ISSN 2723-6471

Big Data Classification of Personality Types Based on Respondents' Big Five Personality Traits

Jennifer Chi [1, *]

[1] School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
[1] Jennifer.Chi@utdallas.edu*
* corresponding author

(Received: August 5, 2021; Revised: January 15, 2022; Accepted: April 21, 2022; Available online: May 24, 2022)

Abstract

A mixed model was introduced in this study, *k*-means clustering analysis for data examination, discriminant analysis for classification, and multilayer perceptron neural network analysis for prediction. After deleted inadequate samples and outliers, total number of observations was 1,009,998 for this study that was collected through on interactive online personality (i.e., big five personality traits) test in 2018. Empirical results based on the *k*-means clustering analysis identified four different personality clusters using the total score of big five personality traits (Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience). Results of the *k*-means clustering analysis were tested for accuracy using the discriminant analysis indicated that cluster means were significantly different, and showed that 95.8% of original grouped cases correctly classified. The multilayer perceptron neural network framework was utilized as a predictive model, showed a 5-5-4 neural network construction, in deciding the personality classification of participants: Training 99.5% of training grouped cases and 99.5% of testing grouped cases correctly classified. Results of this study may provide insight into the understanding of the personality of participants for further psychological, social, cultural, and economic considerations.

Keywords: Big Five Personality Traits; Personality Types; Classification; K-means Clustering Analysis; Discriminant Analysis; Multilayer Perceptron Neural Network.

1. Introduction

The American Psychological Association (APA), adapted from the *Encyclopedia of Psychology*, defines personality as "individual differences in characteristic patterns of thinking, feeling, and behaving" (https://www.apa.org/topics/personality). According to APA Dictionary of Psychology, personality trait defines "a relatively stable, consistent, and enduring internal characteristic that is inferred from a pattern of behaviors, attitudes, feelings, and habits in the individual" (https://dictionary.apa.org/personality-trait). Thus, personality traits reflect people's characteristic patterns of thoughts, feelings, and behaviors that implies consistency and stability -- someone who scores high on a specific trait, like extraversion is expected to be sociable in different situations and over time.

The study of personality traits can be useful in summarizing, predicting, and explaining an individual's conduct that have important implications for behavior. The most popular way of measuring traits is by administering personality tests on which people self-report about their own characteristics. The most widely used system of traits is called the Big Five Personality Test, includes five broad traits that can be remembered with the acronym OCEAN: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

The Big Five Personality Test measures five personality traits, which helps define and expand on each trait and provides a fundamental picture of the individual's personality traits, indicating "the breadth, depth, originality, and complexity of an individual's mental and experiential life" [22]. Each of the major traits from the big five personality traits can be divided into facets that give a more fine-grained analysis of someone's personality.

A brief description of the big five personality traits is (1) *Extraversion*: extent to which individuals engage with the external world and experience enthusiasm and other positive emotions. (2) *Agreeableness*: extent to which individuals value cooperation and social harmony, honesty, decency, and trust worthiness. Agreeable individuals also tend to have an optimistic view of human nature. (3) *Conscientiousness*: extent to which individuals value planning, possess the quality of persistence, and are achievement-oriented. (4) *Neuroticism*: extent to which individuals experience negative feelings and their tendency to emotionally overreact. (5) *Openness to Experience*: extent to which individuals exhibit intellectual curiosity, self-awareness, and individualism/nonconformance [28].

Recently, many studies have been carried out on understanding human personality classification using advanced techniques [2,14,16,25,30,32]. Specifically, Gerlach et al. [16] has proposed four personality types, Role Model, Average, Reserved, and Self-Centered, from more than million participants using four large data sets, which brought more further discussions in academia [13,17,24].

In terms of personality classification, specifically, this study tried to explore segmentation of the participants based on certain perceives of interest regarding the Big Five Personality Traits, and to investigate how the participants' behavior can be identified using the multilayer perceptron neural network framework, based on information obtained from the traditional survey. Furthermore, by learning to recognize the current trends of the participants' perceptions, the multilayer perceptron neural network could make prediction in future outcomes within a campaign.

