Using Machine Learning Techniques to Find the Relationship between University Admission Score and Student Performance－Taking Department of Computer Science and Information Engineering of Tamkang University as an example

Shu-Shiuan Rong
Department of Computer Science and Information Engineering
Tamkang University
Taipei, Taiwan
rss@gms.tku.edu.tw

Abstract: With the impact of the declining birthrate, all higher education institutions are facing enrollment challenges. Collecting useful data can find appropriate students for university admission. In this study, we adopt machine learning techniques like Deep Neural Network(DNN), Random Forest(RF), and Support Vector Machine(SVM) to find the relationship between entrance score and freshman academic performance. Experimental results show that the SVM has the best prediction results on average of total courses in 2-category classification. The RF performs best on mathematical courses and programming courses in 2-category classification. However, on fundamental courses, all methods have the best prediction results in 2-category and perform equally well. Moreover, on other category classifications, like 3-category, 5-category, or 10-category classifications, we cannot find a universal best method. Another result shows that High school GPA has a significant impact on results.

***Keywords*:** Machine Learning, Student Performance, Performance Prediction, Deep Neural Networks, Random Forest, Support Vector Machine.

I. INTRODUCTION

With the reform of education, the General Scholastic Ability Test (GSAT) has become the first major test that most students are exposed to. And two of the three existing admission channels use GSAT score as a criterion. With the impact of the Sub-replacement fertility, the school pays special attention to student abilities in addition to performance. How to choose suitable students to study has also become a major issue for the school.

This work aims to use machine learning models to predict students' academic performance in the freshman year. Our goal is to compare the prediction results of different labeling methods through different machine learning models.

II. LITERATURE REVIEW

In the study proposed by Shahiri et al. in 2015 [1] pointed out that the cumulative grade point average (CGPA) is the most influential feature field, followed by more people using demographic statistics (for example: gender, age, family background, disability) And external evaluation (for example: final exam results for special subjects), and finally behavioral patterns (for example: extracurricular activities, social networking). Shahiri also counted the accuracy of various forecasting methods in the past forecasting research. Among them, the most accurate is the neural network, followed by the decision tree, and then the support vector machine.

1. Deep Neural Network (DNN)

Deep learning is a branch of machine learning. It was proposed by Hinton et al. [2] as early as 2006. The concept is to simulate the neural network of the brain for learning by superimposing multiple hidden layers. In many published studies [3], this method has been used as a method to predict student performance. In the study proposed by Hanan in 2020[8] used three admission scores to predict the cumulative grade point average (CGPA) of students in the first year. In the study, the Artificial Neural Networks (ANN) model was used, and the results showed that the prediction was 79.22% accurate. Accuracy rate. The schematic diagram of the model structure and operation process are as shown in Figures 1 and 2.

