JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 36, 193-203 (2020)
DOI: 10.6688/JISE.202003_36(2).0001

Employing Data Mining to Predict Professional Identity

RAYA MOHAMMED MAHMOOD AND SEFER KURNAZ
Faculty of Engineering and Natural Science
Altinbas University
Istanbul, 34676 Turkey
E-mail: raya.mahmood@ogr.altinbas.edu.tr; sefer.kurnaz@altinbas.edu.tr

Data mining in educational filed becomes more involved and adding value to the educational research. In this study, the development of the professional identity of graduate students in a graduate school of science and engineering (GSSE) / Altinbas University were measured by professional identity-five factor scale (PIFFS). The results of the survey were analyzed statically using SPSS, as a result, four different levels of professional identity were recognized. The collected data analyzed by machine learning algorithms which were deployed using python code to predict student's professional identity levels based on previously achieved results. Various types of algorithms were used and compared in a matter of accuracy and running time to select the most fitted model with the most accurate results.

Keywords: data mining, machine learning, prediction, professional identity, survey analysis.

1. INTRODUCTION

Data mining increasingly becoming a crucial aspect in the data science field where most applications rely on. The availability of widespread platforms that uses data mining algorithms and its flexibility in dealing with various kinds of extensions has attracted most researchers to pursue this track in analyzing their information. However, when it comes to psycho-social sciences and survey analysis, most likely they are using traditional statistics in analyzing and determining trends as well as impact factors. Professional identity development (PID) is one of the psycho-social subjects that dealing with development nature of human behavior and how its impact can be perceived in performing competently and accredited in professional context [1]. Measuring PID for graduate students will reveal learning outcomes and provide an assessment to the learning process as well as examining the eligibility of graduate students to penetrate the employment field. The significance of predicting PID lies in its effect on the evaluation of educational systems as a unit containing students, teachers, and the learning process take into consideration the huge amount of available educational constitution and the need to an efficient way to handle such amount of data accurately.

From a data mining perspective, predicting targeted value like professional identity required less effort than calculating the same value using conventional descriptive statistics. Although inferential statistics like correlation, linear regression is in the vicinity of data mining, there is distinct variance between them when it comes to accurate measurements and performance, even though accuracy can vary in the same data mining algorithm depending on using appropriate values for model factors.

Received September 3, 2019; revised September 11 & 26, 2019; accepted October 11, 2019.
Communicated by Osamah Ibrahim Khalaf.

Therefore, this study will discuss the effect of using various machine learning algorithms on the accuracy of predicted PID values. The collected data stored as CSV files and accessed independently by SPSS for statistical analysis and by python code for data mining analysis where the overall analyzing time and accuracy registered and compared for each algorithm.

2. RELATED WORK

The applied PIFFS survey based on the study presented by [1]. The authors suggested a new scale to measure student's professional identity for students who enrolled in polytechnic in Singapore, however, the scale validated for the intended purposes. As a result, students were separated into two groups depending on their professional identity to high and low professional identity. Whereas this study will not reinvent the wheel by validating the scale again, rather it will focus on the validation of the used data according to the scale, especially correlation and reliability furthermore it will make use of data mining algorithms in predicting student's professional identity as multicategories data [2].

Researchers endeavor to appraise the impact of PID in higher education with various kinds of professions. The literature on professional identity has highlighted the effect of action learning in emphasis on the professional identity potentiality of undergraduate students by assigning them professional rules "consultant" to other students' "clients". In essence, An appositive correlation found between action learning and the development of professional capabilities [3].

There is a unanimity amongst social scientists that students develop their identity during classes, practices as well as engagement with colleagues therefore classes may convey to upholding student's professional identity growth such as mentioned by [4] where they use reflective writing as stimulus in the evaluation and analysis of student's response to assist their evolution in design thinking. Correspondingly, [5] uses the same convergence by deploying reflective activities to boost student's identity development as shown in the case of 66 English students who experience essay reflective writing as an identity expansion tool. The findings of that treatise prove the role of reflective activities in developing a student's identity to prepare them to be highly skilled English teachers.

