An Elaborative and Exploratory Data Analysis of Placement Data

Abhir Mahajan [1], Amarendra Pratap Singh [2], Anas Siddiqui [3], Piyush Srivastava [4], Diksha Daani [5]
1, 2, 3, 4, B. Tech, Dept. of CSE, IPEC Ghaziabad, UP, India
5, Professor, Dept. of CSE, IPEC Ghaziabad, UP, India

Abstract: Traditionally, data analysis was a trial and error-based approach, which made it impossible to compute when heterogeneous and large set of data was provided. Machine Learning being a part of data science has provided better substitutes for exploring large chunks of data by growing rapid and well-organized algorithms and various data-driven models for actual-time processing of data. Campus placement plays an important role in helping each and every educational institution to help the students in achieving their goals. Large scales of information can be extracted from the student's dataset using tools like Data Mining. Various Data Mining models are used to discover knowledge from large datasets containing valuable information. In this paper, we have designed a model, which suggests factors on which the placements depend which may help the students for placement preparation.

Linear Regression, a tool of supervised machine learning can contribute its features to serve the purpose. It is an effective method helping in predicting future trend of student placement based on gender, degree percentage, school percentage, employability test marks and few more features. The result obtained from this will help the students to better understand their weak areas to work upon. Working on these areas will let students achieve better placements in an institution.

So, we decided to select three algorithms, namely Decision Trees, Random Forest, and Logistic Regression and to compare the accuracy levels of each of these algorithms, with respect to our problem and data set. The result of this test would help us in determining which algorithm to use while implementing our model on the placement data set.

Placements are given utmost importance in the present scenario. Our model can help to ensure better placement prospects for students, which would in turn help the students in their careers and to self-estimate where they are lagging in their preparations.

Keywords: *Binary Classification, Placement Factors, Salary Factors, NAN (Not a Number) Values, Pre-Processing, Feature selection, Accuracy.*

I. INTRODUCTION

College days get cohesive as students pioneer on to the road to their Higher Education degree programs.

Educational institutions play an important role in placements to prepare and guide the students by providing them the needed training for getting placed in high-end companies through the process of placements. The different platforms are provided by the placement cell to students so that they can showcase their skills and abilities to the prospective employers. Students can learn how to put forth their knowledge and abilities in the right way ,with the proper placement training, to fetch the best of jobs. The placement cell could opt to analyse their students graph based on performance in placement practice tests before presenting them in front of companies. This graph will be generated on the final year student dataset and this will help students realise which field they should work upon more. This method will be effective for both- the students as well as the institution for holding up a good placement record.

Model which will help placement cell of college to suggest the factors which will play the most important role in the placement of a particular candidate and helps in uplifting their skills before the recruitment process starts. We are using machine learning for our placement predictor model. We use several machine learning algorithms like Decision tree, Logistic Regression, Random Forest to classify students into appropriate clusters and the result would help them in understanding the factors which are most important for their placement so that they can improve their profile. And accuracy of respected algorithms are noted and With the comparison of various machine learning techniques, this would help both placement cell as well as students during placements and related activities. In this paper we have used machine learning techniques to suggest various factors playing a major role in getting students placed based on a dataset. The parameters in the dataset considered for the prediction are high school percentage ,high school board, specialization, degree percentage, degree specialization, work experience and employability test scores. We have done our placement prediction by using various machine learning algorithms like Logistic Regression, Random Forest and Decision Trees. The machine learning algorithm data frame is created using the pandas library based on the above sample dataset. The null data fields are handled by NAN values. We have imported train_test_split present in sklearn for creating training and testing sets from the dataset.

II. METHODOLOGY

Data Analysis can be done using various machine learning algorithms available in Python's Scikit Learn Package. Each algorithm has its own set of advantages and disadvantages. The advantages and disadvantages of an algorithm depends on the type of problem we are handling and the data set we are using. Following are three algorithms were used for the project along with a detailed explanation of what they are.

1. *Decision Trees*

This is a type of Supervised Learning algorithm that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree structured classifier, where internal nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the outcome.

In a decision tree, there are 2 nodes, which are decision node and leaf node. Decision nodes are used to make any decisions and have multiple branches, whereas leaf nodes are the output of those decisions and do not contain any further branches.

⇒ Advantages:

- Decision Trees usually mimic human thinking ability while making a decision, hence it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree like structure.

