ZKG INTERNATIONAL ISSN: 2366-1313

Data Mining Techniques for forecasting College Graduates' Employment

[1]S. SUSMITHA, [2]CH.SAI LAKSHMI

[1,2]Assistant Professor, Dept of CSE, Megha institute of engineering and Technology for women, Ghatkesar (T.S)

Abstract: *It is the students' dream to secure a job right after graduation. However, there are factors that hinder their employability. With the popularity and expansion of better education enrolment in China, college student employment has been a focus of public attention. In this research, taking the 2016 graduates of the Guilin University of Technology as an example, statistical mining techniques are performed to predict employment using five influencing factors. The Gini index was calculated based on the CART algorithm, and a selection tree was constructed. In addition, the Random Woodland algorithm is used to improve the accuracy of the employment forecast value. After collecting, cleaning, and replacing records, 496 employment records were obtained, 70% of which were taken for school samples. The bullet version was tested using the latest samples, and the accuracy reached 81%. Finally, academic compliance and graduation eligibility functions are considered major elements of university student employment. An additive version of the choice tree and random forest region offers a new approach to employment forecasting, which is feasible and adaptable to the career direction of universities.*

Keywords: *Employment prediction, data mining, decision tree, CART algorithm.*

I. INTRODUCTION

There are many reasons why a graduate may want to work. But some of them choose to continue their studies as they have a better aim to polish their skills further so that they can enjoy the process without any hindrance. By reducing the unemployment rate, the kingdom can contribute more to the economy than the goods or services that will be produced. In addition, it can reduce the number of homeless people and the crime rate. Therefore, taking measures to reduce the unemployment rate is very important. Unfortunately, few studies have been conducted on predicting employment status using follow-up research statistics from the Malaysian Ministry of Higher Education. Studying the proposals can provide valuable information for comparing educational outcomes, which can be used for similar changes or the development of institutions to produce good, first-class graduates. This research

ISSN: 2366-1313

focused on tracking the employability of all undergraduate graduates from a public university in Malaysia from 2015 to 2018. This research was conducted to prompt graduates to discover the factors that can contribute to reducing the amount of unemployment in Malaysia. With supervised and unsupervised data mining strategies, job predictions are based on completely different attributes. This study was conducted because additional entries could be included in the analysis. This approach can uncover the underlying variables that may explain patterns of graduate employability as a way to provide more comprehensive factors. In addition, hidden relationships between inputs can be identified by neural networks, which process complex relationships between inputs, a system that can rarely be realized through experimental design. Additionally, a robustness method was included to reduce many levels for categorical inputs within the data set. A spread of variable selection was used in this study to improve the raters' overall performance in predicting job credibility. In addition, this format can select the best raters to assess job credibility. Additionally, new statistics can be obtained when graduates are grouped into similar organizations based on their willingness to direct curricular activities and communication technology (ICT) skills.

Many authors have used decision trees to analyze and predict the employment of university graduates. Liu et al. integrated a database of basic information, a database of rankings, and a database of employment data of university graduates. The ID3 decision tree rule set explored the key factors influencing graduate employment instruments.

Zhang et al. A type decision tree based entirely on the C5.0 algorithm was fitted, and factors influencing the employment path of college graduates, such as their academic performance and CET-4, were explored. And CET-6 score, basis, principal, subjects failed, area of employment, and city.

Tang et al. collected the employment information of college graduates of traditional Chinese medicine, studied the influencing factors of employment based on C4.5 algorithm, and further used random forest algorithm to improve the accuracy of employment forecast.

Data mining techniques have proven their application for analysing employment problems. According to validity, we conclude that logistic regression, decision trees, and random forest algorithms are the best information mining strategies for employment research. However, most researchers specialize in influencing

ISSN: 2366-1313

employment factors and accurately predicting employment trends. Most of them better consider factors, including college student's academic performance, English proficiency, and laptop skills, while a few focus on factors such as their academic performance. Includes periodic activity in circle, family history, and university associations. These elements affect student employability and predictability.

II. LITERATURE SURVEY

Due to the historical development stage and reality, there are few direct studies on the issue of college graduates in foreign countries, and most of the studies regard them as a special group of employment in the rural labour market and place them in the urban-rural mobility of the population. College graduates facing the rural population flow from urban to rural employment is a special performance, and rural population in urbanization is opposite in the direction of the flow from the countryside to city, as a result.

X. Wang et al. explained in the recent years, some domestic experts and scholars have studied the employment of college students from two perspectives: the structure of personnel training and the setting of disciplines.

Hira et al. think the university education system reform and graduate employment system reform of institutions of higher learning are not synchronized, autonomy is relatively small, and admission pipe and professional setting and employment market demand eventually form the situation of college graduates market supply and demand imbalance, the serious influence college students' employment. Some scholars, from the perspective of college students themselves, believe that contemporary college students' employment concept is backward, the professional knowledge system needs to be improved, employment psychology is not mature, and practical operation ability is relatively low, which needs to be further improved.

