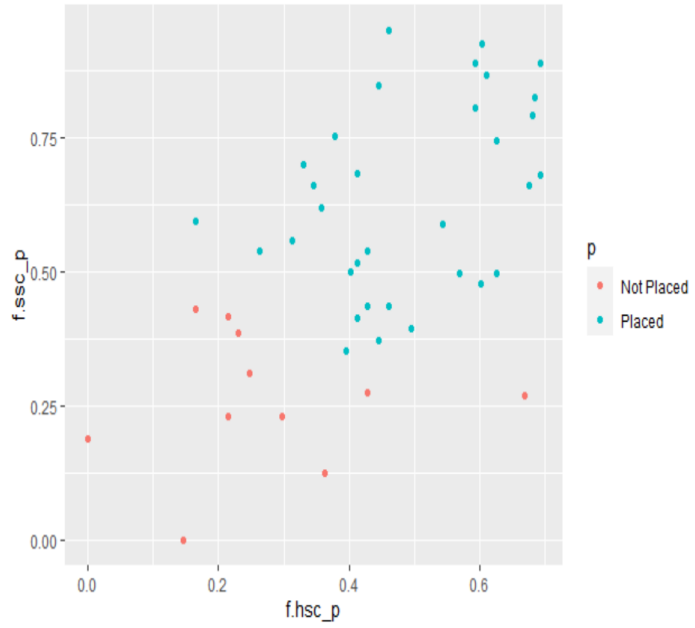
K	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	82.22222	87.09677	71.42857
2	86.66667	90.32258	78.57143
3	86.66667	93.54839	71.42857
4	82.22222	87.09677	71.42857
5	84.44444	87.09677	78.57143
6	88.88889	93.54839	78.57143
7	88.88889	93.54839	78.57143
8	88.88889	93.54839	78.57143
9	93.33333	100	78.57143
10	91.11111	96.77419	78.57143
11	93.33333	100	78.57143

Observation table of KNN



SSC_P Vs HSC_P

This graph shows the actual data points



HSC_P Vs SSC_P

These graph shows the predicted data points by KNN.

From the above two graphs we can compare the actual placement and prediction by KNN.

SVM:

SVM gives the highest accuracy of 95.35% on ideal splitting of our data. The given below is the confusion matrix for 80-20.

Confusion matrix:

test_pred	Not Placed	Placed
Not Placed	6	0
Placed	2	35

Confusion matrix of SVM

Observation:

Splitting Ratio (%)	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)
90-10	95.24	66.67	100	100
80-20	95.35	75	100	100
70-30	90.62	82.35	93.62	82.35
60-40	87.21	76	91.80	79.17
50-50	86.92	78.12	90.67	78.12

Observation table of SVM

Logistic Regression:

Logistic regression gives the accuracy of 93.48% on 80-20(80 % training data and 20% testing). The given below is the confusion matrix of the same.

Confusion matrix:

test_pred	Not Placed	Placed
Not Placed	10	1
Placed	2	33

Confusion matrix of LR

Observation:

Splitting Ratio (%)	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)
90-10	90	75	95.45	85.71
80-20	93.48	83.33	97.06	90.91
70-30	90.91	81.82	94.54	85.71
60-40	86.02	77.78	89.39	75
50-50	83.48	75.86	86.25	66.67

Observation table of LR

Sample dataset :

sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisat	mba_p	status	salary
1	M	67	Others	91	Others	Commerce	58	Sci&Tech	No	55	Mkt&HR	58.8	1	270000
2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	1	200000
3	M	65	Central	68	Central	Arts	64	Comm&M	No	75	Mkt&Fin	57.8	1	250000
4	M	56	Central	52	Central	Science	52	Sci&Tech	No	66	Mkt&HR	59.43	0	288655
5	M	85.8	Central	73.6	Central	Commerce	73.3	Comm&M	No	96.8	Mkt&Fin	55.5	1	425000
6	M	55	Others	49.8	Others	Science	67.25	Sci&Tech	Yes	55	Mkt&HR	51.58	0	288655
7	F	46	Others	49.2	Others	Commerce	79	Comm&M	No	74.28	Mkt&Fin	53.29	0	288655
8	M	82	Central	64	Central	Science	66	Sci&Tech	Yes	67	Mkt&Fin	62.14	1	252000
9	M	73	Central	79	Central	Commerce	72	Comm&M	No	91.34	Mkt&Fin	61.29	1	231000
10	M	58	Central	70	Central	Commerce	61	Comm&M	No	54	Mkt&Fin	52.21	0	288655
11	M	58	Central	61	Central	Commerce	60	Comm&M	Yes	62	Mkt&HR	60.85	1	260000
12	M	69.6	Central	68.4	Central	Commerce	78.3	Comm&M	Yes	60	Mkt&Fin	63.7	1	250000
13	F	47	Central	55	Others	Science	65	Comm&M	No	62	Mkt&HR	65.04	0	288655
14	F	77	Central	87	Central	Commerce	59	Comm&M	No	68	Mkt&Fin	68.63	1	218000
15	M	62	Central	47	Central	Commerce	50	Comm&M	No	76	Mkt&HR	54.96	0	288655
16	F	65	Central	75	Central	Commerce	69	Comm&M	Yes	72	Mkt&Fin	64.66	1	200000
17	M	63	Central	66.2	Central	Commerce	65.6	Comm&M	Yes	60	Mkt&Fin	62.54	1	300000
18	F	55	Central	67	Central	Commerce	64	Comm&M	No	60	Mkt&HR	67.28	0	288655
19	F	63	Central	66	Central	Commerce	64	Comm&M	No	68	Mkt&HR	64.08	0	288655
20	M	60	Others	67	Others	Arts	70	Comm&M	Yes	50.48	Mkt&HR	77.89	1	236000
21	M	62	Others	65	Others	Commerce	66	Comm&M	No	50	Mkt&HR	56.7	1	265000
22	F	79	Others	76	Others	Commerce	85	Comm&M	No	95	Mkt&Fin	69.06	1	393000
23	F	69.8	Others	60.8	Others	Science	72.23	Sci&Tech	No	55.53	Mkt&HR	68.81	1	360000
24	F	77.4	Others	60	Others	Science	64.74	Sci&Tech	Yes	92	Mkt&Fin	63.62	1	300000

Dataset

Our dataset consists of 14 variables and two class(placed or Not placed).

The Correlation matrix is given below:-

```
> cor(p[, unlist(lapply(p, is.numeric))])
              sl_no      gender       X10_p       PU_p
sl_no     1.00000000  0.07430551 -0.07815472 -0.08571138
gender    0.07430551  1.00000000 -0.06896861 -0.02133391
X10_p    -0.07815472 -0.06896861  1.00000000  0.51147210
PU_p     -0.08571138 -0.02133391  0.51147210  1.00000000
degree_p -0.08828133 -0.17321662  0.53840400  0.43420581
workex    0.05915108  0.08515260  0.17567548  0.14102486
etest_p   0.06363593  0.08429352  0.26199269  0.24511293
mba_p     0.02232720 -0.30053122  0.38847755  0.35482260
status   -0.02685942  0.09067040  0.60788873  0.49122794
            degree_p     workex    etest_p       mba_p
sl_no    -0.08828133 0.05915108 0.06363593  0.02232720
gender   -0.17321662 0.08515260 0.08429352 -0.30053122
X10_p     0.53840400 0.17567548 0.26199269  0.38847755
PU_p      0.43420581 0.14102486 0.24511293  0.35482260
degree_p  1.00000000 0.12264755 0.22447017  0.40236377
workex    0.12264755 1.00000000 0.05673469  0.16881149
etest_p   0.22447017 0.05673469 1.00000000  0.21805467
mba_p     0.40236377 0.16881149 0.21805467  1.00000000
status    0.47986099 0.27605996 0.12763937  0.07692165
              status
sl_no    -0.02685942
gender    0.09067040
X10_p     0.60788873
PU_p      0.49122794
degree_p  0.47986099
workex    0.27605996
etest_p   0.12763937
mba_p     0.07692165
status    1.00000000
```