Therefore, the main objectives of this study included (1) to understand participants' perception to the Big Five Personality Traits; (2) to identify participant groups exhibiting common patterns of responses in terms of the Big Five Personality Traits; and (3) to classify participant associated with the Big Five Personality Traits using the multilayer perceptron neural network approach. This paper is organized as follows: the second section shows the data source in terms of the Big Five Personality Traits, while the third section presents the methodological approach. The fourth section demonstrates the empirical results using k-means cluster analysis, discriminant analysis, and multilayer perceptron neural network. The last section provides concluding remarks, and further discussion.

2. Data Source

The data used in this study was extracted from Answers to the Big Five Personality Test, constructed with items from the International Personality Item Pool (https://openpsychometrics.org/_rawdata/). This data was collected through an interactive online personality test done in 2018. Participants were informed that their responses would be recorded and used for research at the beginning of the test and asked to confirm their consent at the end of the test. Respondents were asked to indicate the Big Five Personality Traits containing 50 statements, ten questions that address each personality factor (Table 1), using a five-point Likert scale where 1 = Disagree, 3 = Neutral, 5 = Agree, and 0 = missed.

Table 1. The Big Five Personality Traits

Extraversion = SUM(E1:E10)	
E1	I am the life of the party.
E2	I don't talk a lot.
E3	I feel comfortable around people.
E4	I keep in the background.
E5	I start conversations.
E6	I have little to say.
E7	I talk to a lot of different people at parties.
E8	I don't like to draw attention to myself.
E9	I don't mind being the center of attention.
E10	I am quiet around strangers.
Neuroticism = SUM(N1:N10)	
N1	I get stressed out easily.
N2	I am relaxed most of the time.
N3	I worry about things.
N4	I seldom feel blue.
N5	I am easily disturbed.
N6	I get upset easily.
N7	I change my mood a lot.

N8	I have frequent mood swings.
N9	I get irritated easily.
N10	I often feel blue.
Agreeableness = SUM(A1:A10)	
A1	I feel little concern for others.
A2	I am interested in people.
A3	I insult people.
A4	I sympathize with others' feelings.
A5	I am not interested in other people's problems.
A6	I have a soft heart.
A7	I am not really interested in others.
A8	I take time out for others.
A9	I feel others' emotions.
A10	I make people feel at ease.
Conscientiousness = SUM(C1:C10)	
C1	I am always prepared.
C2	I leave my belongings around.
C3	I pay attention to details.
C4	I make a mess of things.
C5	I get chores done right away.
C6	I often forget to put things back in their proper place.
C7	I like order.
C8	I shirk my duties.
C9	I follow a schedule.
C10	I am exacting in my work.
Openness to Experience = SUM(O1:O10)	
O1	I have a rich vocabulary.
O2	I have difficulty understanding abstract ideas.
O3	I have a vivid imagination.
O4	I am not interested in abstract ideas.
O5	I have excellent ideas.
O6	I do not have a good imagination.
O7	I am quick to understand things.
O8	I use difficult words.
O9	I spend time reflecting on things.
O10	I am full of ideas.

Initial sample size of the pool was 1,015,342. After deleted 5,344 inadequate samples (too many zeros), the total working number of observations was 1,009,998. This amount was used for further analysis to examine the psychometric properties of the big five personality traits, by taking the sum of each trait of the big five personalities, respectively, for this study (Table 2).

Table 2. Descriptive Statistics of the Big Five Personality Traits

	Mean	Standard Deviation
Extraversion	30.33	3.78
Neuroticism	30.34	6.52
Agreeableness	31.66	3.61
Conscientiousness	31.34	3.91
Openness to Experience	32.77	3.87

3. Methods

In this study, a mixed model was introduced – *k*-means clustering analysis for data examination, discriminant analysis for classification, and multilayer perceptron neural network for prediction. Clustering is often used as a market segmentation approach to uncover similarity among customers or uncover an entirely new segment altogether. The *k*-means clustering analysis is used to find clusters which has not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

Empirically, the *k*-means clustering analysis tries to find homogeneous clusters within the data, so that the data points in each cluster consist of similarity within clusters and difference between clusters, according to a similarity measure such as a Euclidean-based distance [4]. Methodologically, *k*-means is an iterative algorithm that form groups of observations around geometric centers called centroids into clusters [6]. The algorithm calculates the centroids, which is determined by the individual conducting the analysis, and assigns a data point to that cluster having least distance between its centroid and the data point. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

Discriminant analysis is often used in combination with the k-means clustering analysis. Discriminant analysis is a statistical technique used to classify the target population into specific categories or clusters based on certain attributes (independent variables) [3]. For any kind of discriminant analysis, some cluster assignments should be known beforehand. Discriminant analysis is also a method of predicting some level of a one-way classification based on known values of the responses. This method is based on how close the measurement variables are to the multivariate means of the levels being predicted. In other words, it is useful in determining whether a set of variables are effective in predicting category membership [8].