⇒ Disadvantages:

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue which can be resolved using the random forest algorithm.
- For more class labels, the computational complexity of the decision tree may increase.

2. *Random Forests*

Random forest is the popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both classification and regression problems in machine learning. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

It is a classifier that contains a number of decision trees on various subsets of the given data set and takes the average to improve the predictive accuracy of that data set. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater the number of trees in the forest leads to the higher accuracy and prevents the problem of overfitting.

⇒ Advantages:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

⇒ Disadvantages:

- Although random forest can be used for both classification and regression tasks, it is not more suitable for regression tasks.

3. *Logistic Regression*

Logistic regression is one of the most popular machine learning algorithms, which comes under the supervised learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either yes or no, zero or one, true or false, etc. But instead of giving the exact value as zero and one, it gives the probabilistic values which lie between 0 and 1.

Logistic regression is much similar to the linear regression except that how they are used. Linear regression is used for solving regression problems, whereas logistic regression is used for solving the classification problems. Next line in logistic regression, instead of fitting a regression line, we fit a "S" shaped logistic function which predicts to well maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its wait a comma etc.

Logistic regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete data sets.

Logistic regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

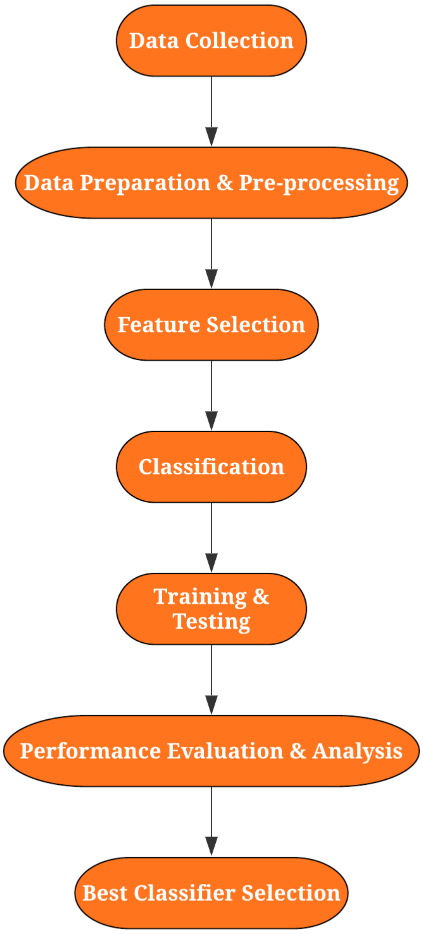
⇒ Advantages:

- Logistic regression is easier to implement, interpret and very efficient to train.
- It makes no assumptions about distributions of classes in feature space.
- It can easily extend to multiple classes and natural probabilistic view of class predictions.
- It is very fast at classifying unknown records.
- It can interpret model coefficients as indicators of feature importance.

⇒ Disadvantages:

- If the number of observations is lesser than the number of features logistic regulation should not be used otherwise it may lead to overfitting.
- It constructs linear boundaries.
- The major limitation of logistic regression is the assumption of linearity between the dependent variable and the independent variables.
- It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as neural networks can easily outperform this algorithm.

The present study focused on real-time institutional data collected from the past performance data of the student in college. After determining various factors which could have a major impact on placement prediction was identified and further selected for the study. Major factors identified for the study were related to academic details, gender and employability test result. The collected data were then pre-processed for quality data extraction. Later, the best suitable features for the present study were identified and selected for further analysis. The figure displays the system architecture of the proposed model.


Data Collection
Data Preparation & Pre-processing
Feature Selection
Classification
Training & Testing
Performance Evaluation & Analysis
Best Classifier Selection

a. Data Collection

Details of various candidates were collected such as Name, Gender, class 10th and 12th percentages along with their boards of education, graduation and post-graduation percentages and many more. From the institution, details of candidates like whether the candidates are placed or not placed and what salary they are getting and their degree percentage and their course specialization. Placement exams like elitmus scores were also collected.

b. Data Preparation and Pre-Processing

Data preparation is a step in a data analysis process in which data is cleaned from one or more sources, transformed and enriched to improve the quality of data prior to its use. The collected data was pre-processed to fill the missing data and is made compatible for further processing.

c. Feature Selection

The data containing fourteen features was further reduced to more significant features with the help of the dimensionality reduction approach. This step was carried out with the help of feature scaling which standardizes. The range of independent variables or features of the data.

d. Classification Methods

The final goal of the classification technique was to deploy NAN values to the non-placed members and remove those features that are distorted to a much greater extent.

