Review on Predictive Analysis of Placement of Students Using Machine Learning Algorithms

Vinod S. Agrawal[1(✉)] and Sarika Shelke Kadam[2]

[1] MGM University, Aurangabad, MH, India
vagrawal@mgmu.ac.in
[2] DR. G.Y. Pathrikar College of Computer Science and IT, MGM University, Aurangabad, MH, India

Abstract. Machine learning is an emerging trend which has proven to learn automatically from past data. Machine learning can be performed using various ways like supervised, unsupervised, reinforcement learning. Machine learning where system computer system learning from the data set available. Machine learning using supervised learning can be performed using either regression or classifier algorithms. In this research paper we have worked on machine learning algorithms which are SVM, Decision Tree, Logistic regression and Random Forest. Using Machine learning algorithms we can build mathematical model and make the predictions using dataset. To make prediction we can go for supervised learning where output will be in the form either yes or No, 0 or 1, in this case either student placed or not placed.

Keywords: Machine learning · Decision tree · Logistic Regression · SVM · Random Forest

1 Introduction

Whenever any student visit for college admission basically looks for placement percentage that college have. All the educational institutes are constantly work hard for placement activities. A detailed literature review has been done to understand the various machine learning algorithms that can be applied. In education sector as well machine learning is used for prediction purpose. Earlier prediction will improve student's learning for placement drives. Using Machine learning algorithms system learns from available dataset and make the prediction once new data received by the system in the form of classification whether student will get placed or not.

Classification of machine learning: 1. Supervised learning in this we consider the labeled data to train to the system. Once system is trained, it generates/predicts the output. It is same like student learns from the teacher. It can be further divided into classification and regression models. In this paper we have worked on classification type as student placement can be classified as either Placed or Not placed. 2. Unsupervised learning As compared to supervised learning set of data provided to machine learning has not been labeled, classified or categorized. In this case we don't have predetermined results. It

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 23–30, 2023.
https://doi.org/10.2991/978-94-6463-136-4_4

is used to find out hidden pattern in the data. It is further divided into Clustering and Association. 3. Reinforcement learning. It is feedback based learning method. In this agent interact with environment and explore it. Decision tree model: They generates rule for the classification of data. It can be used for classification as well as regression, but mostly used for solving classification problems. Dataset is the collection of data in the form of rows and columns as per the problem type specification. Feature is the attribute of the entity. There could be multiple features included into dataset to have great impact on the performance. The feature which is not relevant can be discarded from the dataset. Once decision tree is formed it becomes easy to label for particular case. Random Forest algorithm model: It contains various decision trees for each subsets of the available dataset and then it calculates the average to improve the accuracy of prediction. Support Vector Machine model: Used to create the best line that can segregate n-dimensional space into classes so we can fit new data point in the right category. Logistic Regression model: Used to predict the dependent variable using a set of independent variables. It uses s shaped logistic function which predicts two maximum values i.e. 0 or 1.s shaped function is also called as sigmoid function.

2 Literature Review

According to T. Jeevalatha et al., predicting placement of a student requires lot of parameters to be considered. Personal, Social, Psychological and other variables required for effective prediction of placement of a student. They have worked on decision tree algorithms like ID3, C4.5, and CART. Data preprocessing is required as poor quality of raw data affects the data mining efficiency. Total samples (student details) collected were 1342 and unnecessary columns like residential address removed. Results calculated in the form of recall, precision, accuracy. From the result it is found that ID3 algorithm is appropriate for prediction of placement of student. Efficiency of various decision tree algorithms analyzed and ID3 algorithm provided 95.33% accuracy which was higher than others [1].

According to Dr. B. Muthusenthil et al., they had dataset of 185 students (2018 and 2019 pass out) to improve the accuracy score. They studied algorithms like Linear regression, Decision tree, KNN, Logistic regression and Lasso regression [2]. According to Cong Yu Cai, Huijuan Lu et al., They have worked on linear regression model, K-neighbor regression model, decision tree regression model, XGBoost regression model, gradient boost regression model, light GBM regression model and random tree classifier model [3].

According to Tadi Aravind, placement analysis performed using 2 different datasets. One with simple data and second one is with additional features of students. Root mean square error considered for the study [4]. According to Chandrasekhar Kumbhar, Dr. S. S. Sridhar, they have worked on decision tree, Neural network and Support vector machine algorithms. A dataset has 50 entries in which we have considered 37 entries for training purpose and 13 for testing [5].

