International Journal of Mathematics and
Computer Science, **17**(2022), no. 1, 365–375

Classification models for employability of statistics and related fields graduates from Thailand universities

Thammarat Panityakul[1], **Wisanuwee Suriyaamorn**[2], **Ronnason Chinram**[1], **Noodchanath Kongchouy**[1]

[1]Division of Computation Science
Faculty of Science
Prince of Songkla University
Hat Yai, Songkhla 90110, Thailand

[2]Doctor of Philosophy Program in Data Science
College of Digital Science
Prince of Songkla University
Hat Yai, Songkhla 90110, Thailand

email: thammarat.p@psu.ac.th, wisanuwee.nine@gmail.com,
ronnason.c@psu.ac.th, nootchanath.k@psu.ac.th

(Received July 7, 2021, Accepted September 2, 2021)

Abstract

Thailand's graduate employment data was collected continuously by the Ministry of Higher Education, Science, Research, and Innovation which resulted in a huge, incomplete, and high dimensional data. In this paper, we study a model to predict employability and compare performance usind statistical as well as machine learning models. Our results show that the significant variables are: graduation year, education level, major of study, type of university, ranking, and region. The four models compared are: logistic regression, decision tree, random forest, and KNN models. The comparative result shows that the random forest and KNN models are suitable for employability prediction with 70.554% and 70.158% accuracy, 0.746, and 0.762 in AUC. The

Key words and phrases: Graduates employment, Logistic regression, Classification, Machine learning.
AMS (MOS) Subject Classifications: 62J12.
Corresponding author: Noodchanath Kongchouy noodchanath.k@psu.ac.th
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

Random Forest Model shows the important variables for prediction are: university ranking, master's degree, statistics major, a national university, and the north-east region. Our study implies that the random forest and KNN models can be used to predict employment status for graduating students, the time when they should apply for work, and the academic staff they worked under.

1 Introduction

Employment after university graduation is a big problem in Thailand's society. Despite the report from the Ministry of Higher Education, Science, Research and Innovation (MHESI) showing that during the last 3 years, a student graduating from a higher education institution has a 68%, 69%, 68% chance of employability [8], the news always reports an employment problem [14, 16]. According to the National Statistical Office [10], the most needed occupations in Thailand are service staff, mathematical, statistical and health care careers, in that order. Moreover, if the focus is on specialized careers, then the mathematical and statistical ones are required most in the labor market. Such a background not only is essential for mathematics and statistics fields but also for other careers such as actuarial science and accounting.

As a government organization, MHESI collects data on graduate employment from many sources but this data is huge, incomplete, and multidimensional. The appropriate approach to analyze this data is not only from a statistical sense but also using machine learning methods which are famous in real-life problems such as financial, business, health or climate data. An example of machine learning is to predict the behavior of the stock market [2]. In this paper, statistical and machine learning methods are used to predict the employment outlook for Thailand's graduating university students in Statistics and related fields.

Many kinds of research dealing with graduating students' employment exist but they usually focus on data collected from only one university as it is easy to collect. Naturally, this won't reflect what is happening at a different university [1, 5, 18]. Moreover, the research that uses a machine learning method to analyze students' employment does not constitute enough evidence to explain the result. Consequently, the purpose of this work is to compare and explain the result from a machine learning method as well as a statistical one.

2 Methods

Classification is one type of a predictive model that a response variable is qualitative or categorical. Predicting categorical data can be referred to predict samples from data that assign some categories in those samples. On the other hand, it predicts the probability of each category from a coefficient as linear regression [4]. We use the R software version 3.6.2 to design and implement statistical and machine learning models.

The simplest and commonly used classification model is the Logistic Regression Model. The response variable for logistic regression is a binary dependent variable such as employment or unemployment. This model will predict a probability of the interested event π, also called the odds ratio. Logistic regression follows (equation (2.1)). From [18], the advantage of logistic regression is that the coefficient or odds ratio can describe the rate of change in the probability of employment status. Also, a simple model can be better than a complicated model as in [1] that compares NB, J48, MLP, KNN and logistic regression to predict employment status. As a result, the logistic regression performs better than other models.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_i X + \epsilon_i, \tag{2.1}$$

where $\frac{p(X)}{1-p(X)}$ is odds or probability that an event occurs, β_i is the regression coefficient and ϵ_i is an error term.