The existing literature on professional identity concentrates on the effect of relationships between colleagues inside and out of the educational assembly upon the evolution of student's professional identity specially in higher education [6], Alternatively, some researches focus on the mutual learning in boosting student's professional identity by engaging, communicating with other colleague and observing higher stages students. The gained confidence from learning applied and authenticated procedures in dealing with patients helped pharmacist students in developing their professional identity [7]. These conclusions are similar to those narrated by [8] where the research detected that the professional identity of mathematic students affected by their previous experience that acquired during formal education and the potential effect on their personality and capability in dealing with numerous kinds of students efficiently. Whereas other studies stated the difference between exploratory attitude and formal learning in developing professional identity, also demonstrating the effect of participating inadequate programs that support personality transition to a higher level of understanding and communication intelligence in a certain profession [9].

Some researcher investigates the impact of corporate or organization mentality on the academics in higher education which leads to cleavage in this connection and how this dispute can be solved by bridging the gap between management and academics [10]. On the other hand, counselors in training seemed to examine extra factor that affects professional identity development which is diligence. Accumulative experience, proper coaching, empirical knowledge, participation demeanor along diligence will lead to a huge transformation on a counselor's personality to gain professional identity [11, 12].

From a scientific view, recent attention paid to utilize data mining techniques onward with applied machine learning to expose trends and human behavior in social studies to get deep knowledge on a certain phenomenon or predict a specific behavior [13, 14]. Data mining described by [15] as the science that uses reasoning analysis methods using various kinds of tools and algorithms to detect patterns and predict certain knowledge based on those patterns and relations.

Using data mining to support education (educational data mining EDM) has been widely spread for decades. Many researchers use data mining to predict students' performance student performance by combining multiple data mining techniques such as bridging between classification and clustering [16]. One of the most common ways to collect student's information coming from student's database which makes it easy to apply classification, clustering and outlier detection for a specific behavior, from this point it will be easy to use data mining tool such as WEKA to demonstrate analysis results and obtain the immersed knowledge behind the scene [17]. From this perspective students' intellectual pattern can be predicted using various learning platforms such as e-learning gates used by educational institution (Moodle as an example) [18], using this method will facilitate the optimization of student performance as well as the degree of learning motivation acquired using online learning technology [19, 20].

On the other hand, [21] suggested that with the right questionnaire, we can satisfy both students' and faculty's needs based on the analysis and prediction of their requirements dragged from their answers, which leads to enhancement in student's behavior and performance as well as an institutional value.

The process of educational data mining as illustrated by [22] start firstly by collecting data either by questioner, student's database, repository along with calculating its reliability, on the next step data will be examined statistically to extract important yet general information about the paradigm of the study, then extra knowledge can be derived using prediction to highlight the effective traits. Depending on the type of data a decision made to use either classification or regression or even clustering if it assists the purpose of the study. Finally, it's common to use different kinds of data mining algorithms to scrutinize the precession of the prediction process and adopt the most accurate as an authenticated result.

Many data mining applications presented recently in social science and especially in education such as using data mining and natural processing language in producing a guidance system for students to determine the best institution which reflects their desires and needs [23], another application was presented by [24] which suggest studying the effect of student punctuality on learning process using multiple data mining algorithms such as NB, DT and NN. Another application demonstrated by [25] to analyze postgraduate students' earnings and deploy different kinds of machine learning algorithms in selecting the most effective traits to predict precisely the amount of the income.

3. ANALYSIS SCHEME

This study will measure the professional identity of Altinbaş University students and anatomize the results using data mining techniques for portending students PID. The aim of this study can be achieved by analyzing the collected data in two ways, first statically using descriptive and inferential statistic then using multiple machine learning algorithms for predicting PID without further statistics then compare the results for more accuracy and ideality. The intended procedure used to emphasis the difference in simplicity and accuracy between both methods. The following steps clarify the analysis process.