Since the beginning of the twenty-first century, China has formulated a series of employment policies and management measures to encourage college students to work at the grassroots level and in the central and western regions. The state has carried out projects such as "Selecting and hiring college graduates to work in Villages," "Supporting education, agriculture, medical services and poverty alleviation," "College students volunteer to serve the western Region," and "Rural compulsory education stage school teacher position plan." At the same time, college

ISSN: 2366-1313

graduates are encouraged and supported to find jobs in small and medium-sized enterprises and start their own businesses.

Employment Status

The research done on employment status by [4] classified the status into graduates who are employed, furthering their studies, upgrading skills, waiting for work placement, and unemployed. The study by [5] reported that different groups of age have a different characteristic of the labour market, such that some age groups are more or less likely to be employed, unemployed, or outside the labour force. It also stated that the inclusion or exclusion of the working-age population affects the result of employability.

Factors Contributing to Employment Status

Inputs of age, CGPA, and gender have been found to be significant attributes to the low employability of Malaysian graduates in an early detection and prediction model that merged the data set from Tracer Studies and Malaysian Soft Skills Scale (My3s). In contrast, the result by [6] found that gender and final grades of graduates contribute to perceived employability. Other than that, degree types are found to be a determinant with a positive influence on graduates' employability. It contrasts with the study done by [75], which found that program outcome attainment is not a significant factor in securing employment within the first six months after graduation. From the findings by [8], the chances of employment are found to be significant for graduates associated with high academic performance and who had internship experience. However, high academic performance by itself does not exhibit any significance on employment chances. In addition, proficiency in English, Malay, and other languages is found to be significant in a study done by [9] in predicting employment status within the first six months after graduation. English language proficiency is identified to influence towards predicting the early detection and identification of the employability of graduates. The study stated that knowing a foreign language plays an important role in job hunting, as it becomes an advantage for graduates as one of their special skills. According to the study done by [10], graduates' race and year of graduation are determined to significantly influence unemployment status and identified as the strongest determinants. The research by revealed that satisfaction towards counselling on career and job interviews is found to be a significant determinant on graduates' employability

ISSN: 2366-1313

III. PROPOSED SYSTEM

This study aims to develop a data mining model of predicting students' employment and analyze factors influencing graduates' job-hunting. Data on graduate employment were collected, CART's decision tree rule set was used to account for the five factors affecting employment, and the study's validity was further refined using a random forest algorithm. The employment forecasting model developed in this paper provides a new approach to guide college employment.

Analysis and Model Performance

Data preparation was done before data analysis was conducted. For analysis, supervised data mining partitioned the data into 50:50 due to the massive dataset. The algorithms that were used in this study were decision tree, logistic regression, and artificial neural network, which were mainly used for classification and prediction purposes. Consolidation was introduced in this process because a major problem was that the categorical inputs appeared to be large. There were three types of consolidations used: dataset without consolidation, dataset with manual consolidation, and dataset with tree consolidation. Variable selection was applied on all types of datasets used. The variable selection models used were a model without variable selection, a model with variable selection using stepwise, a model with variable selection using decision tree, and a model with variable selection using logistic regression stepwise. These were applied to compare and obtain the best models at classifying the prediction of employment status. The steps involved in supervised data mining for predicting employment status was described in figure 1. and 2 show the process of supervised data mining ran using SAS Enterprise Miner

This study aims to find a classification model of student's employment and better forecast whether the graduates can find jobs. Fig. 1 explains the research procedures, including data collection, data pre-processing, model construction and model evaluation