The Decision Tree Model looks like a tree flowchart where nodes represent the criteria for variables, branches represent the criteria's results, and leaves represent the categories of a sample [4]. This is a simple model from a machine learning model standpoint that can be interpreted by a tree diagram. In [15, 9], the authors applied the Decision Tree Model and other models to predict employability of graduating students as a decision tree standing for an appropriate model to predict employability with high accuracy and ease. Moreover, the Decision Tree Model can be applied in other education situations such as using it to predict students performance by analyzing their academic status [4] or predicting the university students dropout from the standpoint of background and status [12] thus developing and implementing a new strategy for the educational system of their institutions.

The Random Forest Model is a development of the Decision Tree Model with the idea to construct a set of decision tree models by each decision tree model containing a random of sample set and variable from bootstrap sampling. After constructing a decision tree model, every model and clas-

sified category of sample are combined by using majority vote. To find an appropriate model, we can try to search the number of the optimal values of the Random Forest Model parameters to get some important variables for classification in the model. Parmar et al. [11] studied sentiment analysis on movie review by using the Random Forest Model and concluded that random forest performed well with tuning hyperparameters needing special attention to get an appropriate model. The application of the Random Forest Model to educational research is to predict the graduates' employment from factors such as gender, student degree or dropouts [3].

The k-Nearest Neighbors (KNN) Model is a simple model by learning from the data. The classification of a new unknown sample is compared with the sample set in the data. Then, the most similar sample from data with the unknown samples will classify that sample [4]. To define the similarities among samples, this model uses a distance function to measure the distance among samples. The Euclidean distance (2.2) is commonly used. In addition, the KNN Mmodel compares data samples, normally using an odd number of them in order to classify the output on the basis of a majority vote. On the other hand, the number of k-samples to classify new data may affect the performance of the model to find an optimal value of k to form a good model. Rahman et al. [13] studied both a supervised and unsupervised learning model by comparing decision tree, the naïve Bayes and the KNN model to predict employment status and concluded that the KNN model achieved the highest accuracy.

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2} \tag{2.2}$$

Pre-processing:

The objective was to construct a classification model that predicts the status of graduating students by using data sourced from MHESI. The data was collected after contacting all universities in Thailand from 2013 to 2017 and consisted of 703777 records. The first step was to filter these records arranged by year, students with bachelor's degree in applied mathematics, applied statistics, and statistics major. Then the data underwent a cleaning process to remove those with undefined category, missing value and outlier by applying the KNN imputation method. Minakshi et al. [7] showed that KNN imputation was better than other imputation methods.

Using the data received from every university in Thailand, we can extract a variable for explaining the student employability by using the different regions, university ranking, and the position of a university among 9 that the government classifies as national research universities. After cleaning the

data set, the variable to predict student employment status are: university name, graduating year, education level, major of study, type of university, ranking, national research university, and region of university. The details are described in table 1.

Table 1: List of variables

No.	Variable	Value
1	University name	BUU, CU, CMU,..., TU, TSU
2	Year of graduate	2013 - 2017
3	Education level	Bachelor, Master, Ph.D.
4	Study major	Apply Math, Apply Stat, Statistic
5	Type of university	Government, National university
6	Ranking	In THE ranking, Not in THE ranking
7	National research university	Research, Not research university
8	Region	North, North East, Central,..., South

Variable selection and simple visualization use a confident interval plot for adjusted means for each variable in the univariate model and full model. Then select the significant variable to construct a classification model. The advantage of this approach is clearly in seeing the difference in confident intervals in every variable by using contrast matrices. In addition, this is not considered with a baseline on a dummy variable [17] and can be applied to the classification problem with a confident interval for adjusted proportions [6]. Figure 1 shows the selected variable from the confident interval plot. We can see which confidence interval does not cross the overall means line; that is, this factor gives the different proportions of employment status. The selected variables from figure 1 are: year of graduation, education level, major of study, type of university, ranking, and region. The data set used in this study consists of 3,969 samples and 6 variables.

