Research Article

Applying Machine Learning Algorithms for Student Employability Prediction Using R

G.Vadivu*, K.Sornalakshmi
Information Technology, SRM University, Kattankulathur, Chennai, Tamil Nadu, India.
***Corresponding author's E-mail:** vadivu.g@ktr.srmuniv.ac.in

Received: 28-12-2016; **Revised:** 18-02-2017; **Accepted:** 05-03-2017.

ABSTRACT

In the educational institutions, large amount of data are collected and stored in the databases. These databases are used to monitor their regular academic performance, co-curricular and extracurricular activities. These databases are indirectly useful for predicting the insights like students' performance in the forthcoming semester and the possibility of getting employment through campus. In this paper, the machine leaning algorithms K-Nearest neighbor methods (KNN and Naïve Bayes are used to predict the employability skill based on their regular performance. Algorithms like KNN and Naïve Bayes, are useful to classify the objects into one of several groups based on the values of several variables. The purpose of using KNN, a lazy method, and can be easily implemented also easy to understand. It works on Bayes theorem of probability to predict the class of unknown data set with strong independence assumptions between the variables. Hundreds of thousands of data values and quite a lot of variables are available with the educational databases. In such situation, KNN-easy to implement, Naive Bayes-which is extremely fast relative to other classification algorithm, both are used in this work to predict the performance of the employability opportunity of the students. kNN and naïve Bayes, both of the algorithms are given for the prediction of employability opportunity as 'Yes/No'. This prediction is calculated based on the regular performance in the course level. The courses are categorized as basic concept level, programming, personality development, mathematics, and advanced technical courses. Input scan be included from their school level to predict more accurate prediction of the students' performance, and in this paper the extracurricular activities are also not considered for employability prediction. These are also the important criteria to predict their performance. In future these can also be included to predict their performance.

Keywords: Machine Learning Algorithms, Naïve Bayes, K-Nearest Neighbor (KNN).

INTRODUCTION

In the educational institutions lots of data are collected to monitor their regular performance. In which their personal details like father name, occupation, yearly income, communication address, phone number, mother name, occupation, phone number, and passport size photo are stored, Academic details like their semester, course code, course name, grade, year of passing, co-curricular and extracurricular activity details are stored for understanding the performance of the student. Based on these details Faculty Counselor can able to counsel them for their improvement. These counseling methodologies are useful only after getting their actual data. But there is no method to predict their future performance based on their present performance. Thus in this paper we tried to apply the machine learning algorithms kNN and naïve Bayes to predict their future performance and this will be useful for the students to improve themselves to get placement through campus.

Research Purpose

Most of the current work is to predict the students' grade in the future semester. [1]Abeer Badr El Din Ahmed et.al. Constructed decision tree which is applied on student's database to predict the student's performance on the basis of student's database.

Princy[2] Christy et.al., extensively covers recent studies in predicting student performance by implementing data mining tasks and employed the supervised learning tasks classification, regression and recommender systems. Each of the techniques in their own ways influenced the outcomes of the prediction task.

Anuradha[3] and Velmurugan applied the classification techniques for the prediction of the performance of students in end semester university examinations. Particularly, the decision tree algorithm C4.5 (J48), Bayesian classifiers, k Nearest Neighbor algorithm and two rule learner's algorithms namely OneR and JRip used for classifying the performance of students.

Vivekananthamoorthy[4], et. al., they conducted to understand how young university students use the Social Networking Site LinkedIn and the responses were used to frame a questionnaire. Exploratory Factor Analysis was conducted based on their survey responses to identify the hidden factors associated with the indicator items in the data set. Subsequently, a theoretical model was constructed, the model helped to predict enhancement of student empowerment by using Social Media.

DISCUSSION

The large amount of data collected by the educational institutions to monitor their performance can be used to analyze their future performance. This can be achieved by the popular machine learning algorithms like kNN and naïve Bayes, shown in Fig.1.

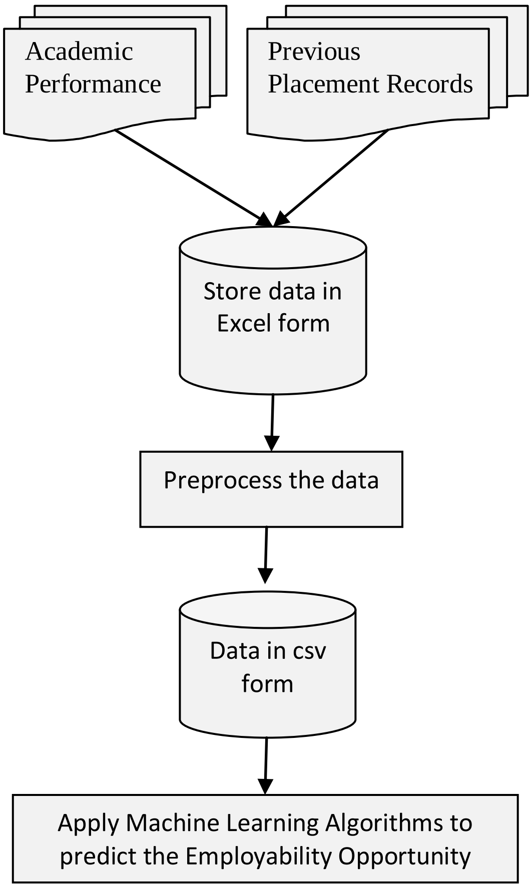
Academic Performance
Previous Placement Records
Store data in Excel form
Preprocess the data
Data in csv form
Apply Machine Learning Algorithms to predict the Employability Opportunity

