Recent Developments in Electronics and Communication Systems
KVS Ramachandra Murthy et al. (Eds.)

doi:10.3233/ATDE221271

Student Placement Prediction Using Machine Learning Models (KNN, SVM, RF, Logistic Regression)

Lutukurthi sathish [a,1] and Tirukoti Sudha Rani [b]
[a]*Aditya Engineering College (A), Surampalem, India.*
[b]*Jawaharlal* Nehru *Technological* University, Kakinada, India.

Abstract: It is a dream for the most students to get placed in the process of fulfilling it we have come out with the application which predicts the placement of each student and also gets the data base of them. The system that could fore-cast the chances or the type of company which predicts the probability a semi-final year student to get placed is called a placement predictor. This prediction system helps an institution in the coming years in terms of academic planning. By using the emerging technologies like deep learning and data mining, a lot of predictor models were coming into picture by analyzing the dataset of passed out student. This project gives a method for using algorithm in deep learning to establish a placement prediction model for semi-final year graduates Unlike other methods in this approach we will try multiple methods and finalize the best method to predict the placements.

Keywords: Machine learning, Artificial intelligence, python, Sklearn , Data Analysis, pandas, classification, K- nearest neighbors (KNN),Naive Bayes, SVM, regression, prediction

1. Introduction

The agenda of higher educational institutions is to provide excellent career options to the students. Placement offers are accounted to be key for all the learners in the college. Institutions are being chosen by students based on number of students got placed from a particular institution. Placements are the key deciding factor in organization rankings. Therefore, every organization is going to be benefitted with this approach of fore- casting the placement probabilities of every job aspirant based on some skills and factors.

The institutions, in order to offer the best training to their students, follow a decision-making process. To back up the decision-making process, different techniques and methodologies involved in education data mining were used for identifying the knowledge by understanding the student databases. The attributes like the performance in the assessment examination conducted by the companies or agencies and communication skills along with the academic capabilities are also very essential to get placed. Now, the decisions are taken towards predicting the parameter which contributes more to become successful in placements along with the chances of getting selected.

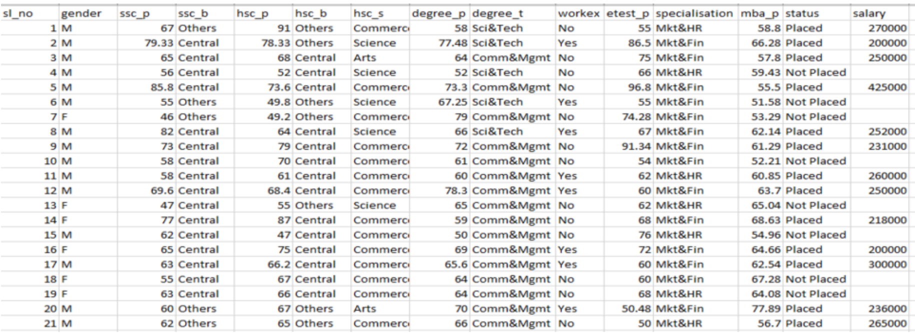
In this project, we gathered the information on 5 different factors of final year student like percentage of marks in 10th standard, Intermediate, graduate level, cube score and an attribute called “Selected” which informs us about the status of the student. Since the data is the most important thing for implementing this model, we gathered the data from various resources from google. And that data set we have lot of factors/features, but after lot of statistical implementations and tests we removed some unwanted features and finally left with most important features. These are the features, which plays major role in learning of algorithms or models to predict the future instances. Some of the main features are gender, department of the student and type of the degree that particular candidate is pursuing and academic track record of the candidates and finally status of placements. The original and fore-casted placement status are compared. The effectiveness and precision of each model is tested and envisaged depending on the factors like execution and final results are discussed.

A. Prediction system:

The different techniques in the machine learning like are Logical Regression, Random Forest, KNN, SVM are used on the dataset to fore-cast the placement status of students. The elements in the dataset that are concentrated mainly for the fore-casting are SSC-P, Intermediate-P, Degree-P, MBA-P, Salary.