According to Athreya Shetty B et al., After many experimentation and research we have found out the most efficient machine learning model suited for the task with comparison to many other machine learning prediction algorithms (Decision Tree) [6].

According to Shashi Sharma et al., In this paper, Predictions are made using various machine learning algorithms. Predicted values are compared with original values. Lesser this distance, more accurate will be the predictions [7]. According to Senthil Kumar et al., The accuracy of 71.66% with tested real life data indicates that the system is reliable for carrying out its major objectives, which is to help teachers and placement cell in an institution to find the prospective students and provide them with better coaching so as to excel in placement processes by various companies [8].

According to Hitarthi Bhatt et al., This paper proposes a system which determines the probability of a student getting placed based on different attributes like programming skills, CGPA, communication skills, internships, backlogs, SSC and HSC marks. For this purpose ID3 classification algorithm is used [9]. According to Pratiyush Guleria, Manu Sood, The main finding of this research is the gathering of knowledge from student's academic performance where Placement Results classify attributes and gives a better understanding of how students should perform and acquire other skills apart from their regular studies to get placed [10]. According to Ajay Shiv Sharma et al., The training and testing accuracy of the algorithm was 98.93% and 83.333% respectively [11].

According to J. Hima Bindu, B. Dushyanth, placement prediction system can be beneficial for the institute as well as student. We can even predict the company in which student may get placed. As per the skills required by particular company institute can train the students [12]. According to Liya Claire Joy, Asha Raj, Dataset contains both academic and placement details of students. Placement prediction system can help students as well as institute for proper academic planning [13]. According to Ravina Sangha et al., The dataset on which algorithm will be tested will be that of MESCOE Pune [14]. According to Mansi Gera, Shivani Goel, placement prediction system will help student to prepare for the predicted company skillsets and improve the chance of being placed [15].

According to Chirag Patel, google form was created to collect the students data so that system can be trained on received data and predictions can be made accordingly. This system was specifically for Post Graduate students. Based on this data, predictive model can be developed to map student placement to particular company [16].

According to Sonali Rawat, Different attributes such as academic results, technical skills, training and projects done, are considered to be desirable for prediction purpose. This paper presents an outlook to differentiate data mining techniques for predictive analytics that can be used in the process of predicting campus placement. Various classification and clustering techniques are inspected to evaluate the performance of students in the recruitment procedure. Using the comparative study amongst these techniques ID3 with accuracy 95.33%, KNN with 97.33%, C4.5 with 88.89%, Naïve bayes with 86.15%, Multilayer perception having 87.395% accuracy is suggested to be the best one [17].

According to Tripti Mishra et al., In traditional education, performance prediction is in matured state with contribution from many researchers. However, there is paucity of research in the field of employability prediction. As both performance and employability of students graduating from an institution decide the market value of the institution,

research is required to develop comprehensive models for performance and employability tool and develop a system that will be able to predict both performance and employability [18].

3 Methodology

Many tools can be used for prediction analysis. Here Python programming used as it has huge packages. Dataset is collected from kaggle which has 215 no of records. Each record has 14 attributes like sno, Gender, 10th %, ssc board, 12th%, 12th stream, Degree %, Degree stream, work experience, specialization, Mba %, status and salary. Jupyter notebook based on python used to work with machine learning models (Fig. 1).

Select classifier and Import required packages. Here all attributes except status are independent variables whereas status is dependent variable. Following are the details in case of all classifiers selected for the study.

3.1 Random Forest

70% data from dataset selected as training data and 30% data as testing data. On the execution of dataset on Random Forest classifier we have got 100% accuracy. Accuracy is measured in the form of precision, recall, F1 score and support (Fig. 2).

