Research and evaluation of factors affecting the employability of university students after graduation

Trần Hải Long, Nguyễn Vinh Quang*, Hoàng Tiểu Bình***

**ThS. Trường Đại học Sư phạm Hà Nội*
***TS. Trường Đại học CMC*

Received: 5/8/2024; Accepted: 10/8/2024; Published:14/8/2024

***Abstract:** In the context of the labor market is volatile, requires candidates to have the ability and skills to match the needs of the employer. However, most recent graduates have difficulty finding jobs that match their abilities. At the same time, assessing the ability of candidates and selecting the right candidates for the job is a huge challenge for the employers. In this paper, we introduce a new approach to early prediction the employable of students after graduation by using some machine learning methods: Stochastic Gradient Descent, Decision Tree, Support Vector Machine and Artificial Neural Network. Experimental results show that, ANN and RF methods bring the best performance to predict student employment status with accuracy up to 70%.*

***Keywords:** Job vacancy, data mining, machine learning...*

1. Introduction

With the strong development of Science - Engineering in general and information technology in particular, the increasingly of popular teaching and learning support software applications have contributed to a rich source of educational data that allows researchers to delve into the topic of educational data mining and apply it more widely.

According to survey data about nearly 2,000 students from the graduation year of 2020 and 2021 after 1 year of graduation conducted by the Center for Quality Assurance of Hanoi National University of Education, there are 10% of students who still can't found a job. The remainder is a group of students working in government and private enterprise environments with 40% and 50% of the total students surveyed, respectively. Hence, if it is possible to apply machine learning and data mining methods to exploit important information in student survey data, it can bring a lot of benefits to the University in making early decisions to overcome unemployment among students after graduation as well as lecturers can actively adjust the teaching content so that after graduation, students can meet the requirements of employers.

2. Content

2.1. Related work

The predicting model about career field that engineering students will choose after graduation is made by Akanksha Pandey et al (Pandey et al., 2022) by using popular classification algorithms such as K-nearest neighbors, SVM, Logistic regression, Decision trees, Gaussian Naive Bayes and Artificial Neural Networks. The study shows that K-nearest neighbors algorithm gives the best results with an accuracy of 63.4% while the artificial neural network model has the lowest accuracy with 45.45%.

Teng Guo et al (Guo et al., 2019) have researched and built a MAYA (Multi-mAjor emploYment stAtus) model to predict the employment status of students after graduation with bias. After the authors used GAN to solve the data imbalance problem, the MAYA model shows outstanding results with up to 88% accuracy.

2.2. Proposed methods

In this study, we propose several methods to build models predicting student employment status from their learning data by selected machine learning-based classifiers. Besides, the original dataset is still kept to build models and compare performance with the other techniques. To have the most intuitive view of the ability of machine learning methods to predict student's employment status, we use artificial neural network and compare with popular machine learning-based classifiers such as Gaussian Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) to find the best method via the confusion matrix, accuracy, precision, recall, Cohen's Kappa and F1-score. Figure 1 describe the research implementation process.

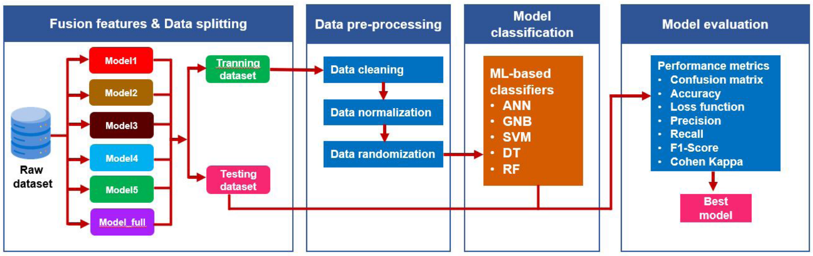


Figure 1: Proposed research method.

2.3. Data collection

The dataset in this research mainly about students from the graduation year of 2020 and 2021. We collecting 19 information fields about student's personal data such as gender, hometown, working environment and student's university score for example foreign language, specialized foreign language, ideology and politics, pedagogical skills, internship and employment status. The goal is trying to explore the potential importance and the internal correlation of these attributes as well as to build a machine learning model for predicting the employment rate of students after graduation.