B. Sample Data’s:

Table :1 Dataset used for Fore-casting



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
1	M	67	Others	91	Others	Commerc	58	Sci&Tech	No	55	Mkt&HR	58.8	Placed	270000
2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000
3	M	65	Central	68	Central	Arts	64	Comm&Mgmt	No	75	Mkt&Fin	57.8	Placed	250000
4	M	56	Central	52	Central	Science	52	Sci&Tech	No	66	Mkt&HR	59.43	Not Placed	
5	M	85.8	Central	73.6	Central	Commerc	73.3	Comm&Mgmt	No	96.8	Mkt&Fin	55.5	Placed	425000
6	M	55	Others	49.8	Others	Science	67.25	Sci&Tech	Yes	55	Mkt&Fin	51.58	Not Placed	
7	F	46	Others	49.2	Others	Commerc	79	Comm&Mgmt	No	74.28	Mkt&Fin	53.29	Not Placed	
8	M	82	Central	64	Central	Science	66	Sci&Tech	Yes	67	Mkt&Fin	62.14	Placed	252000
9	M	73	Central	79	Central	Commerc	72	Comm&Mgmt	No	91.34	Mkt&Fin	61.29	Placed	231000
10	M	58	Central	70	Central	Commerc	61	Comm&Mgmt	No	54	Mkt&Fin	52.21	Not Placed	
11	M	58	Central	61	Central	Commerc	60	Comm&Mgmt	Yes	62	Mkt&HR	60.85	Placed	260000
12	M	69.6	Central	68.4	Central	Commerc	78.3	Comm&Mgmt	Yes	60	Mkt&Fin	63.7	Placed	250000
13	F	47	Central	55	Others	Science	65	Comm&Mgmt	No	62	Mkt&HR	65.04	Not Placed	
14	F	77	Central	87	Central	Commerc	59	Comm&Mgmt	No	68	Mkt&Fin	68.63	Placed	218000
15	M	62	Central	47	Central	Commerc	50	Comm&Mgmt	No	76	Mkt&HR	54.96	Not Placed	
16	F	65	Central	75	Central	Commerc	69	Comm&Mgmt	Yes	72	Mkt&Fin	64.66	Placed	200000
17	M	63	Central	66.2	Central	Commerc	65.6	Comm&Mgmt	Yes	60	Mkt&Fin	62.54	Placed	300000
18	F	55	Central	67	Central	Commerc	64	Comm&Mgmt	No	60	Mkt&Fin	67.28	Not Placed	
19	F	63	Central	66	Central	Commerc	64	Comm&Mgmt	No	68	Mkt&HR	64.08	Not Placed	
20	M	60	Others	67	Others	Arts	70	Comm&Mgmt	Yes	50.48	Mkt&Fin	77.89	Placed	236000
21	M	62	Others	65	Others	Commerc	66	Comm&Mgmt	No	50	Mkt&HR	56.7	Placed	265000

C. Architecture Diagram

The pandas library in python is used to create data frame for the algorithms. fillna(method='ffill') handles the empty data fields. An efficient predictive analysis tool named sklearn is used. To train and create test sets from the datasets we imported train_test_split. The standard scaler does the standardization. The puzzling matrix is used to view the briefness and output can be forecasted depending on the corresponding algorithm used.

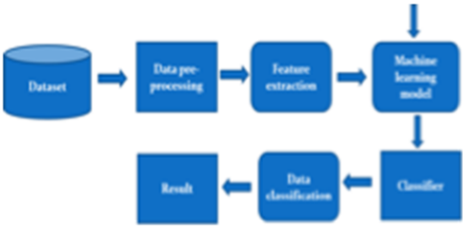


Fig :1 Architecting the data infrastructure, checking accuracy, and training models for prediction

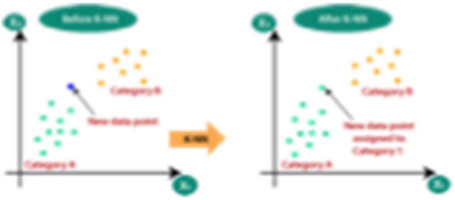


Fig 2:KNN

2. Methodology:

Using the countless obtainable algorithms Machine learning can be gone through. Every algorithm has unique advantages and disadvantages, it will change for the mode of dataset or mode for problem being used what we got at hand some of the algorithms are given below. A comprehensive description long with each algorithm

A. KNN :

KNN abbreviation is, k-nearest neighbors. The K- Nearest Neighbors algorithm is a very popular machine learning algorithm that is very easy to understand and explain. It is also very useful for various classification problems. Although it is regarded as a very intuitive algorithm, it is still very useful for many applications.

If you are new to machine learning, it is important to thoroughly study both the K-Nearest Neighbors algorithm and the other popular metric. There are many different options to choose from. For instance, I will use the K-Nearest side by algorithm to solve the distance query

In this case, we will use the K-Nearest Neighbors algorithm to predict if a customer is likely to purchase our product. To ensure that both the model and the data are fit, I will use the default value of 5

The algorithm:

- This type of model represented using KNN.
- How can a model be obtained and make predictions by using KNN?
- The several nomenclatures for KNN incorporating with how different fields refer to it.
- How the data be prepared to get the maximum from the KNN.
- The references to learn more about KNNs.