The Effective Classification Model not only predicts the data that is used to construct model but also predicts the unknown data. To improve the performance of the classification model and avoid over fitting problems, the data separation approach is used to improve the model. Training data and testing data are separated to a 70:30 ratio before constructing a classification model. Moreover, k-folds cross validation and parameter tuning are used to evaluate the classification model with complicated models.

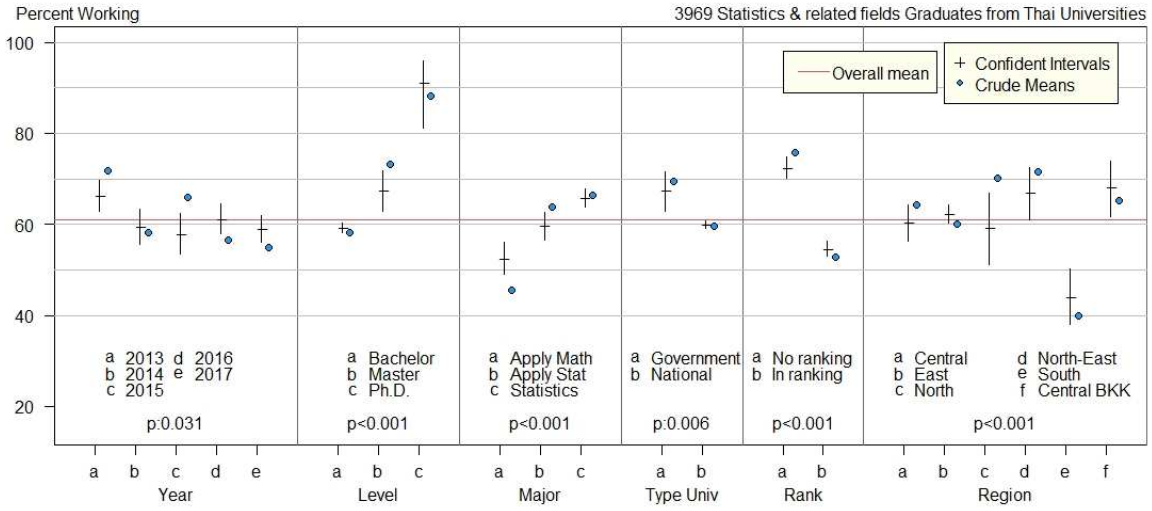

Figure 1: Confident interval plot for classification model.

Performance measure:

This paper uses the measurement to evaluate the performance of a classification model are accuracy, precision and area under ROC curves (AUC). Accuracy is one of the easy measurements that is used to measure correct predict values from a number of all predictions. The calculations are usually presented in percentages. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{2.3}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Accuracy may not yield enough information to compare a model. Therefore, other parameters are also taken into consideration.

Precision is the measurement of predictive model performance. It is considered to be a focused outcome from the predicted result and gets employed by using the formula:

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

The area under the receiver operating characteristic curve (AUC) is a famous measurement for the classification model. A receiver operating characteristic

curve is considered with true positive rate (TPR) and false positive rate (FPR), also known as sensitivity and 1-specificity. For an effective model, a curve will be higher that orthogonal and the area under the curve will close to one value.

3 Results and Discussion

In this section, we compare the results for the four methods. The testing mechanism uses a 10-fold cross-validation with the R programming software. In cross-validation, the data are separated into 10 pieces, 9 of which are used for modeling and the last one is used to test the model. We then repeat the same process by changing the test data to new pieces until the last piece of test data. The data have been applied with four models: logistic regression, decision tree, random forest and KNN model. The performances of classification models are compared for accuracy, precision and area under ROC curve (AUC).

Table 2: Classification accuracy and precision from four models

Methods	Accuracy (%)		Precision (%)	
	Training data	Testing data	Training data	Testing data
Logistic Regression	65.047	65.827	67.136	68.939
Decision Tree	64.795	66.919	70.499	73.913
Random Forest	70.554	68.262	70.921	70.862
KNN	70.158	67.003	70.928	70.125

Table 2 shows the percentage of accuracy from the classification models. Also, the table shows a comparison of training data and testing data to conclude the performance of models. Moreover, the effective model could be predicting new data that the model didn't know similar to a performance from data that use to construct a model. The result from the Random Forest Model achieved the highest accuracy percentage compared to other models, 70.554% in training data and 68.262% in testing data. The second accuracy percentage is from the KNN model, 70.158% in training data and 67.003% in testing data. The first and second models have a few differences from the accuracy percentage. Furthermore, table 2 also shows a percentage of precision to predict "employed" status from the classification model. The highest precision is from KNN model, 70.928% in training data and the second is from Random Forest Model, 70.921% in training data. For testing data, the

highest precision is from decision tree model, 73.913% and the second is from Random Forest Model, 70.862%. The precision represents a performance of the classification model as a focus outcome. In this case, predicting employability status from decision tree, random forest and KNN models are quite high.