3.2 Decision Tree

70% data from dataset selected as training data and 30% data as testing data. On the execution of dataset on Decision Tree classifier we have got 100% accuracy (Fig. 3).

```
In [46]: ad_data
```

Out[46]:

	Sno	Gender	10th %	SSC Board	12th %	HSC Board	12th Stream	Degree %	Degree stream	Work exp	specialisation	Mba %	status	salary
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	Mkt&HR	58.80	1	270000.0
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	Mkt&Fin	66.28	1	200000.0
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	Mkt&Fin	57.80	1	250000.0
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	Mkt&HR	59.43	0	NaN
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	Mkt&Fin	55.50	1	425000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
210	211	M	80.60	Others	82.00	Others	Commerce	77.60	Comm&Mgmt	No	Mkt&Fin	74.49	1	400000.0
211	212	M	58.00	Others	60.00	Others	Science	72.00	Sci&Tech	No	Mkt&Fin	53.62	1	275000.0
212	213	M	67.00	Others	67.00	Others	Commerce	73.00	Comm&Mgmt	Yes	Mkt&Fin	69.72	1	295000.0
213	214	F	74.00	Others	66.00	Others	Commerce	58.00	Comm&Mgmt	No	Mkt&HR	60.23	1	204000.0
214	215	M	62.00	Central	58.00	Others	Science	53.00	Comm&Mgmt	No	Mkt&HR	60.22	0	NaN

215 rows × 14 columns

Fig. 1. DataSet containing 14 columns

```
In [80]: from sklearn.ensemble import RandomForestClassifier

         #Create a Gaussian Classifier
         clf=RandomForestClassifier(n_estimators=100)

         clf.fit(x_train,y_train)
         predictions = clf.predict(x_test)
         from sklearn.metrics import classification_report,confusion_matrix
         print(confusion_matrix(y_test,predictions))
         print(classification_report(y_test,predictions))
```

```
[[20  0]
 [ 0 45]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        20
           1       1.00      1.00      1.00        45

    accuracy                           1.00        65
   macro avg       1.00      1.00      1.00        65
weighted avg       1.00      1.00      1.00        65
```

Fig. 2. Performance of Random Forest algorithm

```
In [69]: from sklearn.tree import DecisionTreeClassifier
         clf = DecisionTreeClassifier()
         clf.fit(x_train,y_train)
         y_pred=clf.predict(x_test)
         from sklearn import metrics
         print("Accuracy:",metrics.accuracy_score(y_test,y_pred))
         from sklearn.metrics import classification_report
         print(classification_report(y_test,y_pred))
```

```
Accuracy: 1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        20
           1       1.00      1.00      1.00        45

    accuracy                           1.00        65
   macro avg       1.00      1.00      1.00        65
weighted avg       1.00      1.00      1.00        65
```

Fig. 3. Performance of Decision Tree Classifier

3.3 Support Vector Machine

70% data from dataset selected as training data and 30% data as testing data. On the execution of dataset on Decision Tree classifier we have got 83% accuracy (Fig. 4).

3.4 Logistic Regression:

70% data from dataset selected as training data and 30% data as testing data. On the execution of dataset on Logistic Regression classifier we achieved 100% accuracy (Fig. 5 and Table 1).

```
In [76]: from sklearn.svm import SVC
         svc_model = SVC()
         svc_model.fit(x_train,y_train)
         predictions = svc_model.predict(x_test)
         from sklearn.metrics import classification_report,confusion_matrix
         print(confusion_matrix(y_test,predictions))

         [[10 10]
          [ 1 44]]

In [77]: print(classification_report(y_test,predictions))

                       precision    recall  f1-score   support

                    0       0.91      0.50      0.65        20
                    1       0.81      0.98      0.89        45

             accuracy                           0.83        65
            macro avg       0.86      0.74      0.77        65
         weighted avg       0.84      0.83      0.81        65
```

Fig. 4. Performance of Support Vector Machine algorithm

```
In [81]: from sklearn.linear_model import LogisticRegression
         clf = LogisticRegression()
         clf.fit(x_train,y_train)
         predictions = clf.predict(x_test)
         from sklearn.metrics import classification_report,confusion_matrix
         print(confusion_matrix(y_test,predictions))
         print(classification_report(y_test,predictions))

         [[20  0]
          [ 0 45]]
                       precision    recall  f1-score   support

                    0       1.00      1.00      1.00        20
                    1       1.00      1.00      1.00        45

             accuracy                           1.00        65
            macro avg       1.00      1.00      1.00        65
         weighted avg       1.00      1.00      1.00        65
```

Fig. 5. Performance of Logistic Regression algorithm

Table 1. Performance Analysis

Sr. No.	Classifier	Result
1	**Random Forest**	**100%**
2	**Decision Tree**	**100%**
3	Support Vector Machine	83%
4	**Logistic Regression**	**100%**

4 Conclusion

Supervised machine learning algorithms like SVM, Decision Tree, Logistic Regression and Random forest algorithm are applied on the sample data set. The accuracy obtained in case of Support Vector Machine is 83% whereas in case of Random Forest, Logistic

Regression and Decision Tree accuracy obtained as 100%. Hence it is better to use Random Forest/Decision Tree/Logistic Regression algorithm in such types of datasets.