Most multivariate analytical techniques can be used in some way to create post hoc market segments. Moreover, neural networks are useful in a broad spectrum of ways, but one of the most popular applications is to the marketing world. Neural networks can be essential in market segmentation because many of them are adopted at the practice of classifying or grouping customers into identifiable groups according to customer characteristics. In fact, neural network is a computing technique designed to simulate the human brain's method in problem-solving. It is one of the most popular machine learning methods which is able to do classification, clustering and prediction tasks.

According to Haykin [19], neural networks form a directed graph by connecting the artificial neurons, the basic information processing components of the network. Mathematically, the output on the neuron can be expressed as follows:

$$f(x) = \varphi(\Sigma_{i=1} x_i w_i + b) \qquad i = 1, \ldots, \mathrm{n} \tag{1}$$

where the x_i are the input features, the w_i are the weights of respective inputs, b is the bias, which is summed with the weighted inputs to form the net inputs, and φ is the non-linear activation function. Bias and weights are both adjustable parameters of the neuron. Thus, it needs a mapping mechanism between the input and output of the neuron. This mechanism of mapping inputs to output is known as activation function [19].

A simple perceptron is a linear classifier that produces a single output based on several real-valued inputs by forming a linear combination using its input weights. Multilayer perceptron (MLP) (Figure 1) consists of multiple layers of working units, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. In the theoretical manner, MLP is a universal approximator, and with respect to its inherent nature, it has a tremendous capacity of constructing any nonlinear mapping to any extent of accuracy. It does not need a priori model to be assumed or a priori assumptions to be made on the properties of data [4].

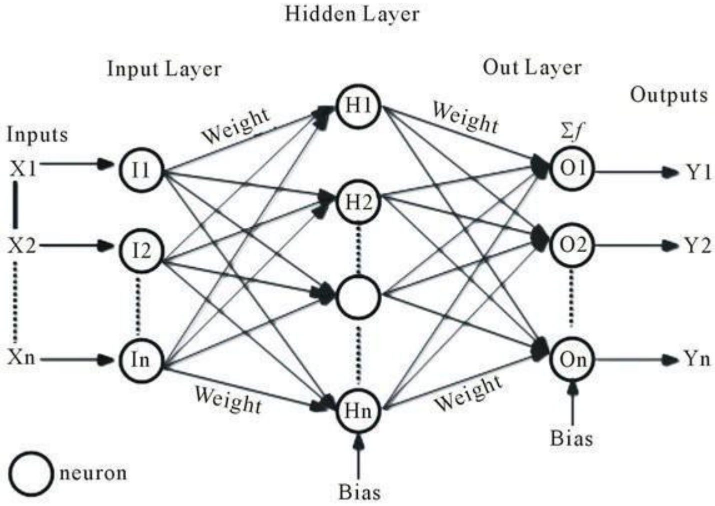

Figure 1. Single Hidden Layer MLP (Adapted from [11])

Gardner and Dorling [15] define multilayer perceptron as: "*a system of simple interconnected neurons, or nodes, which is a model representing a nonlinear mapping between an input vector and an output vector*". MLP is the most utilized model in neural network applications using the back-propagation training algorithm for multilayer feed-forward networks. MLP consists of perceptrons that are organized in layers: an input layer, one or more hidden layers, and the output layer.

Each perceptron calculates the sum of the weighted inputs, and feeds it into its activation function. The result is then passed on to the next layer. The output layer has the same number of perceptrons as there are classes, and the perceptron with the highest activation will be consider the classification of the input sample. Training is achieved by successively feeding all training samples into the network, and comparing the output with the true class label [19].