2.4. Data analysis

The percentage of students graduating with the very good honors accounts for 50.6%, followed by students from the good group with 32.9%, the third place is excellent students by 15.1% and last one are 1.4% for average students.

Next, the issue that the society as well as the university most concerned about is the working environment and employment situation of students after graduation. The working environment of alumni is very diverse, but it also gives a good signal when the students with the right major working environment account for the most about 72.8%. The second position is students working in environment that related with their majors take 12.8%. Followed are the rate of students belong to groups such as not related to major, keep learning and no working environment respectively 5.7%, 4.9% and 3.8%.

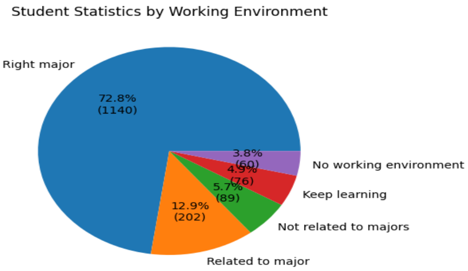


Figure 2. Student's statistics by working environment.

In addition, the employment status of students after one year graduation, as shown in Figure 3. Basically, the employment status of students is divided into three main groups: private enterprises, government and jobless. In which, the number of students working in private enterprises accounts for the mainly about 51%, followed are government officials with the rate of 40% and remain are 9% for the jobless students.

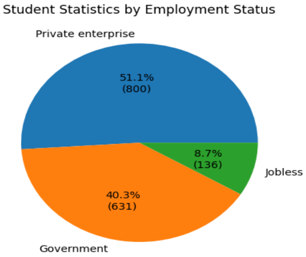


Figure 3. Student statistics by employment status.

Moreover, to get an overview of the employment status of students after graduation, we summarize the employment and unemployment rates of students by graduation grade as shown in Figure 4. It is surprising that the group of students with the least unemployment rate are students graduation with very good grades, not excellent. Beside, students graduate with the very good grade are more likely to work in private enterprises than excellent graduates. The same happens with the group of good and average graduates when the majority of students work in private businesses.

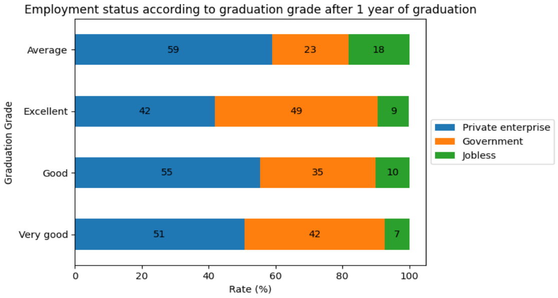


Figure 4. Employment status according to graduation rating after 1 year of graduation.

Last but not least, we create a graph of the correlation between the information fields in the collected dataset, as shown in Figure 5.

It can be seen that subjects in the same field are highly correlated. For example, subjects in foreign language skills have a correlation coefficient of about 0.53 - 0.65, while subjects in the field of education have a correlation coefficient of about 0.4.

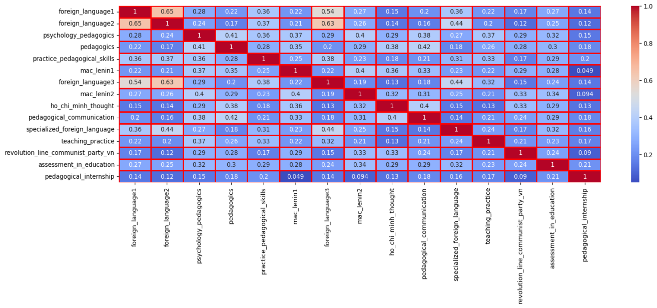
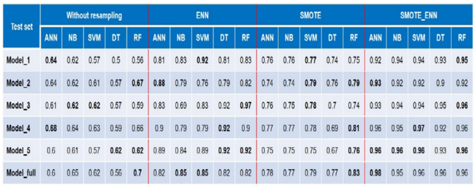
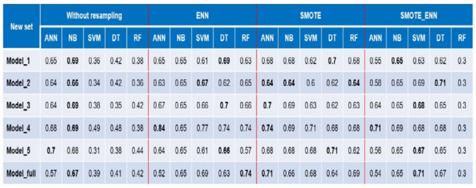


Figure 5. The correlation coefficient between data fields.