References

1. T. Jeevalatha, N. Ananthi , D. Saravana Kumar.: Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms. International Journal of Com puter Applications 108(15), pp. 27–31 (2014).
2. Dr. B. Muthusenthil , VenkatMugesh S, Thansh D , Subaash R.: Predictive Analysis Tool for Predicting Student Performance and Placement Performance using ML algorithms. IJARIIE 6(2), pp. 228–235 (2020).
3. Cong Yu Cai, Huijuan Lu, Ke Yan, Minacho Ye.: Intelligent placement model based on decision tree. International Conference on Information Technology in Medicine and Education 2018, pp. 837–841(2018).
4. Tadi Aravind, Bhimavarapu Sasidhar Reddy, Sai Avinash, Jeyakumar G.: A Comparative Study on Machine Learning Algorithms for Predicting the Placement Information of Under Graduate Students. International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) 2019, pp. 542–546(2019).
5. Chandrasekhar Kumbhar , Dr. S. S. Sridhar.: Trend analysis of university placement by using machine learning algorithms. International Journal of Engineering & Technology 7 (2.4), pp.178–181 (2018).
6. Athreya Shetty B, Akram Pasha, Amith Singh S, Shreyas N I, Adithya R Hande.:Comparative Study of Multiple Machine Learning Algorithms for Students' Performance Data for Job Placement in University. International Journal of Engineering and Advanced Technology, 8(5), pp.182–187(2019).
7. Shashi Sharma, Sunil Kumar Pandey, KumkumGarg.: Machine Learning for Predictions in Academics. International Journal of Recent Technology and Engineering (IJRTE) 8(5), pp. 4624–4627 (2020).
8. Senthil Kumar Thangavel, DivyaBharathi P, AbijithSankar.: Student Placement Analyzer: A Recommendation System Using Machine Learning. International Conference on Advanced Computing and Communication Systems (2017)
9. Hitarthi Bhatt, Shraddha Mehta, Lynette R. D'mello.: Use of ID3 Decision Tree Algorithm for Placement Prediction. International Journal of Computer Science and Information Technologies 6(5), pp.4785–4789 (2015)
10. Pratiyush Guleria, Manu Sood.: Predicting Student Placements Using Bayesian Classification. International Conference on Image Information Processing, pp.109–112 (2015)
11. Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor, Keshav Kumar.: PPS - Placement Prediction System using Logistic Regression. IEEE International Conference on MOOC, Innovation and Technology in Education (MITE) 2014, pp. 337–341(2014)
12. J. Hima Bindu, B.Dushyanth, STUDENT PLACEMENT PREDICTION USING MACHINE LEARNING, © 2021 IJCRT | 9(6),(2021)
13. Liya Claire Joy, Asha Raj, A Review on Student Placement Chance Prediction, 5th International Conference on Advanced Computing & Communication Systems (ICACCS)(2019)
14. Ravina Sangha , Akshay Satras , Lisha Swamy , Gopal Deshmukh, "Student's Placement Eligibility Prediction using Fuzzy Approach", International Journal of Engineering and Techniques 2(6) ,(2016)
15. Mansi Gera, Shivani Goel, A Model for Predicting the Eligibility for Placement of Students Using Data Mining Technique, International Conference on Computing, Communication and Automation (2015)

16. Chirag Patel, Identification of essential parameters for post graduate students' job placement in computer applications in India, Original Research, https://doi.org/10.1007/s41870-018-0136-x ,(2018)
17. Sonali Rawat, Predictive Analytics for Placement of Student- A Comparative Study, International Research Journal of Engineering and Technology (IRJET), 06(06),(2019)
18. Tripti Mishra, Dharminder Kumar and Sangeeta Gupta, Students' Performance and Employability Prediction through Data Mining: A Survey, Indian Journal of Science and Technology, 10(24), (2017)