MLP is the most popular neural network method that has been widely used for many practical applications, and one good reason is that able to learn non-linear representations. It has been widely employed for modeling, prediction, classification, clustering, and optimization purposes [1,5,9,10.27.33].

4. Results

4.1. K-means Clustering Analysis

The k-means clustering analysis techniques assign objects to groups so that there is as much similarity within groups, and difference between groups, as possible. In this study, a k-means clustering analysis was applied to find homogeneous clusters within the 1,009,998 respondents by using the sum of each trait of the big five personalities respectively. Consequently, a four-cluster solution was identified, which was labeled as *Self-Centered*, *Reserved*, *Averag*e, and *Role Models* clusters (i.e., personality types) [16].

The *Self-Centered* personality type: this was the smallest group comprising of 6 percent of the respondents. These respondents received the average scores of all five personality traits were below the average score of the all samples.

The *Reserved* personality type: with 23.5 percent of the respondents, this group was named because the average scores of all five personality traits were above the average score of the all samples.

The *Average* personality type: with 33.4 percent of the respondents, this group was named because the average scores of Extraversion, Agreeableness, Conscientiousness, Openness to Experience were below the average score of the all samples, except the average score of Neuroticism was far above the average score of the all samples.

The *Role Models* personality type: this cluster was the largest group, comprising of 37.1 percent of respondents, named because the average scores of Extraversion, Agreeableness, Conscientiousness, Openness to Experience were above the average score of the all samples, but the average score of Neuroticism was far below the average score of the all samples (Table 3).

Table 3. K-means Clustering Analysis of Respondents' Big Five Personality Traits

	Self-Centered		*Reserved*		*Average*		*Role Models*	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Extraversion	28.17	4.00	31.33	3.75	29.45	3.11	30.85	3.17
Neuroticism	23.48	4.22	37.35	3.98	33.22	3.20	24.43	3.38
Agreeableness	27.72	4.11	33.84	3.39	30.86	3.10	31.63	3.22
Conscientiousness	26.26	3.85	34.35	3.46	30.10	3.14	31.37	3.31
Openness to Experience	27.39	4.15	34.75	3.45	31.44	3.42	33.58	3.19
n = 1,009,998	60,663		237,033		337,572		374,730	
Percentage	6.0%		23.5%		33.4%		37.1%	

4.2. Discriminant Analysis

Discriminant analysis is a statistical technique to classify the target population into the specific categories or groups based on the certain attributes (predictor variables or independent variables) [12,31]. The objective of discriminant analysis is to develop discriminant functions that are nothing but the linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. It enables to examine whether significant differences exist among the groups, in terms of the independent variables. It also evaluates the accuracy of the classification [8].

Results of the k-means clustering analysis were tested for accuracy using the linear discriminant analysis employed as a useful complement to the k-means clustering analysis, which is used primarily to predict membership in two or more mutually exclusive groups. Therefore, a discriminant analysis was employed to classify the 1,009,998 respondents into specific personality types based on their answers related to the big five personality traits. In this case, the Wilk's Lambda scores were 0.159 (χ^2 = 1854179.748, df = 15, p < 0.001), 0.596 (χ^2 = 5230.6.959, df = 8, p < 0.001), and 0.995 (χ^2 = 4978.858, df = 3, p < 0.001) for both discriminant functions, respectively, indicating that group means were significantly different.

The results based on discriminant analysis, 60,663 cases fell into the *Self-Centered* personality type, 237,033 fell into the *Reserved* personality type, 337,572 fell into the *Average* personality type, and 374,730 fell into the *Role Models* personality type in the original row total, which is the frequencies of groups found in the data (Table 4). Across each row, the case amount in the group can be classified by this analysis into each group. For example, of the 60,663 cases that were in the *Self-Centered* personality type, 60,651 were predicted correctly and 12 were predicted incorrectly (12 was predicted to be in the *Average* personality type).

Predicted group membership indicates the predicted frequencies of groups from the analysis. The numbers going down each column indicate how many were correctly and incorrectly classified. For example, of the 90,173 cases that were predicted to be in the *Self-Centered* personality type, 60,651 were correctly predicted, and 29,522 were incorrectly predicted (5,100 cases were in the *Average* personality type and 24,422 cases were in the *Role Models* personality type).