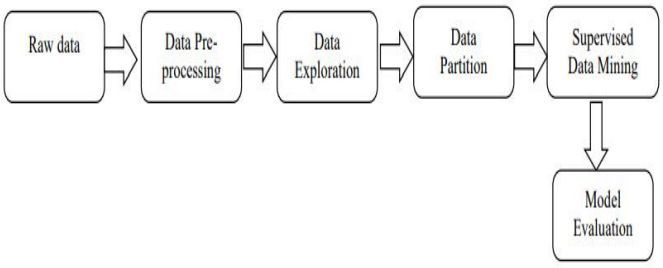

Fig.1 Steps diagram for supervised data mining

ISSN: 2366-1313

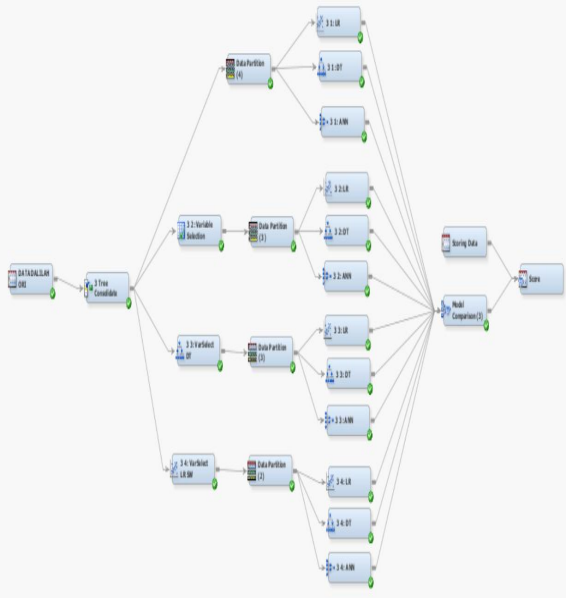

Fig.2 Model with tree consolidation

To choose the best model for model performance (accuracy, misclassification rate, sensitivity, and specificity), model selection was chosen based on parsimony, as a simpler model was preferred and easy to interpret, and model deployment for future use was conducted. As for unsupervised data mining, Decision tree, CART Algorithms was used to find the overall patterns that might reveal hidden factors to give some information. To analyse using Decision tree, CART Algorithm, cluster node is used and then it is applied to segment profile node. To be specific, it was utilized to classify the graduates based on their satisfaction towards curricular activities and ICT skills

Decision tree algorithm

Decision tree is a classical method in machine learning. It is often used to solve classification and prediction problems. The general decision tree consists of one root node, a number of internal and leaf nodes [10]. The internal node of a tree corresponds to a feature and the sample set is divided into the internal nodes according to feature attributes. Leaf nodes indicate the class to be assigned to a sample. The path from root node to leaf node corresponds to classification rules. The decision tree algorithm mainly includes ID3, C4.5 and CART algorithm.

CART Algorithm

The Classification and Regression Trees (CART) uses Gini index as an impurity measure in building the decision tree. Suppose there are K categories, k p is the proportion of the k class samples in the current sample set D (k = 1,2,3,. . . K) , the Gini value of data set D is:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^{\ 2}$$

In (1), the smaller the Gini value is, the lower the impurity is, and the better the feature is. For the sample set D , it is assumed that the discrete attribute a has V possible values $\{a^1, a^2, \cdots, a^V\}$ If it is used to divide the sample set, the splitting Gini index can be calculated as follow

$$Gini(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

ISSN: 2366-1313

IV. RESULTS AND DISCUSSION

The percentage of the graduates' composition comprised 54.2% who are employed and others at 45.8% for IPTA (Institut Pendidikan Tinggi Awam, or Public Higher Learning Institutes) graduates from 2015 to 2018. Figure 5 illustrates the trends in the percentage of graduates according to employment status. The trends indicate that the percentage of employed graduates increased from year 2015 to 2018, i.e., from 9.7%% to the highest peak at 15.4%, respectively; for the percentage for others, there was an increase from 11.1% to 12.1% (2015–2017) but then the percentage dropped by about 1.3% in 2018. By increasing the number of several programs as MySTEP and PROTÉGÉ can reduce the rate of unemployment and at the same time can train fresh graduates to be more marketable.

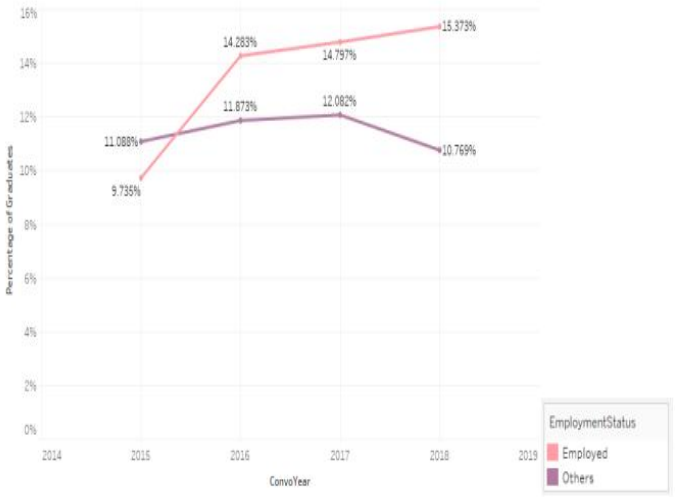

Fig.3 Trends in percentage of graduates according to employment status

Figure 4 shows the percentage of graduates based on working experience as grouped by employment status. Based on the figure shown, both "Employed" and "Others" present a high reading in no working experience and a low percentage in working experience