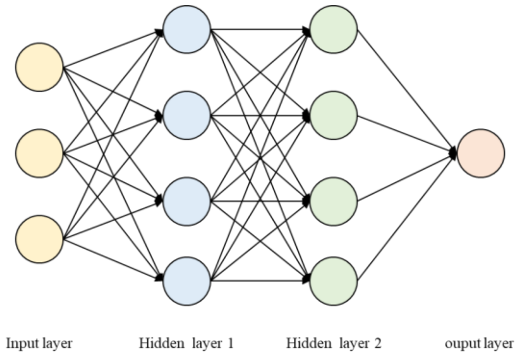


Fig. 1. DNN structure

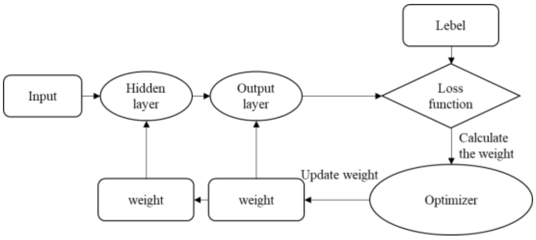


Fig. 2. DNN process

2. Random Forest (RF)

Random forest was first proposed by T.K.Ho of Bell Labs in 1995 [9], and later L. Breiman and Cutler developed and deduced the algorithm of random forest [11]. The basic principle is to combine multiple Classification and Regression Trees (CART), use GINI impurity decision trees, and add randomly allocated training data to greatly improve the final calculation results. It has been widely used in various research related to performance prediction [12]. The study proposed by Anuradha et al. in 2015 used student demographic data and external evaluations to predict the end semester mark (ESM) of students. The decision tree model was used in the research, and the results showed that the accuracy of the prediction was 72.51%. The schematic diagram of the model structure and operation process is as shown in Figure 3.

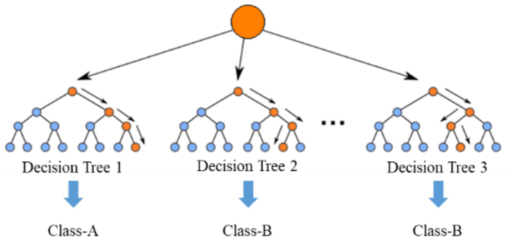


Fig. 3. RF structure

RF summarizes the predictions of all decision trees, and decides the classification results by majority decision. Take Figure 3 as an example, the final classification result is Class-B.

3. Support Vector Machine (SVM)

Vapnik et al. invented the Vapnik–Chervonenkis theory in the 1960s [17]. In 1992, Boser et al. proposed a method of building a nonlinear classifier by applying the kernel technique to the hyperplane of the maximum interval [18]. The predecessor of the current standard (soft interval) was proposed in 1993 by Cortes and Vapnik, and in 1995 Published [19]. It has been widely used in studies related to learning performance prediction[5][13]. Among them, the research proposed by Miguéis[13] used the results of the first two semesters of university and some demographic data (such as gender, marital status, etc.) to predict academic performance after four years. The prediction result of its SVM model reached 92.6%. The schematic diagram of the model structure and operation process is as shown in Figure 4.

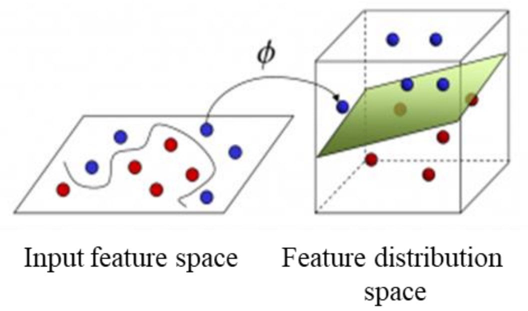


Fig. 4. SVM structure

SVM maps the linearly inseparable samples in the low-dimension space to the high-dimensional space through kernel transformation, and finds a hyperplane to effectively cut these samples (as shown in Figure 4). The distance between the samples on both sides of the hyperplane and the hyperplane itself will determine How good is the model training.

III. DATA PROCESS AND METHODOLOGY

1. Data pre-processing

This work uses data from Center for Institutional Research of Tamkang University and Department of Computer Science and Information Engineering of Tamkang University. Total 328 records. Details in Table 1.

Table 1. Feature and label list

Column name	Type	Pre-processing	Missing
GSAT_Chinese	feature	✓	0
GSAT_English	feature	✓	0
GSAT_Math	feature	✓	0
GSAT_Society	feature	✓	0
GSAT_Science	feature	✓	0
H_PR	feature	✓	0
H_GPA	feature	×	0
Average	label	✓	0
Programming	label	✓	1
Mathematics	label	✓	18
Foundation	label	✓	1

The detail of each feature and label are distributed in the following.

F1. GSAT Scores: GSAT scores are all expressed in grades, which are converted into a unified standard through the percentage of the cumulative number of people.