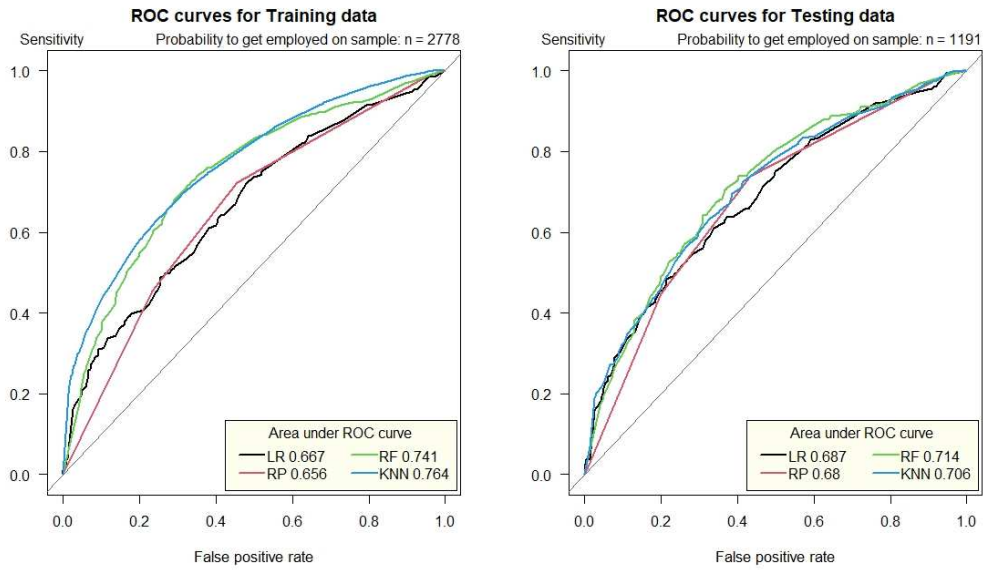

Figure 2: Receiver operating characteristic curve from training data and testing data.

Figure 2 shows the receiver operating characteristic curve with four models and compares them with training data and testing data. ROC curve of random forest and KNN model in training data are higher than logistic regression and decision tree models. That makes the area under the curve also high. Consider an area under ROC curve. If an area is close to one, then that represents an effective classification model. The result from the KNN model achieved the highest value of AUC, 0.762 in training data and 0.733 in testing data. The second AUC is a Random Forest Model, 0.746 in training data and 0.727 in testing data.

From both classification models, there were differences. The Random Forest Model was complicated and took a long time to construct a model by using a bootstrap method that was hard to represent but showed an important variable to use in the model. However, the KNN model was a simple

model learning from the data sample but unable to show the model's structure and the important variable as it didn't use the information from data for construction.
The most important variables from the Random Forest Model were: ranking, master level, Statistics major, national university, and North-East region, respectively. As we were unable to know the rate that affects the opportunity of employability, we had to consider this model with a statistical one to compare.

Back to figure 1, we can see that a high adjusted mean of employment rate from a confident interval plot is: Ph. D. level, no ranking, master level, Statistics major, and government university. The result is similar to the Random Forest Model which considers an important variable. The Random Forest Model provides an important variable in: ranking, master level, statistics major, and national university. This is an important variable for the opportunity to get employment. The reason that an important variable from the Random Forest Model is different from the Logistic Regression Model is an algorithm to set up baseline for each categorical variable. Besides, we can interpret an important variable from the Random Forest Model with the result from figure 1.