Covariance matrix is:-

```
> cov(p[, unlist(lapply(p, is.numeric))])
              sl_no      gender       X10_p        PU_p
sl_no    3870.0000000  2.21495327 -52.6413551 -58.1060280
gender      2.2149533  0.22960226  -0.3578129  -0.1114001
X10_p     -52.6413551 -0.35781287 117.2283767  60.3483733
PU_p      -58.1060280 -0.11140013  60.3483733 118.7557058
degree_p  -40.4136449 -0.61077505  42.8971367  34.8198200
workex      1.7523364  0.01943056   0.9057885   0.7318502
etest_p    52.5561682  0.53622626  37.6592247  35.4616777
mba_p       8.1023364 -0.84003673  24.5359520  22.5558460
status     -0.7757009  0.02016953   3.0555023   2.4851491
            degree_p     workex    etest_p      mba_p
sl_no    -40.4136449 1.75233645  52.5561682  8.1023364
gender    -0.6107750 0.01943056   0.5362263 -0.8400367
X10_p     42.8971367 0.90578852  37.6592247 24.5359520
PU_p      34.8198200 0.73185025  35.4616777 22.5558460
degree_p  54.1511028 0.42979548  21.9294686 17.2720200
workex     0.4297955 0.22677679   0.3586855  0.4689450
etest_p   21.9294686 0.35868550 176.2510184 16.8869734
mba_p     17.2720200 0.46894501  16.8869734 34.0283757
status     1.6393106 0.06103021   0.7866701  0.2083106
              status
sl_no    -0.77570093
gender    0.02016953
X10_p     3.05550228
PU_p      2.48514910
degree_p  1.63931058
workex    0.06103021
etest_p   0.78667007
mba_p     0.20831058
status    0.21551837
```

V. CONCLUSION :

A placement prediction algorithm forecasts final-year B-Tech students' placement status. In the RStudio environment, multiple machine learning methods are employed for data analysis and prediction. The accuracy of several algorithms is examined above. SVM clearly has a precision of 95.35%(80-20)The accuracy of Logistic Regression is likewise good, with a score of 93.48%based on the dataset. Machine learning algorithms' accuracy may vary depending on the dataset. The results of our investigation show that SVM, Logistic Regression, Random Forest, and KNN are all suitable for binary classification issues, since they all have an accuracy of over 90-95%percent.

VI. REFERENCES:

[1] Nikhil Kumar, Thirunavukkarasu K., Ajay Shanker Singh, E. Rajesh, "Campus Placement Predictive Analysis using Machine Learning", 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020

[2] Dr. B. Muthusenthil , Venkat Mugesh S , Thansh D , Subaash R, "Predictive Analysis Tool for Predicting Student Performance and Placement Performance using ML algorithms", 2020

[3] Mrs. J. Samatha, D. Manjusha, B. Pooja, A. Usha, "STUDENT PLACEMENT CHANCE PREDICTION", May 2020

[4] Irene Treesa Jose , Daibin Raju , Jeebu Abraham Aniyankunju, Joel James, Mereen Thomas Vadakkel," Placement Prediction using Various Machine Learning Models and their Efficiency Comparison", International Journal of Innovative Science and Research Technology, 2020

[5] Pothuganti Manvitha and Neelam Swaroop, "Campus Placement Prediction Using Supervised Machine Learning Techniques", 2019.

[6] Abhishek S. Rao, Aruna Kumar S V, Pranav Jogi, Chinthan Bhat K, Kuladeep Kumar B, Prashanth Gouda, "Student Placement Prediction Model: A Data Mining Perspective for Outcome-Based Education System", International Journal of Recent Technology and Engineering (IJRTE), 2019.