Table 4. Classification Results[a] Based on the Discriminant Analysis

		Personality Type	Predicted Group Membership				
			Self-Centered	*Reserved*	*Average*	*Role Models*	Total
Original	Count	*Self-Centered*	60651	0	12	0	60663
		Reserved	0	234457	2297	279	237033
		Average	5100	6597	325741	134	337572
		Role Models	24422	477	3327	346504	374730
	%	*Self-Centered*	100.0	0.0	0.0	0.0	100.0

		Reserved	0.0	98.9	1.0	0.1	100.0
		Average	1.5	2.0	96.5	0.0	100.0
		Role Models	6.5	0.1	0.9	92.5	100.0

a. 95.8% of original grouped cases correctly classified

4.3. MLP Neural Network

After the formation of the identified four personality types, an MLP neural network was employed as a predictive model in deciding the classification of the respondents based on their perceptions toward the big five personality traits. The MLP Module of IBM SPSS Statistics 26 was used as the tool to build the neural network model and to test its accuracy. The MLP neural network model, trained with a back-propagation learning algorithm which uses the gradient descent to update the weights towards minimizing the error function [21].

Initially, the data was randomly assigned to training (70%) and testing (30%) subsets. The training dataset was used to find the weights and to build the neural network model, while the testing data was used to find errors and to prevent overtraining during the training mode. Randomly, 1,009,998 data samples were divided into 708,107 data samples for the training, and 301,891 data samples for the testing. The neural network model is constructed with the multilayer perceptron algorithm.

In order to find the best MLP neural network, disparate possible networks were tested and it concluded that the MLP neural network with a single hidden layer was the best option for this study. Sheela and Deepa [29] pointed out that as the number of neurons or the number of layers of a neural network increase, the training error also increases due to the overfitting. It is clear that using a single input layer, a single hidden layer, and a single output layer in the MLP neural network will help to decrease the probability of overfitting and will require relatively lower computational time.

The MLP Module of IBM SPSS Statistics 26 was used as the tool to choose the best architecture model automatically and it built the network with one hidden layer. The hyperbolic tangent was used as the activation function in the hidden layer, while the softmax function was used as the activation function in the output layer. Cross-entropy was used as error function because of the use of softmax function. Intuitively, the cross-entropy loss function is used to measure the error at a softmax layer, typically the final output layer in a neural network.

One of the most salient considerations in the construction of neural networks is choosing activation functions for hidden and output layers that are differentiable. The results showed that in this study, the hyperbolic tangent activation function can be used for the single hidden layer because it cannot be used in networks with many layers due to the vanishing gradient problem. Also, the rectified linear activation function can be used for the output layer not only because it overcomes the vanishing gradient problem, but allows models to learn faster and perform better [18].

From the five independent variables in the input layer, the architecture automatically selected five nodes in the hidden layer, and the output layer had four nodes as the dependent variable named *Cluster*. The network diagram showed the five input nodes, the five hidden nodes and the four output nodes representing the four identified personality types. In the architectural point of view, it was a 5-5-4 neural network, means that there was total of five independent (input) variables, five neurons in the hidden layer, and four dependent (output) variables (Figure 2).