4 Conclusion

The main issue of this study was predicting employability. This study attempts to explain the significant variables and visualization from classification models based on sample data obtained from MHESI. In order to construct an effective model from different methods, we designed and implemented to predict employment. Our result showed that the Random Forest Model was superior to the KNN Model with accuracy, precision percentage, and area under ROC curve. The result on the confident interval plot was similar to descriptive statistics to show the difference in employment rate from each variable level and with important variables from the predictive model. As the Random Forest and KNN models can be used to predict employability, the person in charge at a university can then advise his/her students to follow the right paths to secure a good future.

Acknowledgment. We would like to thank Prof. Don McNeil for his helpful guidance.

References

[1] M. T. R. A. Aziz, Y. Yusof, Graduates employment classification using data mining approach, AIP Conference Proceedings, **1761**, (2016), ID 020002.

[2] R. Choudhry, K. Garg, A Hybrid machine learning system for stock market forecasting, Proceedings of World Academy of Science, Engineering and Technology, (2008), 315–318.

[3] F. J. Garcia-Peñalvo, J. Cruz-Benito, M. Martin-González, A. Vázquez-Ingelmo, J. C. Sánchez-Prieto, R. Therón, Proposing a machine learning approach to analyze and predict employment and its factors, International Journal of Interactive Multimedia and Artificial Intelligence, **5**, no. 2, (2018), 39–45.

[4] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning with applications in R, Springer, New York, 2013.

[5] B. Jantawan, C.-F. Tsai, The application of data mining to build classification model for predicting graduate employment, International Journal of Computer Science and Information Security, **11**, no. 10, (2013), 1–7.

[6] N. Kongchouy, U. Sampantarak, Confidence intervals for adjusted proportions using logistic regression, Modern Applied Science, **4**, no. 6, (2010), 2–7.

[7] Minakshi, R. Vohra and Gimpy, Missing value imputation in multi attribute data set, International Journal of Computer Science and Information Technologies, **5**, no. 4, (2014), 5315–5321.

[8] Ministry of Higher Education, Science, Research and Innovation, The increase rate and percentage of students in the academic year 2008-2018 classified by educational level and fileds, (2019). http://stiic.sti.or.th/stat/ind-lf/ind-lf-g001/lf-t002/

[9] T. Mishra, D. Kumar, S. Gupta, Students' employability prediction model through data mining, International Journal of Applied Engineering Research, **11**, no. 4, (2016), 2275–2282.

[10] National Statistical Office, A study of labor demand trends in the labor market in Thailand between 2017-2021, Forecasting Bureau, National Statistical Office, (2017). http://www.nso.go.th/sites/2014/DocLib13

[11] H. Parmar, S. Bhanderi, G. Shah, Sentiment mining of movie reviews using random forest with tuned hyperparameters, International Conference on Information Science, Kerala, (2014).

[12] M. Quadri, N. V. Kalyankar, Drop out feature of student data for academic performance using decision tree techniques, Global Journal of Computer Science and Technology, **10**, no. 2, (2010), 2–5.

[13] N. A. B. A. Rahman, K. L. Tan, C. K. Lim, Supervised and unsupervised learning in data mining for employment prediction of fresh graduate students. Journal of Telecommunication, Electronic and Computer Engineering, **9**, (2017), 155–161.

[14] P. Rujivanarom, Fresh graduates hardest hit as joblessness rises, The Nation Thailand, (2015).
http://www.nationthailand.com/national/30260918

[15] M. A. Sapaat, A. Mustapha, J. Ahmad, K. Chamili, R. Muhamad, A data mining approach to construct graduates employability model in Malaysia, International Journal on New Computer Architectures and Their Applications, **1**, no. 4, (2011), 1086–1098.

[16] Thaipbsworld, Ministry to launch volunteer projects to ease graduate unemployment, ThaiPBS, (2019). http://thaipbsworld.com/ministry-to-launch-volunteer-projects-to-ease-graduate-unemployment

[17] P. Tongkumchum, D. McNeil, Confidence intervals using contrasts for regression model, Songklanakarin Journal of Science and Technology, **31**, no. 2, (2009), 151–156.

[18] X. Xu, W. Zhang, Analysis of the influence factors of the ability of graduate employment based on the Logistic Regression Model, International Journal of Control and Automation, **8**, no. 9, (2015), 405–412