This is the easiest algorithm for implementation of KNN. Based on value selected for K the fallacy of the algorithm depends. It's recommended to execute the algorithm with the different inputs of K, in order to find the K for the data given. The least fallacy of the K can be found

In the above figure , if the new data point class we need to find out then we will calculate the distance of that new data point to all near by data points and based on the k value and distances we will find out whether the new data point belongs t category A or category B. There are different kind of distance measures but mostly we will use Euclidean distance. Let us assume that (x1,y1) (x1,y1) and (x2,y2) (x2,y2) are two points in a two-dimensional plane. Here is the Euclidean distance formula.

The Euclidean distance formula says:

d = square root of [(x22 – x11)2 + (y22 – y11)2]

where,

- (x11, y11) are the coordinates of one point.
- (x22, y22) are the coordinates of the other point.
- d is the distance between (x11, y11) and (x22, y22).

➢ *Advantages:*

- KNN is a relatively simple algorithm that can be used to build a model without requiring additional knowledge. It can be used in various applications such as classification and regression.

➢ *Dis-advantages:*

- The algorithm speed reduces as the no. of variables and examples increases.

Although KNN is an ideal method for identifying class labels, it is not very practical for use in situations in which predictions are needed to be made much faster. Therefore, if anyone has the necessary command power, KNN can be used in other applications.

B. SVM :

SVM stands for Support Vector Machine. Among the most popular Supervised Learning algorithms, SVM is the one used for Classification and also for Regression problems. But, primarily, it's usage id for machine learning classification problems. A data item in the n-dimensional space refers to the number of features that it has. This number is computed by taking into account the value of each feature. After plotting the data item and performing classification, we find the hyper-plane that is most likely to differentiate it from the other classes. Now the problem lies in finding which hyper-plane t must be chosen such as right 1. A library in python is Scikit-learn that can be used timplement different machine learning algorithms, SVM also could be utilized with the scikit-learn it is a library in python.

➢ *Advantages:*

- It is effective in spaces with high dimensions.
- This method is still applicable when the number of dimensions exceeds the number of samples.
- The support vectors are therefore memory efficient because they use a subset of training points.
- The decision function can be specified with different Kernel functions. Custom kernels can also be specified in addition to the common kernels.
- Higher speed and better performance with a limited number of samples (in the thousands) compared to neural networks

➢ *Disadvantages:*

- It is crucial to avoid over-fitting in choosing Kernel functions and regularization terms if the number of features is more than the number of samples..
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross- validation

C. Logistic Regression:

When the dependent variable is dichotomous (binary), logistic regression is the appropriate regression analysis to perform. Like all regression analyses, logistic regression is a predictive analysis. An example of a logistic regression is a comparison between an independent binary variable and a nominal, ordinal, interval or ratio dependent variable

The Intellects Statistics software helps you performthe analysis and interpret the output in plain English, so that you can easily understand the results.

The statistical model which is used to forecast the result depending on binary dependent variables is called logistic regression. As this model has a non-independent variable with just two outputs, it is primarily used to model the probability of events resulting from win/lose, alive/death. Logistic function is used by this model to estimate the probability between one or more dependent variables and find the similarities in their relationships.

The following is defined as logistic *Regression*

➢ *Advantages*

- Using Logistic Regression, you can predict the label of a variable based on its probability.
- The distribution of features is not based on any underlying assumptions.
- The binomial correlation coefficient can be applied to multiple cases
- A logistic regression on a variable predicts its label on the basis of a probability distribution.

(x) = 1/(1+exp(-x))

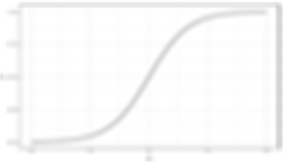
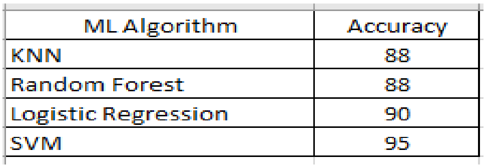


Fig 3: - Logistic regression



ML Algorithm	Accuracy
KNN	88
Random Forest	88
Logistic Regression	90
SVM	95

Table :2 Accuracy of algorithms using Machine Learning

➢ *Disadvantages*

- For high-dimensional datasets, Logistic Regression is over-fitted. In these scenarios, regularization (L1 and L2) techniques can be used to prevent over-fitting.
- When linear decision surfaces can't be drawn, that is,when data can't be linearly separated, the model