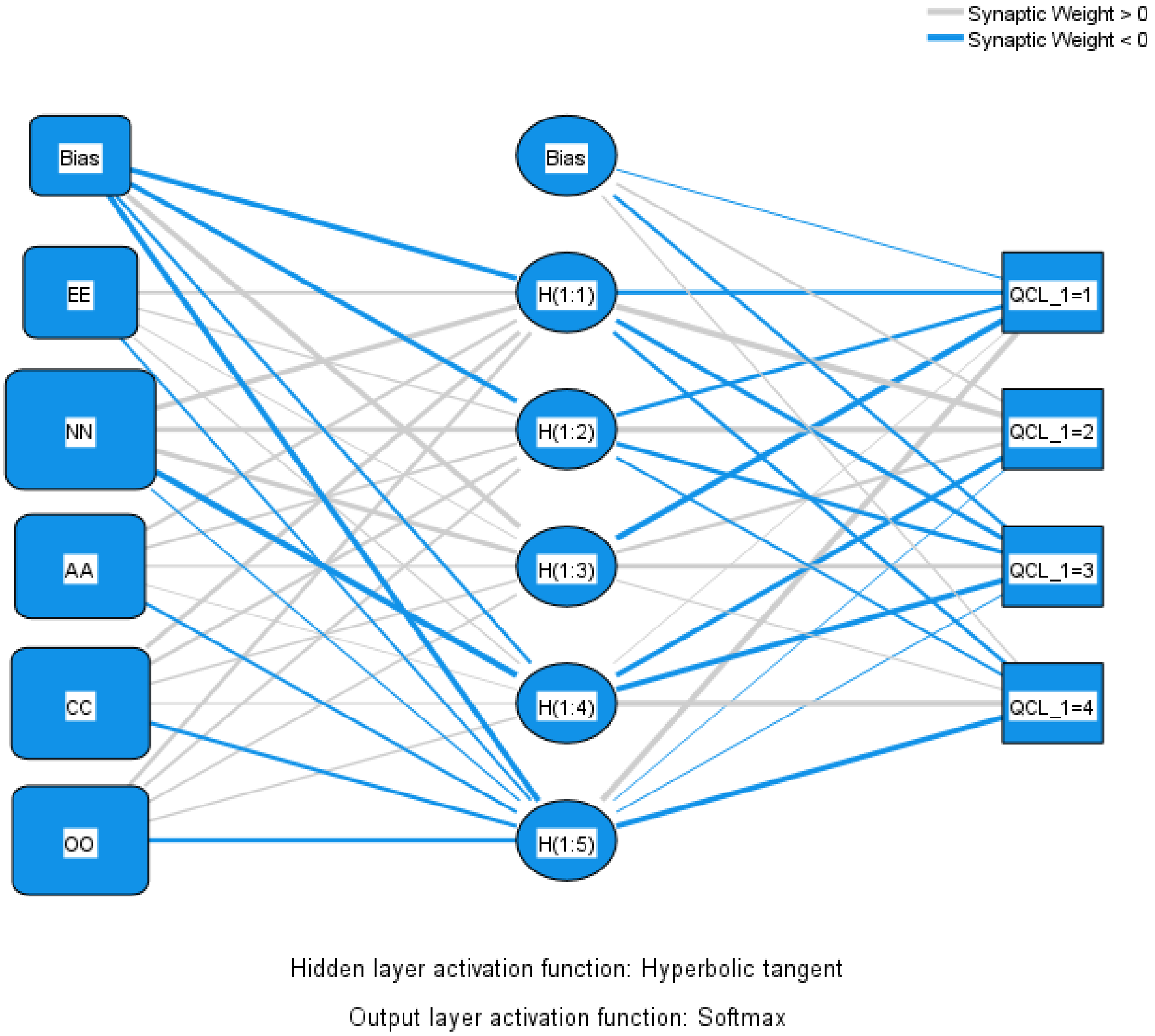

Figure 2. Network Diagram

The model summary provided information related to the results of training and testing sample (Table 5). Cross entropy error is displayed because the analysis is based on the softmax activation function, and is given for both training and testing sample since is the error function that minimizes the network during training phase [21]. The value of cross entropy error (= 12661.166) indicated the power of the model to predict the four identified personality types. The cross entropy error was less for the testing sample compared with the training data set, meaning that the network model had not been over-fitted to the training data and has learned to generalize from trend. The result justified the role of testing sample which was to prevent overtraining.

In this study, the percentage of incorrect prediction was equal to 0.5% in the training sample. Therefore, the percentage of correct prediction was 99.5% which is an excellent prediction in a qualitative study for determining the results of the big five personality traits for the four identified personality types. The learning procedure was performed until one consecutive step with no decrease in error function was attained from the training sample.

Table 5. Model Summary (Dependent Variable: Cluster)

Training	Cross Entropy Error	12661.166
	Percent Incorrect Predictions	0.5%
	Stopping Rule Used	Maximum number of epochs (100) exceeded
	Training Time	0:03:51.81
Testing	Cross Entropy Error	5479.766

	Percent Incorrect Predictions	0.5%

Using the training sample only, the MLP neural network utilized synaptic weights to display the parameter estimates that showed the relationship between the units in a given layer to the units in the following layer (Table 6). Note that the number of synaptic weights can become rather large, and that these weights are generally not used for interpreting network results [21].