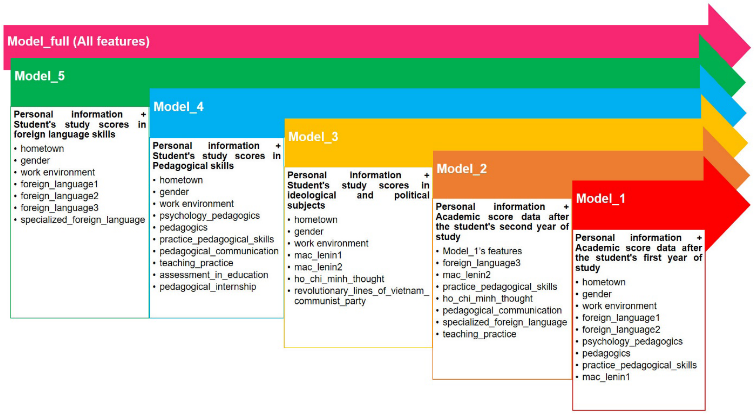
Test set	Without resampling					ENN					SMOTE					SMOTE_ENN				
	ANN	NB	SVM	DT	RF	ANN	NB	SVM	DT	RF	ANN	NB	SVM	DT	RF	ANN	NB	SVM	DT	RF
Model_1	**0.64**	0.62	0.57	0.5	0.56	0.81	0.83	**0.92**	0.81	0.83	0.76	0.76	**0.77**	0.74	0.75	0.92	0.94	0.94	0.93	**0.95**
Model_2	0.64	0.62	0.61	0.57	**0.67**	**0.88**	0.79	0.76	0.79	0.82	0.74	0.74	**0.79**	0.76	**0.79**	**0.93**	0.92	0.92	0.9	0.92
Model_3	0.61	**0.62**	**0.62**	0.57	0.59	0.83	0.69	0.83	0.92	**0.97**	0.76	0.75	**0.78**	0.7	0.74	0.93	0.94	0.94	0.95	**0.96**
Model_4	**0.68**	0.64	0.63	0.59	0.66	0.9	0.79	0.79	**0.92**	0.9	0.77	0.77	0.78	0.69	**0.81**	0.96	0.95	**0.97**	0.92	0.96
Model_5	0.6	0.61	0.57	**0.62**	**0.62**	0.89	0.84	0.89	**0.92**	**0.92**	0.75	0.75	0.75	0.67	**0.76**	**0.96**	**0.96**	**0.96**	0.93	**0.96**
Model_full	0.6	0.65	0.62	0.56	**0.7**	0.82	**0.85**	**0.85**	0.82	0.82	0.78	0.77	0.79	0.77	**0.83**	**0.98**	0.95	0.96	0.96	0.96