Figure 1: Flow diagram of the machine learning process

Machine Learning Algorithms

In machine learning, build predictors that allow classifying things into categories based on some set of associated values. Many algorithms are available for automated classification, includes random forests, support vector machines, Naïve Bayes classifiers, and some types of neural networks. Simple and easy to implement is k-Nearest Neighbours (kNN) algorithm. kNN classifies new instances by grouping with the existing instances with the most similar group instances. The kNN task is divided into three major functions:

1. Distance between any two points has to be calculated.
2. Finding the nearest neighbours based on the distance value.
3. Assign the class labels based on the nearest neighbour value.
4. This kind of excel data is taken as input, verified for blank values, non-numeric values, invalid values and converted into csv format. R[6]programming language is used to read the csv data file as input for kNN.
5. >GVind<-sample(2,nrow(GV sample), replace =TRUE, prob =c (0.67, 0.33))
6. The index value was calculated with the probability of 0.67, 0.33.

IT104	PD10	MA10	IT10	IT10	LE12	IT10	Placed
90	80	90	90	80	70	90	1
90	80	90	90	80	70	90	1
40	50	40	40	50	40	40	0
90	50	60	70	70	70	90	1
90	70	70	80	70	80	90	1
90	80	80	80	80	80	90	1

> GVsample.training <- GV sample[GVind==1, 1:250]

In the above command first 250 rows are taken for training set.

GVsample.test <- GVsample[GVind==2, 1:250]

In the above command first 250 rows are taken for test dataset.

GVsample.trainLabels<-GVsample[GVind==1, 59]

Column 59 is having the detail of 'Placed or not' in terms of '1/0'. Those values are given as trainLabels.

GVsample.testLabels <- GVsample[GVind==2, 59]

Test labels will be calculated for the 'testdataset' and the result can be stored in column 59.

GV_pred <- knn(train = GVsample.training, test = GVsample.test, cl = GVsample.trainLabels, k=10)

> GV_pred

[1] 1 1 1 0 1

Levels: 0 1

The sample output shows whether the student will get placed or not 1is for Yes/0 is No, using kNN.

Another algorithm Naïve Bayes is used for the same kind of work, which is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows:

$$P(H|X) = P(X|H)\ P(H)/P(X)$$

Where,

P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X.

P (H) is the prior probability, or apriori probability, of H.

P (X|H) is the posterior probability of X conditioned on H

Naive Bayes algorithm in question classification proceeds by finding out the feature associated with each category. Features are the Headword occurring in questions pertaining to one particular category expressed in terms of their relevance in particular question.

```
>train data <-as.data.frame (sample [1:59,])
```

```
S,B,B,B,B,C,S,S,S,A,A,A,C,C,D,S,A,S,S,B,A,A,A,S,B,B,B,B,S,B,
A,S,B,A,S,A,A,A,B,A,S,A,A,B,A,A,A,S,A,S,S,A,A,S,B,B,S,Yes

S,B,B,B,B,C,S,S,S,A,A,A,C,C,D,S,A,S,S,B,A,A,A,S,B,B,B,B,S,B,
A,S,B,A,S,A,A,A,B,A,S,A,A,B,A,A,A,S,A,S,S,A,A,S,B,B,S,No

D,D,C,D,C,C,A,S,S,A,A,D,D,D,D,C,D,C,B,B,B,D,S,C,D,D,D,D,B
,B,C,C,D,D,C,C,C,C,D,D,D,C,C,D,C,C,C,S,D,B,C,B,A,C,B,B,S,Y
es

C,D,C,D,C,D,A,S,A,S,A,C,D,C,C,B,D,C,B,B,B,D,S,C,D,D,C,D,B,
B,C,C,D,D,C,B,C,B,D,C,C,B,B,C,C,C,C,S,D,C,B,B,A,B,C,B,S,Yes
```

In the above dataset, the input values are shown as grades in letters, and the last value is whether student is placed in terms of Yes/No.

```
> test data <- as. data. Frame (sample [16,])
```

Test data is created for the record number 16 along with all 59 attributes.

```
> Model <- naïve Bayes (Placed ~ IT1001 + ME1005 +
LE1001+MA1002+ME1001+EE1001+IT1002+…..+IT1049,
train data)
```

Naïve Bayes model is created based on the 'Placed' attribute associated with 59 attributes.

```
> Results <- predict(model, test data)
```

Test data has been stored with the dataset to be verified, that has been given as the input to naïve bayes model to predict the result.

```
> Results
[1] Yes
```

Based on the model the placement opportunity for the given test data is 'Yes', means the student will get the employability opportunity. If the result shows 'No' then that student has to be monitored and given counseling to improve his performance.