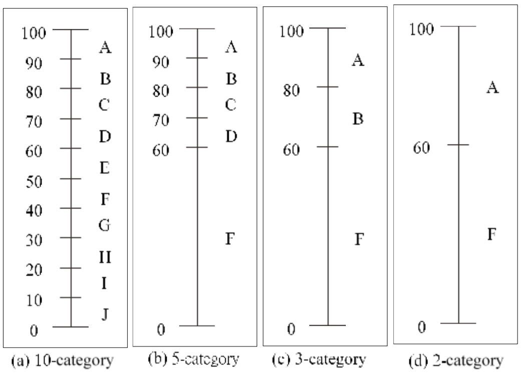
F2. H_PR: H_PR is converted from the school's name. Some private schools can only be collected from the Internet, so that the value is less reliable.

F3. Average: Select the professional subjects according to the list provided by the department office, and then average the professional subjects.

F4. Professional subjects: According to the course classification provided by the department office, the professional subjects are divided into three categories: Programming, Mathematics, and Foundation, and then averaged individually.

2. Labeling methods

The labeling methods are 10-category, 5-category, 3-category, and 2-category, as shown in Figure 5.



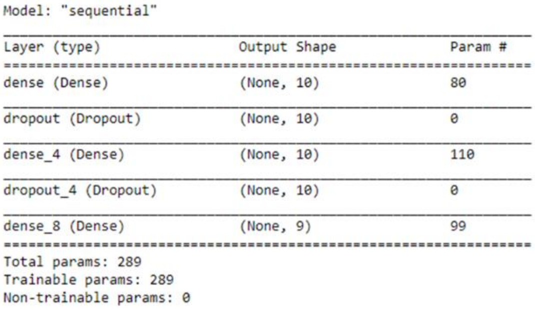
(a) 10-category (b) 5-category (c) 3-category (d) 2-category

Fig. 5. Labeling methods

3. Model settings

1) Deep Neural Network (DNN)

