Personality Forecast of Bachelor Degree Student Using Decision Tree Technique

Busarin Eamthanakul[1],
Montean Rattanasiriwongwut[2],
Faculty of Information Technology,
King Mongkut's University of Technology North Bangkok, Thailand
[1]busarin.ea@gmail.com
[2]montean@it.kmutnb.ac.th

and Monchai Tiantong[3]
Faculty of Technical Education,
King Mongkut's University of Technology North Bangkok, Thailand
[3]monchai@kmutnb.ac.th

***Abstract* - An objective of this research is to forecast a student personality using a decision tree technique by C4.5 algorithm. Also, it measures for an efficiency of a developed model. This case study collects a data from a personality questionnaire. A sample group who do this questionnaire is a group of 1,873 undergraduate students in Suan Sunandha Rajabhat University. A model of personality forecast is created and tested by a K-Fold Cross Validation method, a Percentage Slip method, and a Training Set and Test Set method. As a result of this research, it describes that a learning data set and a testing separation method can be used for a student personality forecast model development. As well, this model development uses a decision tree technique that has a high accuracy for a student personality forecast.**

***Keywords* - Big Five Personality, Classification Rules, Decision Tree, Forecast**

I. INTRODUCTION

An education is a main factor to develop oneself and a country increasingly. There are many factors to development a student to be a person who is good at many things. For example, academy, moral, ethics, wisdom, emotion, social ability, etc. A personality development is one of those factors. There is a research report that a personality effects to abilities of problem encountering, obstacle coping, and boring in work [1]. Moreover, a personality is a factor of a working career success. Since a good personality makes an attractiveness to that person also an influence to persuade the other. It is a real power of success [2].

There are many theories for a personality study to know what kind of personality each person should be. But the most popular and standard theory is The Big Five Personality by Costa and McCrae. There are five kinds of personality in this theory as Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. In order to forecast a student personality, the researcher uses a data mining technique to classify a data and forecast a student personality after that. It can be used in a university for an educational planning, a teaching preparation, and a consultation of education suitable for each student in the future. Furthermore, each student can improve themselves for their personality later.

II. LITERATURE REVIEWS

A. Decision Tree 4.5 Algorithm

A decision tree [3] is a technique that gives a result in a form of a tree structure. If a user has a data to arrange into group, a user will

compare a data attribute with a tree route until a destination class having the same data group. Inside a tree, it consists of a node, a branch, and a leaf. Each node has an attribute for testing. A branch is a possible value of a testing attribute. A leaf is a lowest part in a decision tree. It refers to a group of class or a forecast result from a root node which is in a highest part of a decision tree as shown in Fig. 1.

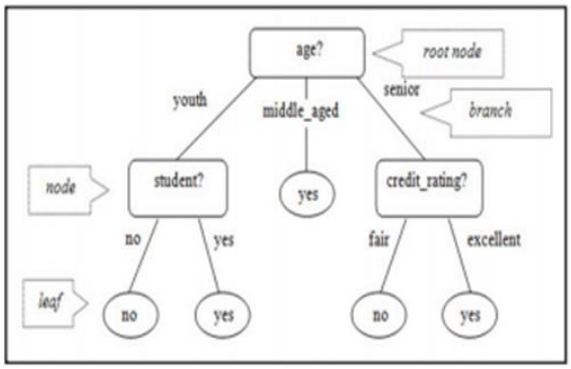

Fig. 1 A Decision Tree Diagram [4]

Algorithm C4.5 is developed from ID3 by Quinlan [5]. This algorithm solves a problem in a case of scattered data in a data set or not grouping yet. Then, a classification is inclined. So, it increases split information that calculates as in Equation 1.

$$\text{SplitInfo}_A(d) = -\sum_{i=1}^{n} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \quad (1)$$

After split information implementation, it refers to a data scattering attribute. After that, it uses split information values to divide with a standard gain value for solving an inclination problem. And then, it comes out with a standard gain ratio as shown in Equation 2.

$$GainRatio\ (A) = \frac{Entropy(D) - Entropy_A(D)}{SplitInfo_A(D)} \quad (2)$$

After getting a gain ratio, it chooses a highest value as a first node to implement also creates later nodes. It uses a remained attribute until receives a completed tree that has the least scattering data or the most grouping.