CONCLUSION

The students' employability is very much important as institution point of view as well as student point of view. In this regard to improve the students' performance, the academic performance has been analyzed and predicted using the algorithms KNN and naïve Bayes. The algorithms are applied on the data set of 250 students with 59 attributes. The accuracy obtained after analysis for KNN is 95.33% and for the naïve Bayes is 97.67%. Hence, from the above said analysis and prediction it would be better if the naïve Bayes is used to predict the student's employability results.

Acknowledgement: We acknowledge with thanks Management and faculty of SRM University for the support and guidance and for providing opportunity. We also acknowledge the student community for their enthusiasm and overwhelming response to this project.

REFERENCES

1. Abeer Badr El Din Ahmed1, Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology, 2(2), 2014, 43-47.
2. Princy ChristyA. and RamaN., "Relevance and Resonance of Data Science in Performance Prediction and Visualization", Indian Journal of Science and Technology, 2016, Vol 9(19).
3. Anuradha C. and VelmuruganT., "A Comparative Analysis on the Evaluation of Classification Algorithms inthe Prediction of Students Performance", Indian Journal of Science and Technology, 2015,Vol 8(15).
4. SenB., UcarE., and DelenD., "Predicting and analyzing secondary education placement-test scores: A data mining approach" Expert Systems with Applications, Vol. 39, 2012, pp. 9468-9476.
5. Vivek anantha moorthyN. NaganathanE. R. and Raj kumar. "Determinant Factors on Student Empowerment and Role of Social Media and eWOM Communication: Multivariate Analysis on LinkedIn usage", Indian Journal of Science and Technology, Vol 9(25), 2016.
6. VenablesW. N., Smith D. M. , R Core Team, "An Introduction to R Notes on R: A Programming Environment for Data Analysis and Graphics", 2015, Version 3.2.1.
7. Elbadrawy A, Studham RS, Karypis G.,"Collaborative multi-Regression models for predicting students, performance in course activities", Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, 2015, p. 103–7.
8. Pal AK, Pal S., "Analysis and Mining of Educational Data for Predicting the performance of Students", International Journal of Electronics Communication and Computer Engineering, 4(5), 2013, 1560–5.
9. Ajith P, Tejaswi B, Sai MSS, "Rule Mining Framework for Students Performance Evaluation", International Journal of Soft Computing and Engineering, 2(6), 2013, 201–6.
10. Agrawal BD, GuravBharti B., "Review on Data Mining Techniques used For Educational System", International Journal of Emerging Technology and Advanced Engineering, 4(11), 2014, 325–9.
11. AffendeyL., ParisI., MustaphaN., SulaimanM., MudaZ., "Ranking of influencing factors in predicting students' academic performance", Information Technology Journal, 2010, Vol. 9 No. 4, pp. 832-837.
12. Bakar A., Noor A., Mustapha A., NasirK. M., "Clustering analysis for empowering skills in graduate employability model", Australian Journal of Basic and Applied Sciences, 2013, Vol. 7, No. 14, pp. 21-24.

13. KabakchievaD., "Predicting student performance by using data mining methods for classification", Cybernatics and Information Technologies, 2013, Vol. 13, No. 1, pp. 61-71.

14. RomeroC., VenturaS., "Educational data mining: A survey from 1995 to 2005", Expert systems with applications, Vol. 33, 2007, No. 1, pp. 135-146.

15. SembiringS., ZarlisM., HartamaD., RamlianaS., WaniE., "Prediction of student academic performance by an application of data mining techniques", Proc. International Conference in Management and Artificial Intelligence, 2011, Vol. 6, pp. 110-114.

16. Nina Zumel, John Mount, "Practical Data Science with R", Manning Publications, 2014.

17. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, "Mining of Massive Datasets", Cambridge University Press, 2014.

18. Mark Gardener, "Beginning R - The Statistical Programming Language", John Wiley & Sons, Inc., 2012.

19. Jure Leskovec, Stanford Univ.Anand Rajaraman,Milliway Labs, Jeffrey D. Ullman, "Mining of Massive Datasets", Cambridge University Press, 2 edition, 2014.

20. Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, "Practical Data Science Cookbook", Packt Publishing Ltd., 2014.

21. Nathan Yau, "Visualize This: The FlowingData Guide to Design, Visualization, and Statistics", Wiley, 2011.

22. Alberto Cordoba, "Understanding the Predictive Analytics Lifecycle", Wiley, 2014.

23. Eric Siegel, Thomas H. Davenport, "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die", Wiley, 2013.

24. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 3rd ed, 2010.

25. Lior Rokach and Oded Maimon, "Data Mining and Knowledge Discovery Handbook", Springer, 2nd edition, 2010.

26. Ronen Feldman and James Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", Cambridge University Press, 2006.

27. Vojislav Kecman, "Learning and Soft Computing", MIT Press, 2010.

28. Jared Dean, "Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners", Wiley India Private Limited, 2014.

Source of Support: Nil, **Conflict of Interest:** None.