3.1 Data Collection, Cleaning, Coding

PIFFS were submitted to both master and doctoral students in GSSE/Altinbaş University, some students fill the paper survey in class while others prefer electronic version presented online using google forms. The survey contains twenty-five questions distributed in five groups [1]. deliberately, the average of each group calculated to get the overall average. Extra personal student information combined with the scale and average fields to perform a single excel file which contains forty attributes for four-hundred tuple (student). Since data were collected in various ways then some inputs were mistakenly entered, spelled or left vacant. These mistakes resolved by the following steps:

1. Unreliable numeric input such as age (entering 2 instead of 25) or vacant answer, handled by calculating the average of ages which was 30 and replace unreliable values by this number.
2. Different spelling of country names, the native language was handled by uniform writing of these names in this stage.
3. Vacant choices in survey questions entry were handled by taking the average answers and fill it in the place.

The final step at this stage was coding the entries thus each choice in the survey has numeric code to facilitate calculation and analysis in the upcoming stages.

3.2 Statistical Analysis

This phase accomplished using SPSS software. At this phase, the type of data will be the key factor in analyzing survey entries, therefore, data analyzed on three different groups according to data types as below.

(a) Categorial data, such as gender, nationality, program type, stage, native language and studying language analyzed by counting frequencies of occurrence and cumulative percentage to obtain a general view of the collected information.
(b) Quantitively data, such as age is described also using frequencies and descriptive statistics as (mean, median, mode, *etc.*).
(c) Survey statistics (correlation, reliability, validity), At this stage the answers to the five-factor scale items analyzed using statistical tests to obtain the significance of the results, the reliability of the collected information and the validity.

Firstly, the PIFFS analyzed by descriptive statistics, each question has proposed an answer as a scale from one to five (1. Never True, 2. Not Really True, 3. Neutral Somewhat, 4. True, 5. Definitely True). The answers in each group averaged for each student and the results of five groups averaged consequently for each student to get a single value to represent student response for the overall survey. To get a professional identity attribute which is a single number that indicates the level of the proficiency of each student, the overall scale means calculated to be 3.56 then the student divided based on their average answers into quartiles around the mean to get multi categorical attribute. Thus, instead of telling that student with high or low professional identity, the professional identity represented a scale from 1 to 4 were (1. Low PI, 2. Fair PI, 3. Moderate PI, 4. High PI), the first quartile < 3.28, second quartile between 3.28 and 3.6, third quartile between 3.6 and 3.92, fourth quartile > 3.92.

Secondly, correlation analysis carried out between main survey fields to indicate the type of the relation among survey variables however this step clarified which subset of questions is more likely related, as a result, the more correlation is close to 1 the more related variable exists. Normally this step is important in survey design and analysis however it included in this study for comparison purposes.

The third measurement includes building the reliability of the model under collected data within Cronbach's Alpha Based measure equal to 0.7, this means that the survey is consistent and will generate the same result if it is applied on different sample, then the validity of the survey result created by taking the square root of reliability of the scale where the increased alpha value indicates high validity. Finally, the regression analysis applied to measure the leverage of independent variables on the PID attribute as the dependent variable and check model fitting.

3.3 Feature Selection

Since this study aimed to use a machine learning algorithm in predicting professional identity, therefore, feature selection applied using WEKA to get benefit from embodied feature selection algorithms also to compare the results to normal correlation achieved using inferential statistics. It's worth to mention that at this stage, the average fields for every group in the PIFFS was omitted as its added for statistical purposes and to emphasis the fact that using machine learning algorithms will simplify gaining final results without extra calculation.

3.4 Prediction Using ML Algorithms

Since this study uses labeled data, classifiers were used to find relations between the provided data which lead to more accurate predictions. Then, this knowledge is used to anticipate new tuples. Furthermore, it is essential to evaluate the execution of the classifiers, by finding how accurate their predictions are. The usual procedure used to measure the performance of classifiers is by splitting labeled data into two sets, the first set is used for the training set and the second set is used for evaluating the classifier performance. The procedure that is used to find the performance of classifiers is shown in Fig. 1.