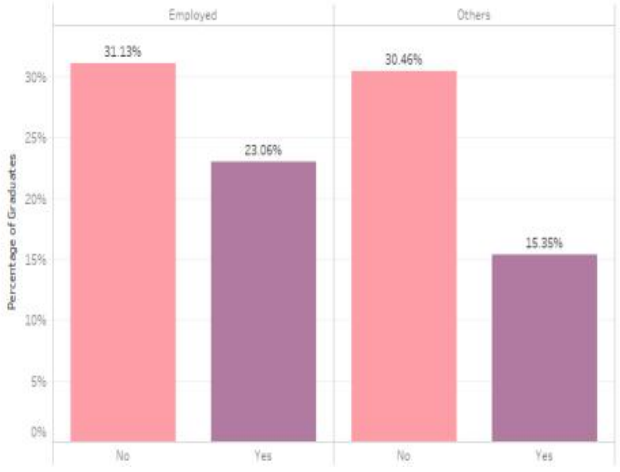

Fig.4 Bar chart of percentage of graduates based on working experience according to employment status

ISSN: 2366-1313

For supervised data mining, to find the attributes that contributed towards employment status, the significant inputs were obtained from logistic regression based on their significant values; for decision tree and artificial neural network, the significant inputs were obtained through variable importance. All these outputs for determining the significant variables were executed without using any variable selection and without consolidating the inputs' categories. Logistic regression result shows that age, course, CGPA, year of convocation, English (SPM), third language, experience with counselling, facilitators, facilities, funding, gender, ICT skills, library, OKU status, race, study mode, system, and working experience were contributors to employment status since the p-values for each input were lower than the alpha value of 0.05. By referring to the variables' importance provided by decision tree, course, prerequisite, working experience, year of convocation, age, race, study mode, funding, marital status, and gender were found to be important towards employment status. As for artificial neural network, significant inputs identified were course, working experience, year of convocation, prerequisite, third language, age, race, marital status, facilities, English (SPM), CGPA, and funding. The common inputs that were found to be significant in all two models were course, working experience, year of convocation, age, and race, which highly contributed towards employment status

V. CONCLUSION

As the situation of undergraduate employment becomes more and more severe, employment instruction is an important part of students' management in colleges and universities. How to apply data mining algorithm to forecast students' employment is an urgent task. Based on the CART decision tree algorithm, this paper establishes an employment prediction model and analyses the influencing factors of employment. Then we further adopt random forest algorithm to improve the learning accuracy. The experimental results show that the combination of decision tree and random forest algorithm can effectively predict students' employment situation. The model proposed provides a new method for employment prediction and has practical application value.

REFERENCES

[1] C.J. Yue, J. Xia, W.Q. Qiu, "An empirical study on graduates' employment: Based on 2019 national survey," Journal of East ChinaNormal University(Educational Sciences), no. 4, pp. 1–17, 2020.

ISSN: 2366-1313

[2] M. T. R. and Y. Yusof, “Application of data mining in forecastinggraduates employment,” Journal of Engineering and Applied Sciences, vol. 12, pp. 4202-4207, 2017.

[3] K. C. Piad, M. Dumlao, M. A. Ballera and S. C. Ambat, "Predicting ITemployability using data mining techniques," 2016 Third International Conference on Digital Information Processing, Data Mining, andWireless Communications (DIPDMWC), Moscow, Russia, 2016, pp. 26-30.

[4] Z. Liu and Z.G. Zhao, “Analysis and calculation of high school graduatestudent based on data mining,” Journal of Shenyang Normal University(Natural Science Edition), vol. 34, pp. 105–108, January 2016.

[5] L.Y. Zhang, F.C. Wang and Z.Y. Han, “Analysis of the influencingfactors of university graduates employment based on decision treealgorithm--A case study of information college of Beijing ForestryUniversity,” Forestry Education in China, vol. 35, pp. 46 – 51, March2017.

[6] Y. Tang and P. Wang, “Study on employment forecasting of graduatesof traditional Chinese medicine based on C4.5 and random forest algorithm,” China Medical Herald, vol. 14, pp. 166-169, August 2017

[7] L. Cai and X. Wang, “Prediction and influencing factors of college students’ career planning based on big data mining,” Mathematical Problems in Engineering, vol. 2022, Article ID 5205371, 11 pages, 2022

[8] A. Bai and S. Hira, “An intelligent hybrid deep belief network model for predicting students’ employability,” Soft Computing, vol. 25, no. 14, pp. 9241–9254, 2021

[9] B. Yang, “Internship effect prediction for physical education majors based on artificial neural network,” International Journal of Emerging Technologies in Learning, vol. 16, no. 24, 2021

[10] F. Yang, “Decision tree algorithm-based university graduate employment trend prediction,” Informatica, vol. 43, no 4, pp. 573-579, 2019.

[11] Prasadu Peddi (2022), A Hybrid-Method Neighbor-Node DetectionArchitecture for Wireless Sensor Networks, ADVANCED INFORMATION TECHNOLOGY JOURNAL ISSN 1879-8136, volume XV, issue II.