B. The Big Five Personality

The Big Five Personality means a human behavior that expresses in each element by a concept of Costa and McCrae [6-7]. It consists of five kinds of personality. First is Neuroticism. Neuroticism is a person whose emotion is so various, sad, angry easily, and worry. In addition, he or she cannot control for a stimulation or desire of his or her emotion. Second is Extraversion. Extraversion is a person who likes to interact with other person. He or she enjoys with an activity, frankly, and optimistic. Third is Openness to Experience. Openness to Experience is a person who is interested in a new thing, accepts for a new thinking, finds for a new experience, fanciful, and revelation. Forth is Agreeableness. Agreeableness is a person who is open-minded, friendly, generous, helpful, and tender-minded. Fifth is Conscientiousness. Conscientiousness is a person who has a clear target, like to plan in advance, can manage for his or her life, high responsibility, respects for an ethics.

III. RELATED RESEARCH

Payoon [8] develops a data mining system using a decision tree technique. Also, a tree structure will express in a structure having different rules by a using target or objective. After that, it classifies a group of data as specific before. As a result, the system can classify in a good level. As well, the developed system can be used with other kinds of data.

Wasana [9] develops the system of psychological research paper searching. The system uses a decision tree technique for a research paper arrangement. As a result, it can separate a research paper in each classification more efficient. Also, it comes out with a condition for searching a research paper more effective.

Natthaphol and et al. [10] study for a mental problem of students in Phramongkutklao College of Medicine before and after taking a military course. Moreover, they study for a relation between a personality and other factors that are effective in a metal problem.

Aranya [11] studies for a relation between a personality, an atmosphere in organization, a relationship with an organization in each

employee, and a good member behavior of teachers in a vocational school of Nakhon Si Thammarat Province.

Monchai and Chusri [12] develop a readymade software for a personality testing using a Sixteen Personality Factor Test by Raymond B Cattel in Thai edition. This software is a prototype to develop a system for a personality testing for user who is sixteen years old or over.

Gokul [13] investigates for a relation between result data after a data mining implementation in a mobile phone. Moreover, he investigates for a personality attribute of a mobile phone owner by a Big-Five Personality Traits.

Johann [14] checks for a relation between a personality and Five-Factor Model. Furthermore, it compares with an Internet using style also data openness behaviour in a social network.

Yoram [15] studies for a relation between a Facebook using behaviour and that Facebook owner according to the Big Five Personality Model. In addition, Data in this research are personality data and information inside his or her Facebook.

IV. METHODOLOGY

There are three processes in this research as a data preparation, a forecast model creation and testing, and a forecast model efficiency measurement.

A. Data Preparation

The researcher collects a data using a personality questionnaire for 1,873 undergraduate student of Suan Sunandha Rajabhat University in Year 2013-2015. The factors that are used for a data analysis are gender, age, level, average grade, faculty, department, religion, sibling, birth sequence, hometown, bringing up style, father education, mother education, father occupation, mother occupation, average family income, and pocket money from parents. After that, data from student questionnaires are transformed to *.CSV file in order to prepare for a model creation and testing by WEKA Programming.

B. Forecast Model Creation and Testing

To create that forecast model, it uses a WEKA Programming to help for a model creation and testing by a decision tree technique in C4.5 algorithm. A received prototype will be in a form of data classification rules that learning by a training set or a prototype creation data set. Next, it is tested by a test set using a K-Fold Cross-Validation method, a Percentage Split method, and a Training and Test Set method.

The first method is a K-Fold Cross-Validation method. This method separates all data of N set in k fold. Each part has N size divided by k. After that, the first part of data is tested with the other parts until finished. And then, it changes to use the second part of data for testing instead until completed in k parts. For this method, the researchers specifies for k values as 5, 10, and 20 in sequence. So, it gets three forecast models.