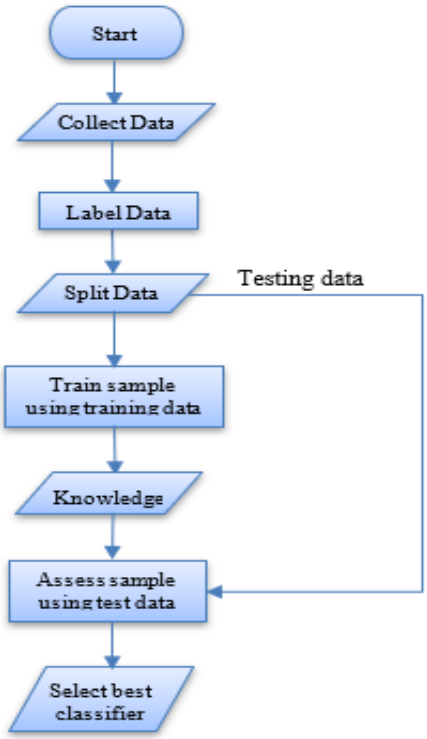


Fig. 1. Classifiers performance evaluation.

By using a scikit-learn library in Python, which delivers rigid implementations of a variety of machine learning algorithms, four classifiers were used for evaluating each model and select the best classifier depending on the accuracy. These classifiers are:

(a) Decision tree, this classifier computes the information gain for each feature that plays a role in the classification procedure. Information gain means the ratio of the amount of information that can be recovered from a specific feature to the overall information that may be recovered from the whole dataset.
(b) K-nearest neighbor, the main idea of this classifier is based on choosing k tuples of the training set, a dataset which is the most similar to the new tuple being classified, a predicted class is accomplished by using the dominant class among these tuples. Measuring a similarity is done by finding Euclidean distance between every tuple in the dataset and the new tuple.
(c) Naïve Bayes Classifier, the mechanism relies on building a classifier from a probabilistic model using Bayes theorem. Naïve Bayes consider that the value of a special feature is independent of the value of other feature, given the class variable.
(d) Support Vector Machine (SVM), SVM defined as a discriminative classifier by the notion of a separating hyperplane which helps in the classification of new data points. It can achieve a non-linear classification in additional to linear classification by using a kernel trick technique.

4. RESULTS

4.1 Statistical Analysis Results

The statistical analysis shows that most students are male and comes from Arabian countries. however, the variance of student's nationality which contains twelve different nationalities with a total of seven different languages emphases the flexibility of studying curriculum in connecting these variables by providing the English curriculum as unified language. The preferable program was ECE as it is an engineering program and most students attending in MSc stage. The vast majority of students were between 23 and 30 which reflects the professionality of young students and its effect on the learning process.

From observed results of PIFFS analysis, the answers to the first factor questions almost follow the normal distribution and the highly frequent answers reflect student's high knowledge in practicing professionally. The second group analysis showed that most students have a high experience in the profession. Most students look positively in having a role model in the profession as described in the third group, students look to the fact that having a role model could energize them to be more professional. Normal distribution reflected in the answers of students on the fourth group, where high amount of student feels neutral about their self-efficacy. In the fifth group it can be seen that most students have their preference for the profession and know exactly the type of profession they will go for.

In correlation analysis results shows that the first three factors are within a significant level at 0.05 which is assigned by [1]. Although the fourth and fifth factors are not highly correlated according to the results. moreover, a high correlation between first and third factor is observed, in contrast between first and (second, fourth, fifth) lower correlation registered. the results highlighted that the first factor is the effective factor in determining a student's professional identity. The survey results found reliable according to Cronbach's Alpha indicator with 0.721 and valid with 0.849 while the increased alpha emphasis survey validity. The model fitted according to regression analysis with a reasonable fit of 0.79 with an error of 0.62 at 0.05 confidence interval.

In compared to correlation, feature selection results carried out by WEKA was done by seeking various algorithms, filter method correlation by ranking using cross-validation was the algorithm that produces the most authenticated result depending on the rank of each item in the scale. The highly ranked features used in the prediction process, mentioning that the only effective feature in personal information was age, in contrast, the fifth factor found ineffective according to its low rank.

4.2 Prediction Results

To apply prediction classifiers, the whole dataset split up to a training set which takes 80% from dataset and testing set which takes the remaining 20%. A confusion matrix is used to estimate each classifier's performance.

In KNN it is important to use a grid search for obtaining the optimum number of nearest neighbors which giving the highest accuracy. As indicated in Fig. 2, the relationship between testing accuracy and the number of neighbors is clarified using grid search. it is obvious that the highest accuracy (0.8375) occurred when the number of K is 9 and 10.