New set	Without resampling					ENN					SMOTE					SMOTE_ENN				
	ANN	NB	SVM	DT	RF	ANN	NB	SVM	DT	RF	ANN	NB	SVM	DT	RF	ANN	NB	SVM	DT	RF
Model_1	0.65	**0.69**	0.36	0.42	0.38	0.65	0.65	0.61	**0.69**	0.63	0.68	0.68	0.62	**0.7**	0.68	0.55	**0.65**	0.63	0.62	0.3
Model_2	0.64	**0.66**	0.34	0.42	0.36	0.63	0.65	**0.67**	0.62	0.65	**0.64**	**0.64**	0.6	0.62	**0.64**	0.58	0.65	0.69	**0.71**	0.3
Model_3	0.64	**0.69**	0.38	0.35	0.42	0.67	0.65	0.66	**0.7**	0.66	**0.7**	0.69	0.63	0.62	0.63	0.64	0.65	**0.68**	0.65	0.3
Model_4	0.68	**0.69**	0.49	0.48	0.38	**0.84**	0.65	0.77	0.74	0.74	**0.74**	0.69	0.71	0.68	0.68	**0.71**	0.69	0.68	0.68	0.3
Model_5	**0.7**	0.68	0.31	0.38	0.44	0.64	0.65	0.61	**0.66**	0.57	0.68	0.68	0.68	**0.71**	0.62	0.56	0.65	**0.67**	0.65	0.3
Model_full	0.57	**0.67**	0.39	0.41	0.42	0.52	0.65	0.69	0.63	**0.74**	**0.71**	0.66	0.68	0.64	0.69	0.54	0.65	**0.71**	0.67	0.3

2.4. Create predicting model

To predict the employment situation of students, we create six models in which three models to early predict the students employment status from learning data of first academic year (Model_1), second year (Model_2), fourth year (Modelfull) and three more models have been created by field of subjects about thought and politics (Model_3), pedagogical skills (Model_4), foreign language skills (Model_5). Figure 6 shows how the features are used for the respective models.



Model_full (All features)

Model_5

Personal information + Student's study scores in foreign language skills
- hometown
- gender
- work environment
- foreign_language1
- foreign_language2
- foreign_language3
- specialized_foreign_language

Model_4

Personal information + Student's study scores in Pedagogical skills
- hometown
- gender
- work environment
- psychology_pedagogics
- pedagogics
- practice_pedagogical_skills
- pedagogical_communication
- teaching_practice
- assessment_in_education
- pedagogical_internship

Model_3

Personal information + Student's study scores in ideological and political subjects
- hometown
- gender
- work environment
- mac_lenin1
- mac_lenin2
- ho_chi_minh_thought
- revolutionary_lines_of_vietnam_communist_party

Model_2

Personal information + Academic score data after the student's second year of study
- Model_1's features
- foreign_language3
- mac_lenin2
- practice_pedagogical_skills
- ho_chi_minh_thought
- pedagogical_communication
- specialized_foreign_language
- teaching_practice

Model_1

Personal information + Academic score data after the student's first year of study
- hometown
- gender
- work environment
- foreign_language1
- foreign_language2
- psychology_pedagogics
- pedagogics
- practice_pedagogical_skills
- mac_lenin1

Figure 6: Features used to train prediction models.

2.5. Evaluate prediction models

The prediction results of the models after applying different machine learning methods are shown in. When compared with other machine learning methods, ANN models are not too superior, but show their strengths in some specific cases. With the test dataset, ANN and RF methods give the highest accuracy. Model_full with RF bring 70% of accuracy and Model_4 with ANN is 68%.

Compare the accuracy of all models when applying different machine learning methods and resampling techniques on the test dataset.

3. Conclusion

In this article, we have analyzed and given statistics on the learning and employment status of students at Hanoi National University of Education. At the same time, an approach in building models to predict the working status of students after one year of graduation was introduced. The results show that the models predicting early employment status from first year student and second year student data have an accuracy of 64% for ANN and 67% for RF when predicting the test dataset. The best model is Modelfull using RF on the test dataset with prediction accuracy of 70%.

This work has received support from the project "Research and evaluate factors affecting ability employability of students after graduation - SPHN23-15" funded by Ministry of Science and Technology of the Socialist Republic of Vietnam.

References

1. Casuat, C. D. (2020). Predicting Students' Employability using Support Vector Machine: A SMOTE-Optimized Machine Learning System. International Journal of Emerging Trends in Engineering Research, 8(5), 2101–2106. https://doi.org/10.30534/ijeter/2020/102852020

2. Casuat, C. D., & Festijo, E. D. (2019, December). Predicting Students Employability using Machine Learning Approach. 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS). https://doi.org/10.1109/icetas48360.2019.9117338

3. Chawla, N. v, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. https://doi.org/10.48550/ARXIV.1106.1813