

Evaluation of Placement Predictions with KNN and Logistic Regression Algorithms

¹Kodukula Sirisha, ²Erukulla Sai Tharun and ³N. Suresh Kumar

¹⁻³Department of Computer Science and Engineering, GITAM Institute of Technology, Visakhapatnam
sirishaa1133@gmail.com, erukullasaitharun15@gmail.com, nskgitam2009@gmail.com

Abstract—Placements in campus is an important factor for students to choose that educational institution as their choice. To understand the standards of the institution, placements play a vital role. Using machine learning algorithms such as classification techniques can predict the status of the students regarding placements. In this paper a models designed with help of classification algorithms which predicts weather a student is placed or not. This helps students to understand the institute's placement capability. This model will also suggest few categories which are helpful in order to get placed such as which degree should be opted, choice of specialization, percentage requirement etc: This model is a comparison between two main classification algorithms: KNN and Logistic regression which will further help us in estimating the accuracies of both models.

Index Terms— Classification, KNN, Logistic regression, Accuracy.

I. INTRODUCTION

Placements are the main factor in every institution. Majority of students tend to attract to a particular institute if it has high placement rate. So, predicting and analysing a model which tells about the additional skills required in order to get placed.

The main objective of this model is to predict the student is placed or not by the percentage or his/her degree branch and high school grades and also specialization chosen. Algorithms used in this model are KNN and Logistic Regression. These both are supervised learning; it is a technique in which train the model by giving input and output so that it can learn from it. Section 2 gives us the idea about the related work done and the methodology so far for building this model. Section 3 gives us idea about the literature survey which means the study of related work done so far, and contains the algorithm part and final section contains the work done and result [1][2].

II. RELATED WORK AND METHODOLOGY

A. Data Gathering and cleaning

The dataset is adopted from online sources for evaluation [3]. The data collected will never be ready to start the process. Correct the data as it contains lot of inconsistencies such as missing values, redundant variables, duplicate values, etc. Removing such inconsistencies is very essential as it may end up in showing errors [4] [5].

B. Exploratory Data Analysis

EDA or Exploratory Data Analysis is one of the stages of Machine Learning. It includes knowing about the data. At this stage, drawn and correlate the data between the variables. The dimensionality of the data is 215X15. The categorical values such as gender, Degree_t, workex, Specialization, and status are categorical values need to be converted into int datatype (0,1). The rest of the columns are used for visualization with graphs and charts [6]. Some of the points observed from the data as follows:

- i) Male students are double than female and they are getting placed for a bit higher salary than females.
- ii) Central board students are more in number but compared to them other board students are also getting placed well.
- iii) Most popular branch of degree is commerce and management.
- iv) Most popular specialization is mkt&fin and student who did this specialization are getting placed.
- v) employability test percentage, high school branch does not affect more on campus placements.
- vi) The factors that affect the student placement are gender, school percentage, high percentage, degree percentage, degree branch, work experience, specialization, mba percentage.

C. Building a Machine Learning Model

The attribute values are derived during data analysis. The hypothesis selected at data processing is applied to machine learning model.

All the patterns derived during Data Exploration are used to build the Machine Learning Model. This stage we began by splitting the data set into two parts, training data, and testing data. Training model is used to analyse the model. Testing data is used to test the model. The instance space is divided into two parts, one is training part and the second one is testing part. The data analysis is done with two classification models KNN and Logistic regression models. The two models are evaluated for their performance comparison [7].

III. LITERATURE SURVEY

Student academic performance prediction is one of the important attribute among all attributes. Most of the existing techniques considered classification techniques for machine learning models. The academic performance can be derived from the student result. The student results is most important factor need to be considered for their evaluation. The previous year grades and internal marks of the current semester plays very important role in evaluating the student performance [7][8].

A. Knn algorithm

Knn is the supervised learning algorithm and it is used to assume similarity between the new case and available data and then places the new data point into the existing classes.

Knn is also known as lazy learning algorithm and also is a nonparametric algorithm [2][4]

Steps in knn algorithm:

Step 1: Choose K value to select nearest neighboring clusters from the data set

Step 2: Calculate the Euclidean distance from data point of a new student to available data points in the data set.

Step 3: Chose the K number of clusters grouping nearest data points with respect to new data point.

Step 4: Among these k neighbors' points, count the number of data points in each category and add new data point to that cluster, where the count is maximum.

Step 5: Repeat the above steps for different k values and find the best k value in terms of accuracy.

In the present model after importing knn classifier the fit function is implemented where it tries to fit out train data of independent and dependent variable. After completing fitting of the data, get the best K value in our model In the present model maximum accuracy 0.93 is found when K=4. The best accuracy is evaluated from confusion matrix. The confusion matrix consists of parameters like True positive and True negative with False positive and False negative. In the present work the parameters are derives as follows,

True positive: These are the values that are true for example students who are placed is predicted as placed.

True negative: These are values that are false and predicted as false example students who are not placed predicted as not placed.