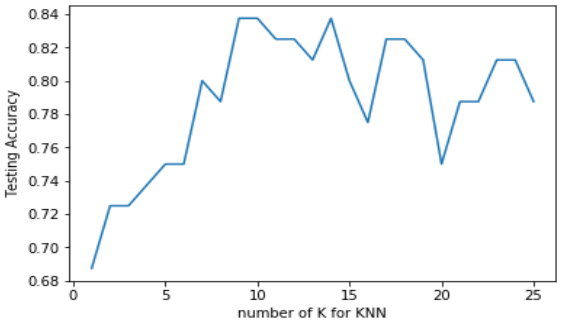


Fig. 2. Grid search for KNN.

In SVM, the RBF kernel is implemented as a non-linear SVM to map out the actual dataset into a higher dimensional extent for making it a linear dataset. Table 1 presents the essence of classification outcomes.

Table 1. Classification results summary.

	Execution time	Accuracy
Decision Tree	0.230767 second	1
KNN	0.312469 second	0.8375
Naïve Bayes	0.281224 second	0.76
SVM	1.187218 second	0.8875

The different accuracies of tested classifiers indicate the capability of the decision tree of deciding on divergent features in a shorter time and a very good execution by selecting the attribute with the peak information gain as a filter to be node or leaf relying on the results of classifying the remaining tuples.

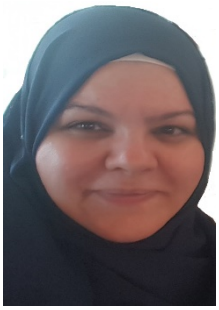
5. CONCLUSIONS

To sum up, using data mining in social and educational topics can be beneficial in focus on certain issues also in facilitating result gaining in comparison to standard statistical methods. Moreover, detecting influential traits using a data mining framework can idealize prediction in later stages. The type of the collected data determines the analysis path, the feature selection algorithm, the data mining method (classification in this case) and the algorithm used in prediction. Most educational data mining surveys use DT where the algorithm generates high accuracy with low running time in most cases as in this survey. However, it's worth mentioning that full accuracy may indicate overfitting in DT learning algorithm according to the limited number of tuples that can be handled in future work using Neural Network. Although, it's important in each case to use various kinds of machine learning algorithms to get authenticated output. It can be seen that PID predicted easily and accurately based on prior knowledge and can be used for huge number of unseen students in future by just filling the survey and running the DT algorithm, moreover student's development can be tracked by observing their transition in four-stages scale instead of assigning them to high or low PI.

REFERENCES

1. C. Tan, H. T. van der Molen, and H. G. Schmidt, "A measure of professional identity development for professional education," *Studies in Higher Education*, Vol. 42, 2017, pp. 1504-1519.
2. S. Kurnaz and R. Mahmood, "Methodology preview on predicting students professional identity using data mining techniques," in *Proceedings of the 4th International Conference on Engineering & MIS*, 2018, Article 56, p. 6.
3. A. Lizzio and K. Wilson, "Action learning in higher education: an investigation of its potential to develop professional capability," *Studies in Higher Education*, Vol. 29, 2004, pp. 469-488.
4. M. Tracey and A. Hutchinson, "Reflection and professional identity development in design education," *International Journal of Technology and Design Education*, Vol. 28, 2018, pp. 263-285.
5. I. Ivanova and I. Skara-minc, "Development of professional identity during teacher's practice," *Procedia – Social and Behavioral Sciences*, 2016, pp. 529-536.
6. V. Sweitzer, "Towards a theory of doctoral student professional identity development: A developmental networks approach," *The Journal of Higher Education*, Vol. 80, 2009, pp. 1-34.
7. S. J. Bridges and B. Hons, "Research in social and administrative pharmacy professional identity development: Learning and journeying together," *Research in Social and Administrative Pharmacy*, Vol. 14, 2018, pp. 290-294.
8. G. Gowri, R. Thulasiram, and M. A. Baburao, "Educational data mining application for estimating students performance in Weka environment," *IOP Conference Series: Materials Science and Engineering*, Vol. 263, 2017, pp. 1-9.
9. K. Nesje, E. Canrinus, and J. Strype, "Trying on teaching for fit, development of professional identity among professionals with multiple career opportunities," *Teaching and Teacher Education*, Vol. 69, 2018, pp. 131-141.
10. R. Winter, "Academic manager or managed academic? Academic identity schisms in higher education," *Journal of Higher Education Policy and Management*, Vol. 31, 2009, pp. 121-131.
11. H. Woo, J. Lu, C. Harris, and B. Cauley, "Professional identity development in counseling professionals," *Counseling Outcome Research and Evaluation*, Vol. 8, 2017, pp. 15-30.
12. S. Dong, A. Campbell, and S. Vance," Examining the facilitating role of mindfulness on professional identity development among counselors-in-training: A qualitative approach," *The Professional Counselor*, Vol. 7, 2017, pp. 305-317.
13. A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, Vol. 5, 2017, pp. 15991-16005.
14. I. Shingari and D. Kumar, "A survey on various aspects of education data mining in predicting student performance," *Journal of Applied Science and Computations*, Vol. 5, 2018, pp. 38-42.
15. K. Kushwaha and P. Mishra, "A survey on data mining using machine learning techniques," *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, 2016, pp. 177-180.
16. B. K. Francis and S. Babu, "Predicting academic performance of students using a