C. *Random Forest*

There are a variety of classification algorithms that can be used to classify different types of objects, such as SVM, Logistic regression, and Naive Bayes. In the pecking order of these algorithms, the unmethodically the forest classifier is close to the top. This paper will talk about how decision trees work. The decision tree structure is a flowchart-like framework that shows the various nodes that are involved in a feature. They are then used to represent the class label of the feature. The branches of these represent features that lead to the class of labels. A decision tree is composed of rules that describe the paths that the feature takes from its root node to its leaf node. The goal of random forest is to provide a collection of decision trees that are based on the wisdom of the crowds. In a forest, every tree with its own classified rules and properties will try to find a suitable classes labels for the issue. A voting process is then carried out in the forest to see which tree gets the most votes. The result of the voting determines which class label gets the most votes and is considered the final label for the problem. This method provides a more accurate model for predicting class labels

➢ *Advantages:*

- It can balance errors in data set where classes are imbalanced
- It handles the lengthy data sets with greater dimensionality.
- It handles more than 10 hundreds of input variables and could notice the most prominent variables and as such, it is a good method to reduce the dimensionality.

➢ *Disadvantages:*

It does better of a job for classification problems rather than regression problems as it finds it harder to produce continuous values rather than discrete ones

3. Result:

The end result of the different machine learning algorithms performing is tabulated below. Machine learning algorithms like KNN, Logistic Regression, Random Forest and SVM are used for analyzing the data. We tested and forecasted the position of students regarding placements depending on the same dataset and accomplish the True Positive, True Negative False Positive, False Negative, and precision of each and every algorithm Based on the confusion matrix results we found out the accuracy of each model and tabulated. Based on those things we can observe that every model is good for the data set but Support Vector Machine giving the best result followed by Logistic Regression.

4. Conclusion

Placement prediction system is the one which forecasts the level of final year graduates regarding placements. Different Machine learning algorithms were implemented using python for the data analysis and to forecast the results. We examine the precision of various algorithms and it’s tabulated above. It’s evident that SVM gives 95% exactness. Logistic Regression was also a better technique which gives 90% correctness depending on the dataset. The precision of Machine learning algorithms might vary based upon the dataset. By observing the results, it is evident that, KNN, Random Forest, Logistic Regression, SVM in the mentioned order are efficient for the problems involving binary classification because theyall have a precision more than 87%. Few recruiters may take GATE scores and history of backlogs into the consideration which were not present in our dataset. The results might deviate from the actual in such cases.

References:

[1] Molina, M. M., Luna, J. M., Romero, C., & Ventura, S., 2012, “Meta-learning approach for automatic parameter tuning: a case of study with educational datasets”, in Proceedings of the 5th international conference on educational data mining, pp.180- 183.

[2] T. Jeevalatha, and N. Ananthi,D. Saravana Kumar, “Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms”, International Journal of Computer Applications (0975 – 8887),Volume 108 – No 15, December 2014. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350..

[3] NeelamNaik and SeemaPurohit, “Prediction of Final Result and Placement of Students using Classification Algorithm”, International Journal of Computer Applications (0975 – 8887), Volume 56– No.12, October 2012.

[4] Ajay Kumar Pal and Saurabh Pal, “Classification Model of Prediction for Placement of Students”, I. J. Modern Education and Computer Science, 2013, 11, 49-56.

[5] Ajay Shiv Sharma, Swaraj Prince, ShubhamKapoor and Keshav Kumar, “PPS - Placement Prediction System using Logistic Regression”, IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), 2014.

[6] BahaSen ,EmineUcar and DursunDelen , “Predicting and analyzing secondary education placement-test scores: A data mining approach”, International journal of Expert system with applications, Volume 3 , 2012,Issue 10, pgno: 9468-9476

[7] Mangasuli Sheetal B1, Prof. 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 Vol. 5, Issue 6, June.

[8] S.Hari Ganesh A.Joy Christy “Applications of Educational Data Mining: A Survey “, IEEE Sponsored 2nd International Conference ICIIECS-15

[9] John Jacob, Kavya Jha,Paarth Kotak, Shubha Puthran “Educational Data Mining Techniques And Their Applications”, IEEE International Conference On Green Computing and Internet Of Things (ICGCIoT),2015. https://doi.org/10.1109/ICGCIoT.2015.7380675

[10] Ramanathan et al. “Mining Educational Data for Students' Placement Prediction using Sum of Difference Method”, International Journal of Computer Applications (0975– 8887) Volume 99– No.18, August 2014.