The second method is a Percentage Split method. This method separates all data in a learning set and a testing set by a sampling method. A data size is specified as 20 in percentage, 66 in percentage, and 80 in percentage for getting three forecast models.

The third method is a Training and Test Set method. The researcher specifies for a ratio between two data set as 80 in percentage and 20 in percentage from all 1,873 data set. As a result, it gets 1,498 data set of learning and 375 data set of testing in sequence.

C. Model Efficiency Measurement

The researcher measures an efficiency of a model from a model forecast result of a confusion matrix. It implements for finding four values as accuracy value, precision value, recall value, and f-measure value. As well, it calculates from data in Table I with equations from 3 to 6 sequentially as shown below [16].

TABLE I
A CONFUSION MATRIX

		(Predicted Class)	
		Class = Positive (+)	Class = Negative (-)
(Actual Class)	Class = Positive (+)	True positive (TP)	False negative (FN)
	Class = Negative (-)	False positive (FP)	True negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$F - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

V. RESEARCH RESULTS

Here are the results from a student personality forecast. This research uses a decision tree technique by a K-Fold Cross-Validation method, a Percentage Split method, and a Training and Test Set method. There are 1,873 data set that express 4 efficiency values as accuracy value, precision value, recall value, and f-measure value as shown in Table II.

TABLE II
EFFICIENCY VALUES FROM MODEL TESTING BY DIFFERENT METHODS

Forecast Model Method	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
5-Fold Cross Validation	78.06	74.70	78.10	76.00
10-Fold Cross Validation	78.06	75.00	78.10	76.10
20-Fold Cross Validation	77.68	74.90	77.70	75.90
20 in Percentage Split	64.09	61.00	64.10	61.40
66 in Percentage Split	77.39	74.60	77.40	74.90
80 in Percentage Split	78.40	76.30	78.40	75.50
Training and Test 80:20	88.80	86.30	88.80	87.00

From Table II, the results of this research explains that a student personality forecast model developed by a training and test set method has a higher efficiency to classify than by other methods. An accuracy value is 88.80 in percentage, a precision value is 86.30 in percentage, a recall value is 88.80 in percentage, and an f-measure value is 87.00 in percentage. Then, a training and test set method can be used to develop a student personality forecast model by a decision tree technique effectively.

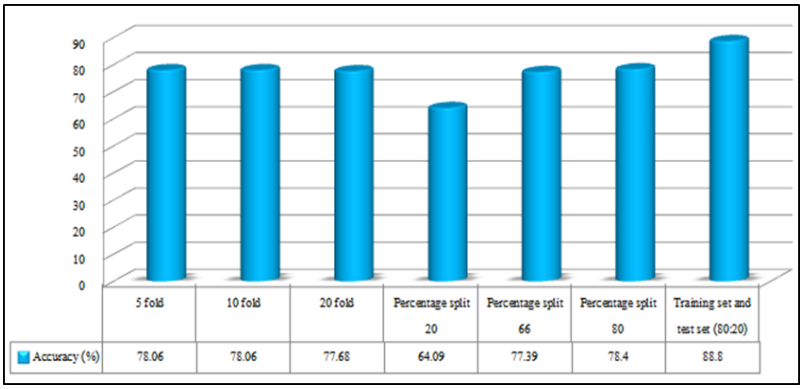

Fig. 2 Comparison Graph between Accuracy Values of Personality Classification by Decision Tree Technique

From Fig. 2, a forecast model that is developed by a training and testing set method has a highest accuracy value at 88.80% in percentage. It refers that a model that has a high accuracy value will classify data better than a model that has a low accuracy value.