Among the various classification algorithm items, the proposed work utilizes random forest and decision trees which are considered to be the best-suited algorithms for the present study of classification. A total of 215 data samples were collected from various departments of the institution. The one which has a maximum distance from the data points was chosen for better classification in the case of multiple hyperplanes. The dimension of the hyperplane depends on the total number of features. In the proposed study, the linear regression algorithm used various mathematical functions (kernels) like radial basis function (RBF), linear, sigmoid, and polynomial. Further, experimental investigation on various kernels was applied to the data samples and results were compared based on the performance metrics to identify the best kernel to classify the data.

e. Training and Testing

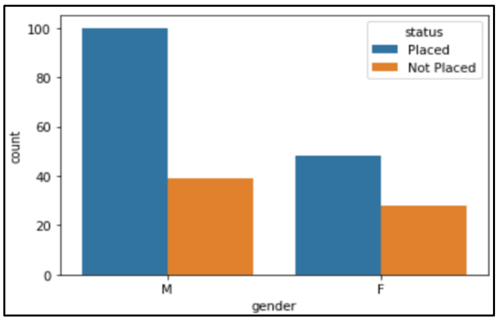
For better model validation, the dataset in the present study was split two times into training and testing with the help of SciKit Learn library for better evaluation of the data. Different proportions like 80:20 and 60:40 ratios were made and utilized for the study. A proportion of 80:20 signifies that 80% data is considered as training data and rest 20% data is considered as testing data.

f. Performance Evaluation and Analysis

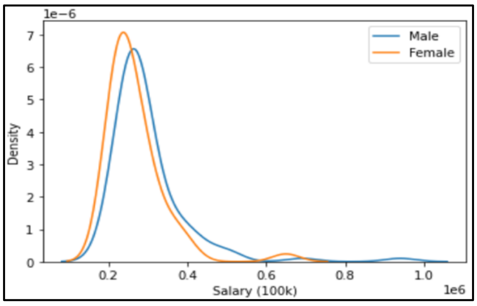
In this part, we are using algorithms like decision tree and random forest to get the importance of the every individual feature present and how much are these Features affecting the placement of the students.

g. Experiments and Results

Feature: Gender

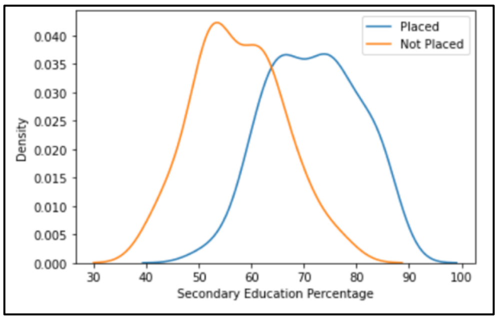


- We have samples of 76 Female students and 139 Male students.
- 30 Female and 40 Male students are unplaced.
- Female students have a comparatively lower number of placements.
- Females have greater chances of getting placed than males.

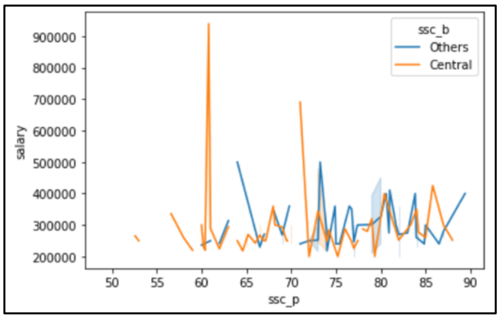


- More outliers on Male students show that males are getting higher CTC jobs.
- Male students are offered a slightly greater salary than females on average.

Feature: Senior Secondary Education Percentage

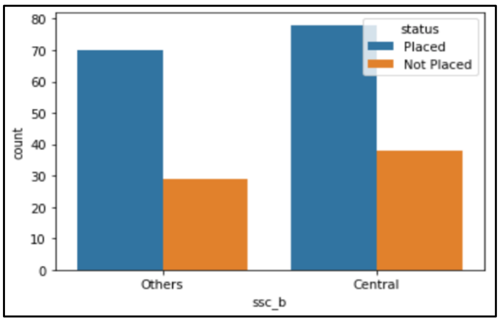


- All students with Secondary Education Percentage above 90% are placed.
- Almost all students with Secondary Education Percentage below 50% are not-placed
- Students with a good Secondary Education Percentage are placed on average.