Figure 6 shows the DNN structure of the experiment. To prevent over-fitting, set the dropout to 0.4. The optimization function uses SGD, the loss function uses categorical_crossentropy, the activation function for each hidden layer uses ReLU, and the activation function of the output layer is softmax.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 10)                80
_________________________________________________________________
dropout (Dropout)            (None, 10)                0
_________________________________________________________________
dense_4 (Dense)              (None, 10)                110
_________________________________________________________________
dropout_4 (Dropout)          (None, 10)                0
_________________________________________________________________
dense_8 (Dense)              (None, 9)                 99
=================================================================
Total params: 289
Trainable params: 289
Non-trainable params: 0
```

Fig. 6. The summary of DNN structure

2) Random Forest (RF)

n_estimators is mainly used to limit the number of decision trees and avoid over-fitting. In this experiment, set this parameter to 5. Set the Criterion to “gini”.

3) Support Vector Machine (SVM)

The kernel function mainly assists in mapping features to high dimensions, In this experiment, set this parameter to “poly” to mean linear segmentation in high dimensions. The parameter degree is only related to poly, set to 3. The parameter decision_function_shape set to “ovr”, which means that the n-classification problem is treated as multiple binary classification problems, and each binary problem is divided into two categories: “one” and “rest”.

4. Evaluation

In the experiment, we apply two accuracy rates. The first accuracy rate is the ratio of the full answer rate (bingo), the prediction result, and the test set answer, the “completely consistent” result to the total number of samples in the test set. The second accuracy rate is 1-away, which is the prediction result and the test set answer. The “matching before and after” means “the ratio of the test set answer, plus one and minus one” and the ratio of the test sample to the total number of samples in the test set.

IV. RESULT & DISCUSSIONS

The evaluation of the experimental results is illustrated by the standard values of four items: Average, Programming, Mathematics, and Foundation. The four labeling methods are presented separately in the value of each item. Finally, the importance of features is shown.

1. Average:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-away accuracy into Figure 7.

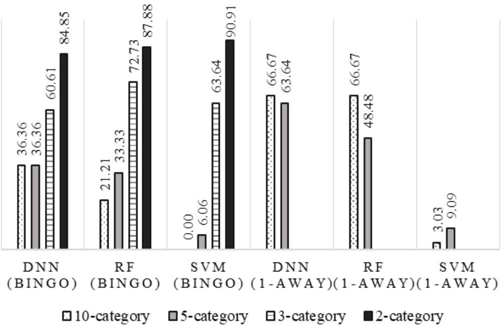


Fig. 7. Result of Average

2. Programming:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-away accuracy into Fig.8.

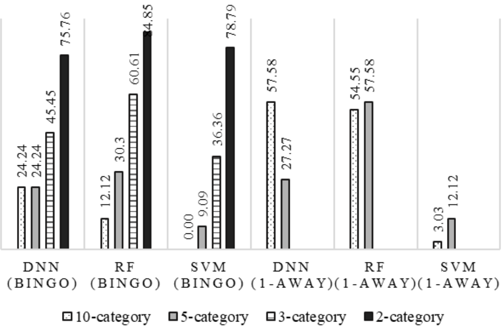


Fig. 8. Result of Programming

3. Mathematics:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-away accuracy into Figure 9.

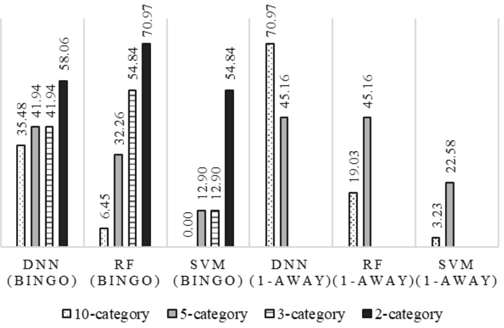


Fig. 9. Result of Mathematics

4. Foundation:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-Away accuracy into Figure 10.

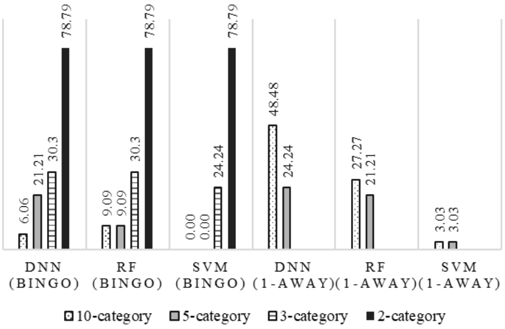


Fig. 10. Result of Foundation

5. Feature Importance in Random Forest:

Obtain the Feature Importance through the feature_importances_ function in RF model. shown in Figure 11.

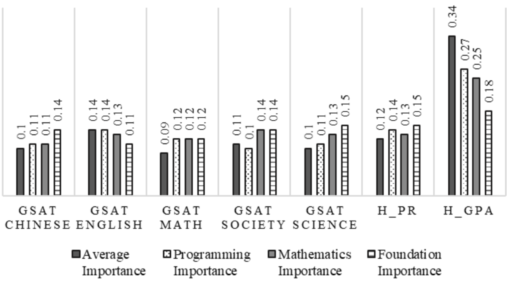


Fig. 11. Result of Foundation

According to the prediction of performance rankings add up organized into Figure 12.

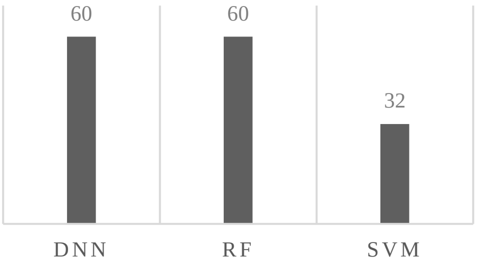


Fig. 12. Prediction of performance rankings add up

Based on the above chart, the following results are obtained:

1) By organizing the experimental results into rankings, we can see that the DNN and RF perform better.
2) After using the 1-Away accuracy, a higher accuracy rate can be obtained without losing the meaning of labeling method. The 1-Away accuracy of "10 cat." is particularly improved.
3) Reducing the label category helps to improve accuracy.