False positive: These are type I error means model predicted as positive but the actual value is negative. Example students who is not placed but predicted as placed.

False negative: These are type II error means model predicted as negative but the actual values is positive. Example student who is placed predicted as not placed

B. Logistic regression

The supervised binary class Machine learning algorithm Logistic regression predicts the probability of an event in the range of 0 to 1. It maps the predicted value to class value. Here sigmoid function is used to map the values in between 0 and 1. Here sigmoid function is used in order to map the values in between 0 and 1. It shapes like symbol 'S'. Sigmoid function is also called logistic function [5].

Dataset Description:

This dataset contains of placement data of students. It includes high school and secondary school percentages and specialization. It also includes type of graduation and obtained grades; Work experience and salary offers the places students [1]. The head and tail of the data set is shown in figure 1.

Data source

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
210	211	M	80.60	Others	82.00	Others	Commerce	77.60	Comm&Mgmt	No	91.0	Mkt&Fin	74.49	Placed	400000.0
211	212	M	58.00	Others	60.00	Others	Science	72.00	Sci&Tech	No	74.0	Mkt&Fin	53.62	Placed	275000.0
212	213	M	67.00	Others	67.00	Others	Commerce	73.00	Comm&Mgmt	Yes	59.0	Mkt&Fin	69.72	Placed	295000.0
213	214	F	74.00	Others	66.00	Others	Commerce	58.00	Comm&Mgmt	No	70.0	Mkt&HR	60.23	Placed	204000.0
214	215	M	62.00	Central	58.00	Others	Science	53.00	Comm&Mgmt	No	89.0	Mkt&HR	60.22	Not Placed	NaN

215 rows × 15 columns

Figure 1: Features of Data set

IV. RESULT

After implementation of KNN algorithm on student data set it is found with 91% of accuracy predicted the accuracy. In the present work the logistic regression is also tested and analysed for comparison with KNN algorithm. The maximum accuracy is obtained 79.62% with logistic regression. The confusion matrix derived for KNN and Logistic regression is shown in figure 2a and Figure 2b.

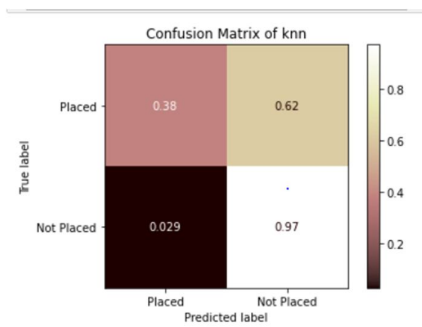


Figure 2a: confusion matrix for KNN

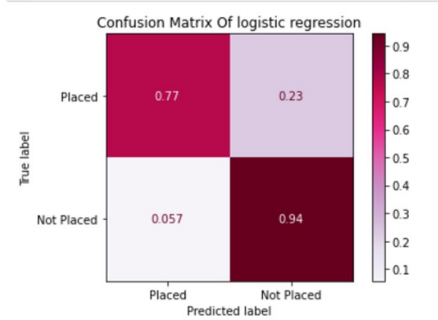


Figure 2b: confusion matrix for Logistic regression

The accuracies observed with different k-values are shown in table I.

TABLE I: ACCURACY OBSERVED IN KNN

k-value	Accuracy
3	0.9161676646706587
4	0.9341317365269461
5	0.8862275449101796
6	0.9101796407185628
7	0.8862275449101796

V. CONCLUSION

Comparing the accuracies of KNN with Logistic Regression, it is concluded that K- Nearest Neighbour is best choice for this dataset as it has accuracy of 91%. From the data set it is observed that board studies do not have effect on placement percentage. On the other hand, the specialization of the graduation plays very important role and clear indicator for placements. The percentage of scholl does not have impact on placement, but percentage in graduation have impact on placements. It is also observed that students with no experience are also placed. It indicates experience is not a big criterion to get the placements for freshers.

REFERENCES

- [1] International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 9 (2019) pp. 2188-2191 © Research India Publications. Campus Placement Prediction Using Supervised Machine Learning Techniques
- [2] International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019 Student Placement Prediction Model: A Data Mining Perspective for Outcome-Based Education System
- [3] <https://www.kaggle.com/benroshan/factors-affecting-campus-placement>
- [4] N. Suresh kumar, P. Praveena, "Evolution of hybrid distance based kNN classification ", IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 10, No. 2, June 2021, pp. 510~518
- [5] Divya Sai Patnana, et al., "Logistic Regression Analysis on Social Networking Advertisement", Journal of Critical Reviews, VOL 7, ISSUE 04, 2020
- [6] Samriddhi Volume 12, Special Issue 2, 2020 Print ISSN: 2229-7111 Online ISSN: 2454-5767 Student Placement Prediction System using Machine Learning
- [7] Educational data mining for student placement prediction using machine learning international Journal of Engineering & Technology 7 (1.2) (2018) 43-46
- [8] 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.