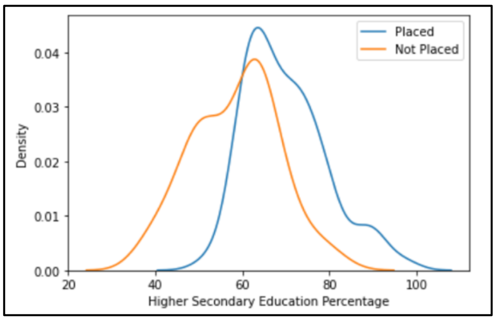
- No specific correlation (pattern) between senior secondary education percentage and salary of placed student.
- Board of education is not affecting salary.

Feature: Senior Secondary Education Board

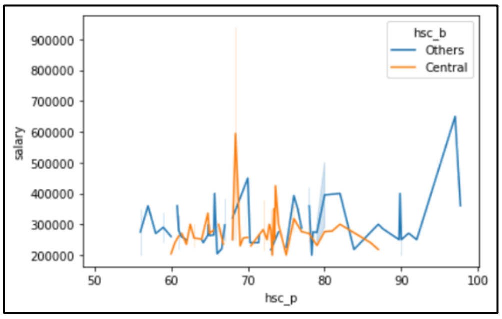


- Board of secondary education is not affecting the placement of a candidate significantly.

Feature: Higher Secondary Education Percentage

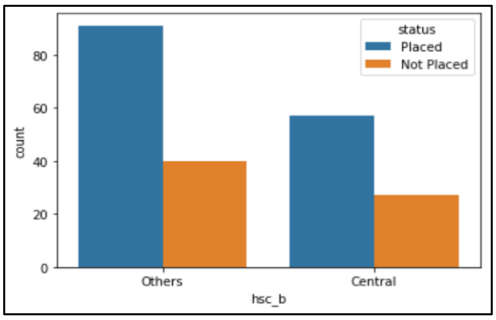


- More students are placed with higher secondary education percentage above 60%.
- Straight abrupt drop below 60% higher secondary education percentage.
- Higher Secondary Education percentage must be at least 60% for better chances of placement.



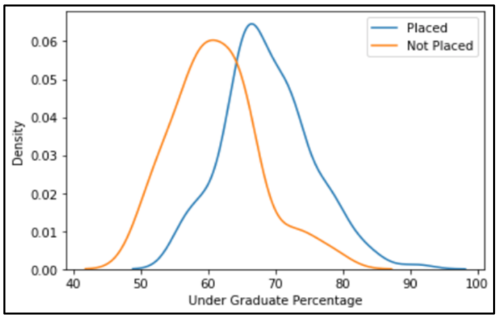
- High salary from both central and other boards of education.
- Again, no correlation (pattern) between higher secondary education percentage and salary of placed students.

Feature: Higher Secondary Education Board



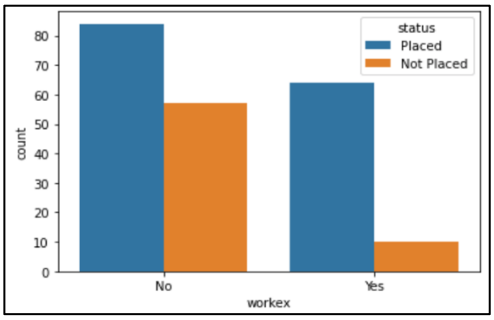
- Board of education is not affecting the placement of a candidate significantly.

Feature: Under Graduation Percentage

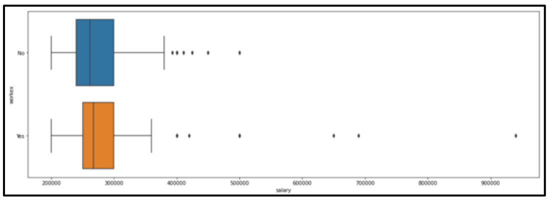


- All students with under graduation percentage 65% and above are placed.
- A steep drop is found below 65% showing less number of placements.
- Under graduation percentage should be at least 50% to be placed.