4) There is a correlation between the input feature and performance, especially the high school GPA shows its importance in each performance classification.

V. CONCLUSIONS

We use different labeling methods to predict the results through different models when the features and the number of samples are limited, and obtains a good prediction performance by appropriate use of accuracy evaluation criteria.

Follow-up recommendations from the researchers can begin has aspects:

1) Gather more valuable features, such as student's demographic, external assessments, and personal interest and behavior.
2) If the features are sufficient, also can try different input combinations to improve accuracy.

VI. REFERENCES

[1] A.M.Shahiri, W.Husain, N.A.Rashid, "A review on predicting Student's performance using data mining techniques," Procedia Computer Science, Vol.72, p.414-422, 2015.

[2] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, K.Lang, "Phoneme recognition: neural networks v.s. hidden Markov models,"International Conference on Acoustic, Speech,Signal Processing, p.107-110,1988.

[3] M.Mayilvaganan, D.Kalpanadevi, "Comparison of classfication techniques for predicting the performance of students' academic environment," International Conference on Communications, Computation, Networks and Technologies, Sivakasi, India, p.113-118, 2014.

[4] P.M.Arsad, N.Buniyamin, J.L.A.Manan, “A neural network students' performance prediction model(NNSPPM),” IEEE International Conference on Smart Instrumentation, Measurement and Applications(ICSIMA), Kuala Lumpur, Malaysia, p.1-5, 2013.

[5] F.Marbouti, H.A.D.Dux, K.Madhavan, “Models for early prediction of at-risk students in a course using standards-based grading,” Computers & Education, Vol.103, p.1-15, 2016.

[6] G.Gray, C.McGuinness, P.Owende, “An application of classifcation models to predict learner progression in tertiary education,” IEEE International Advance Computing Conference(IACC), p.549-554, 2014.

[7] E.N.Maltz, K.E.Murphy, M.L.Hand, “Decision support for university enrollment management: Implementation and experience,” Decision Support Systems, Vol.44, No.1, p.106-123, 2007.

[8] A.M.Hanan, “Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems,” IEEE Access, Vol.8, p.55462–55470, 2020.

[9] T.K.Ho, “Random decision forests,” Proceedings of 3rd International Conference on Document Analysis and Recognition, p.14-16, 1995.

[10] T.K.Ho, “The random subspace method for constructing decision forests,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.20, Issue 8, p.832-844, 1998.

[11] L.Breiman, “Random Forests,” Machine learning, Vol.1, Issue 45, p.5-32, 2001.

[12] C.Anuradha and T.Velmurugan, “A comparative analysis on the evaluation of classication algorithms in the prediction of students performance,” Indian Journal of Science and Technology, Vol.8, No.15, p.974-6846,2015.

[13] V.L.Miguéis, A.Freitas, P.J.V.Garcia, A.Silva, “Early segmentation of students according to their academic performance: A predictive modelling approach,” Decision Support Systems, p. 36-51, 2018.

[14] H.Guruler, A.Istanbullu, M.Karahasan, “A new student performance analyzing system using knowledge discovery in higher educational databases,” Computers & Education, Vol.55, No.1, p.247-254, 2010.

[15] S.Natek, M.Zwilling, “Student data mining solution–knowledge management system related to higher education institutions,” Expert Systems with Applications, Vol.41, p.6400–6407, 2014.

[16] S.Fong, R.Biuk-Aghai, “An automated university admission recommender system for secondary school students,” The 6th International Conference on Information Technology and Applications, 2009.

[17] V.N.Vapnik, “The Nature of Statistical Learning Theory. Springer,” New York. for medical diagnosis-application to congenital heart disease. Journal of the American Medical Association.

[18] B.E.Boser, I.M.Guyon, C.Vapnik, “V.N.A training algorithm for optimal margin classifiers.,” Proceedings of the fifth annual workshop on Computational learning theory, COLT.92, p.144, 1992.

[19] C.Cortes, V.Vapnik, “Support-vector networks.,” Machine Learning, Vol.20, p.273-297, 1995.