hybrid data mining approach," *Journal of Medical Systems*, Vol. 43, 2019, pp. 1-15.
17. S. Aher and L. M. R. J. Lobo, "Data mining in educational system using WEKA," in *Proceedings of International Conference on Emerging Technology Trends*, Vol. 3, 2011, pp. 20-25.
18. S. Huang, "Analysis of students' learning behavior based on association rule mining algorithm in moodle network platform," in *Proceedings of the 5th International Conference on Electrical and Electronics Engineering and Computer Science*, 2018, pp. 472-475.
19. M. Jovanovic, M. Vukicevic, M. Milovanovic, and M. Minovic, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study," *International Journal of Computational Intelligence Systems*, Vol. 5, 2012, pp. 597-610.
20. K. Kularbphettong and C. Tongsiri, "Mining educational data to analyze the student motivation behavior," *International Journal of Information and Communication Engineering*, Vol. 6, 2012, pp. 1032-1036.
21. A. Bajaj, "Use of data mining techniques in assessing student and faculty needs," *International Journal of Innovative Research and Development*, Vol. 6, 2017, pp. 199-202.
22. A. Manjarres, L. Gabriel, and M. Sandoval, "Data mining techniques applied in educational environments: Literature review," *Digital Education Review*, No. 33, 2018, pp. 235-266.
23. A. Panchal and R. Nair, "College recommendation system using data mining and natural language processing," *International Journal of Engineering Science and Computing*, Vol. 8, 2018, pp. 18863-18866.
24. I. Siddiqui, Q. Arain, and S. Bhutto, "Analyzing students' academic performance through educational data mining," *3C Tecnología. Glosas de innovación aplicadas a la pyme*, 2019, pp. 402-442.
25. E. Wright, Q. Hao, K. Rasheed, and Y. Liu, "Feature selection of post-graduation income of college students in the United States," in *Proceedings of International Conference on Social Computing*, *Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 2018, pp. 38-45.



Raya Mohammed Mahmood was born in 1981 in Baghdad, Iraq. She received the B.S. degree in Computer and Software Engineering from Al-Mustansiriya University, Iraq and M.S. degree in Computer Science / Computer Networks from Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics, Iraq. She is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, Altinbas University, Turkey. Her research interests include applied machine learning and datamining in educational field.

Sefer Kurnaz was born in April 19, 1956 in Giresun, Turkey. He has a retired colonel from Turkish Air force Command. In the past, he was in charging of as Director of Aeronautics and Space Technologies Institute (ASTIN) in Turkish Air force Academy, Istanbul. He has mainly specialized in space technology application. He has graduated Ph.D. at the İstanbul University and specialized in computer science. He was worked at 6 Allied Air Tactical Force (6ATAF) in NATO, Izmir as a system support section chief. He is a founder and before retiring was a general chair Resent Advances in Space Technologies (RAST) International Conference, Turkey. He is an author of two books in Turkey and United States of America. Under his leadership have published over than 20 scientific papers.