Feature: Work Experience

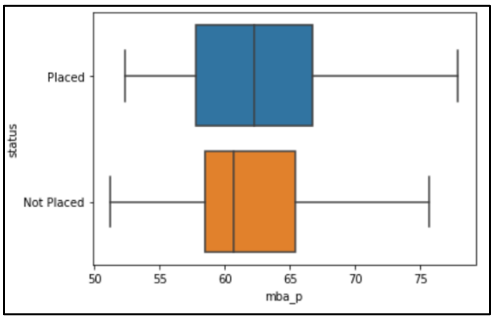


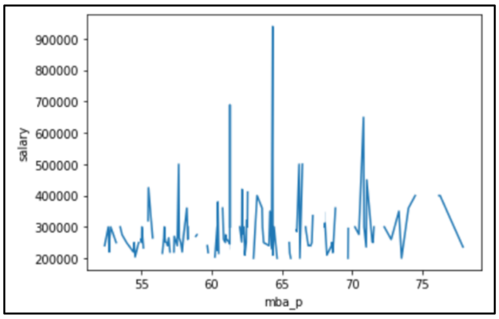
- This affects placements. Very few candidates can be seen to have work experience but unplaced.
- Work placement increases the chances of placement at a considerably higher rate.



- Outliers (higher salary than average) on both end but students with work experience get much higher salary.
- Average salary as well as base salary are both higher for candidates with work experience.

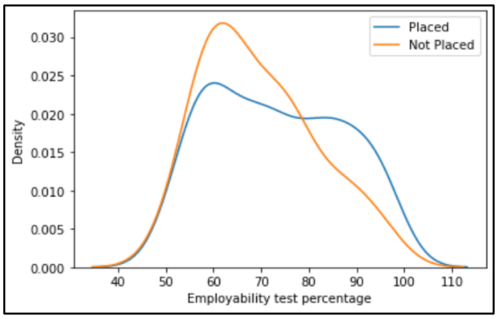
Feature: MBA Percentage



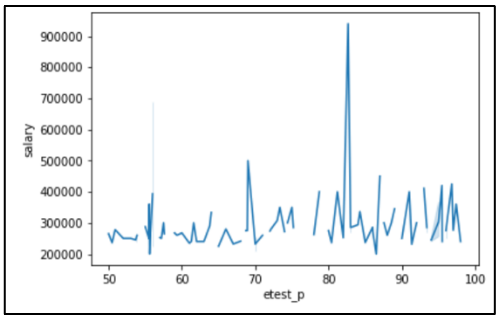


- MBA percentages affects both salary and placements of a particular candidate significantly.
- MBA candidates seem to have a bit higher salary than other candidates.

Feature: Employability Test Percentage



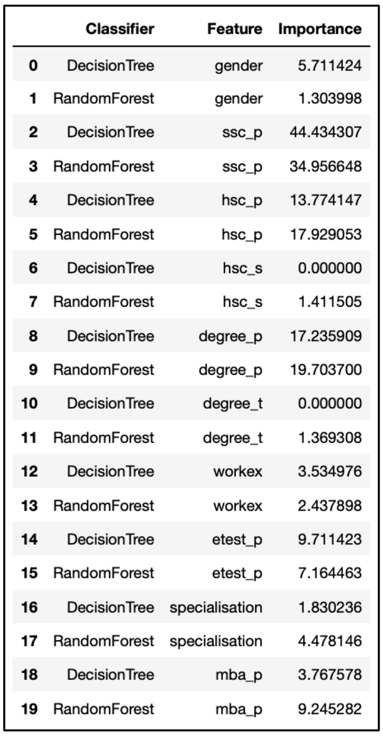
- A high overlap is seen between placed and unplaced students under employability test percentage.
- Employability test percentage does not affect salary much.
- Placed students are seem to have 80% and higher in this test.
- Unplaced students are seem to have percentage from range 50% to 79%.



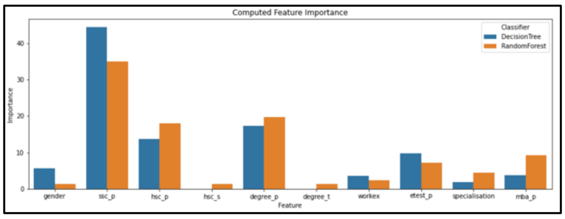
- This test surprisingly does not affects one salary.
- Hence, for salary, employability test percentage must not be taken into account.