VI. CONCLUSIONS

An undergraduate student personality forecast model uses a decision tree technique in C4.5 algorithm. It is separated in three methods for a forecast model development. These are a K-Fold Cross-Validation method, a Percentage Split method, and a Training and Test Set method. According to the efficiency and accuracy measurement of a developed model, a training and test set method has an accuracy value at 88.80 in percentage, a precision value at 86.30 in percentage, a recall value at 88.80 in percentage, and an f-measure value at 87.00 in percentage. Moreover, all efficiency values are higher than other methods. Furthermore, there are 224 rules of data classifications. In addition, a training and test set method can be used to forecast a student personality in Suan Sunandha Rajabhat University that is more effective than other methods.

If bringing 224 classification rules from a forecast model to create an application for forecasting a student personality, a user will get information about a student personality by a test of student data with these 224 rules. In addition, this research can be useful for student, teacher, administrator, and student consultant in a university. They will use for an education plan, a study plan, an education consult according to a student personality in an overview idea. And then, an educational management in a university will be effective in the future.

VII. ACKNOWLEDGMENT

The authors of this research express their sincere appreciation to the Institute of Research and Development, Suan Sunandha Rajabhat University, Thailand for financial support of the study.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Saowaros, A. (2007). "Relation between Big Five Personality and Boring in Career: Case Study of One Individual Hospital". Master Degree Thesis of Thammasat University, Thailand.

[2] Wijit, A. (2002). "Personality Development Technique and Method (2nd Ed.)". Bangkok: O.S. Printing House.

[3] Kaewchinporn, C. (2010). "Data Classification with Decision Tree and Clustering Techniques". Thesis in Computer Science, King Mongkut's Institute of Technology Ladkrabang, Thailand.

[4] Uhmn, S., Kim. D.H., and Kim. J. (2007). "Chronic Hepatitis Classification using SNP data and Data Mining Techniques". IEEE computer society, pp. 81-86.

[5] Quinlan, J.R. (1986). "Induction of decision trees". Machine Learning, Vol. 1, pp. 81-106.

[6] Costa, P.T. and McCrae, R.R. (1992). "Revised NEO Personality Inventory and NEO Five-Factor Inventory Professional Manual". Odessa FL: Psychological Assessment Resources.

[7] Howard, P.J. and Howard, J.M. (2004). "The Big Five Quickstart: An Introduction to the Five-Factor Model of Personality for Human Resource Professionals (Revised)". Charlotte, North Carolina: Centacs.

[8] Phayoon, P. (2005). "Data Mining System Development by Decision Tree". Master Degree of Science System Development Project, Department of Information Technology, King Mongkut's Institute of Technology Ladkrabang.

[9] Wasana, W. (2009). "Psychiatry Research Paper Searching System Using Decision Tree: Case Study of Somdet Chaopraya Institute of Psychiatry". Master Degree of Science Special Case, Department of Information System Management, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok.

[10] Natthaphol, C. and et al. (2015). "Correlation Between Peraonality Profile and Psychological Distress in Medical Cadets and Medical Students during Basic Military Training". Journal of the Psychiatric Association of Thailand, Vol. 58, pp. 147-156.

[11] Aranya, C. (1997). "The Relationship between Personality, Organizational Climate, Organizational Commitment and Organizational Citizenship Behaviors of Private Vocational School Teachers in Nakhon Si Thammarat". Journal of Southern Technology, Vol. 6, pp. 59-66.

[12] Monchai, T. and Chusri. (1998). "PC Program for Personality Test". Journal of Technical Education Development, Vol. 11, pp. 7-11.

[13] Gokul, C., Jan, B., and Daniel, G. (2013). "Mining large-scale smartphone data for personality studies". Journal of Personal and Ubiquitous Computing, Vol. 17, pp. 433-450.

[14] Johann, S., Christina, K., and Manfred, T. (2009). "Personality Traits, Usage Patterns and Information Disclosure in Online Communities". HCI 2009 – People and Computers XXIII – Celebrating people and technology, pp. 169-174.

[15] Yoram, B., Michal, K., and Thore, G. (2012). "Personality and Patterns of Facebook Usage". Web Science'12, June 22-24, 2012, Evanston, IL, USA.

[16] Eakasith, P. (2014). "Introduction to Data Mining and Big Data Analytics". <http://dataminingtrend.com/2014>.