Results: *We are going to use only those features which we found out to be considerably affecting one's placement and salary. Hence, ignoring boards of education of senior secondary education and higher secondary education as they don't seem to have much affects, we are left with following features:*

1. *Gender*
2. *Senior Secondary Education Percentage*
3. *Higher Secondary Education Percentage*
4. *Under Graduate Degree Percentage*
5. *Work Experience*
6. *Employability Test Percentage*
7. *MBA Percentage*



	Classifier	Feature	Importance
0	DecisionTree	gender	5.711424
1	RandomForest	gender	1.303998
2	DecisionTree	ssc_p	44.434307
3	RandomForest	ssc_p	34.956648
4	DecisionTree	hsc_p	13.774147
5	RandomForest	hsc_p	17.929053
6	DecisionTree	hsc_s	0.000000
7	RandomForest	hsc_s	1.411505
8	DecisionTree	degree_p	17.235909
9	RandomForest	degree_p	19.703700
10	DecisionTree	degree_t	0.000000
11	RandomForest	degree_t	1.369308
12	DecisionTree	workex	3.534976
13	RandomForest	workex	2.437898
14	DecisionTree	etest_p	9.711423
15	RandomForest	etest_p	7.164463
16	DecisionTree	specialisation	1.830236
17	RandomForest	specialisation	4.478146
18	DecisionTree	mba_p	3.767578
19	RandomForest	mba_p	9.245282



III. CONCLUSION

The proposed work was an attempt to predict his students placement with the help of a data mining approach. In this work, the three best suited classification algorithms - logistic regression, decision tree and random forests were used. We analysed the accuracy of different algorithms. Logistic regression and Random Forest were good with same accuracy of 75.35% whereas decision tree gave an accuracy of 81.53% based on the given data set. The accuracy of machine learning algorithms may differ according to the data set. From the result of our analysis it is clear that logistic regression, random forest, decision tree are good for binary classification problems since they all give accuracy of above 75%.

Therefore the predictive model proposed in this study would help the educational institutions to give quality inputs prior to placements in order to help the students to work on their weak areas for getting placed in their dream companies which would indirectly help the institution in achieving the milestone. From the promising results obtained, it can be concluded that the proposed model could be implemented with the help of other classification algorithms also. The present study is focused on new few major departments with limited features hence, if the data set is further strengthened, the model can be successfully utilised for other educational institution related applications in result analysis which is biggest challenge for any educational institution.

Top 5 features affecting placements we identified are:

1. ssc_p → Senior Secondary Percentage
2. degree_p → Under Graduation Percentage
3. mba_p → MBA Percentage
4. hsc_p → Higher Secondary Percentage
5. etest_p → Employability Test Percentage

Top 5 features affecting placements we identified are:

1. gender → Gender
2. degree_p → Under Graduation Percentage
3. mba_p → MBA Percentage
4. hsc_p → Higher Secondary Percentage
5. etest_p → Employability Test Percentage

IV. REFERENCES

1. Alfiani, Ardita Permata, and Febriana Ayu Wulandari. "Mapping a student's performance based on a data mining approach (a case study)." Agriculture and Agricultural Science Procedia 3 (2015): 173-177.
2. Liu, Yang, et al. "The Application of Data Mining Techniques in College Students Information System." 2018 International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018). Atlantis Press, 2018.
3. Gilbert, Noah. "Predicting Success: an Application of Data Mining Techniques to Student Outcomes." International Journal of Data Mining & Knowledge Management Process (IJDKP) 7.2 2017: 1-20.
4. Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. "Classification and prediction based data mining algorithms to predict slow learners in the education sector." Procedia Computer Science 57 (2015): 500-508.
5. Sheetal, M. B, Savita, Bakare. “Prediction of Campus Placement Using Data Mining Algorithm-Fuzzy logic and K nearest neighbor.” International Journal of Advanced Research in Computer and Communication Engineering 5.6 2016: 309-312.
6. Pothuganti Manvitha and Neelam Swaroopa, Campus placement prediction using supervised machine learning techniques, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 9 (2019) pp. 2188-2191.
7. Ajay Kumar Pal and Saurabh Pal, Classification model of prediction for placements of students, I.J.Modern Education and Computer Science, 2013, 11, 49-56.
8. Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A and Tojo Mathew, Student placement prediction using machine learning, (IRJET), e-ISSN: 2395-0056, pISSN: 2395-0072, Volume: 06 Issue: 04 | Apr 2019.