Table 6. Parameter Estimates

		Predicted								
		Hidden Layer 1					Output Layer			
Predictor		H(1:1)	H(1:2)	H(1:3)	H(1:4)	H(1:5)	Cluster1	Cluster2	Cluster3	Cluster4
Input Layer	(Bias)	-3.897	-3.536	4.047	-1.617	-5.466				
	EE	1.146	0.703	0.372	0.583	-1.083				
	NN	3.720	3.366	3.459	-7.624	-0.640				
	AA	1.808	1.145	0.795	0.261	-1.538				
	CC	2.682	1.844	0.989	0.478	-2.066				
	OO	2.074	1.435	1.032	0.905	-2.519				
Hidden Layer 1	(Bias)						-0.425	1.078	-1.802	0.672
	H(1:1)						-2.568	6.760	-3.090	-1.910
	H(1:2)						-2.080	5.807	-2.671	-1.100
	H(1:3)						-6.088	2.044	3.125	0.517
	H(1:4)						0.083	-3.647	-5.360	8.388
	H(1:5)						5.980	-0.034	-0.403	-5.032

Based on the MLP neural network, a predictive model was developed and displayed a classification table (i.e., confusion matrix) for categorical dependent variable the four identified personality types, by partition and overall (Table 7). As can be seen, the MLP neural network correctly classified 704,872 participants out of 708,107 in the training sample and 300,499 out of 301,891 in the testing sample. Overall, 99.5% of the training and 99.5% of the testing cases were correctly classified. The predictive model developed had excellent classification accuracy.

Table 7. Predictive Ability and Classification Results (Dependent Variable: Cluster)

Classification						
Sample	Observed	Predicted				
		Self-Centered	*Reserved*	*Average*	*Role Models*	Percent Correct
Training	*Self-Centered*	42023	0	117	264	99.1%
	Reserved	0	165721	129	303	99.7%
	Average	74	219	236053	333	99.7%
	Role Models	275	1009	512	261075	99.3%
	Overall Percent	6.0%	23.6%	33.4%	37.0%	99.5%
Testing	*Self-Centered*	18081	0	45	133	99.0%
	Reserved	0	70675	54	151	98.7%
	Average	34	106	100612	141	99.7%
	Role Models	127	373	228	111131	99.3%
	Overall Percent	6.0%	23.6%	33.4%	37.0%	99.5%

Using the training sample only, it was able to classify 261,075 *Role Models* participants in the *Role Models* personality type, out of 262,871. It held 99.3% classification accuracy for the *Role Models* personality type. Similarly, the same model was able to classify 236,053 *Average* participants in the *Average* personality type out of 236,679, 165,721 *Reserved* participants in the *Reserved* personality type out of 166,153, and 42,023 *Self-Centered* participants in the *Self-Centered* personality type out of 42,404. It was able to generate 99.1% classification accuracy for the *Self-Centered* personality type, and 99.7% classification accuracy for both the *Reserved* and *Average* personality types.

The Receiving Operating Characteristic (ROC) curve is a two-dimension graph commonly used to measure the performance of classification problems [23,26,34]. A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations. Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations. For example, in medical testing, the true positive rate is the rate in which people are correctly identified to test positive for the disease in question.

The ROC curve is a diagram of sensitivity (or TPR) versus specificity (1 – FPR) that shows the classification performance for all possible cutoffs. A commonly used approach when selecting a cut-off point is to give equal weight to the importance of sensitivity and specificity by choosing the point nearest to the top-left most corner of the ROC curve. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test [23,26,34].

In order to check or visualize the performance of the multi-class classification problem, the area under the ROC curve (AUC) is a performance measurement for the classification problems at various threshold settings [23,26,34]. It tells how much the model is capable of distinguishing between classes. By analogy, the higher the AUC, the better the model is at distinguishing between patients with the disease and no disease. Thus, an excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has AUC near to the 0 which means it has the worst measure of separability.

As illustrated in Figure 3, the result showed the classification in this study performed excellent to distinguishing between personality types. At the same time, the result showed that there was AUC = 1.000, indicated that the classifier was able to perfectly distinguish between all the positive and the negative class points correctly.

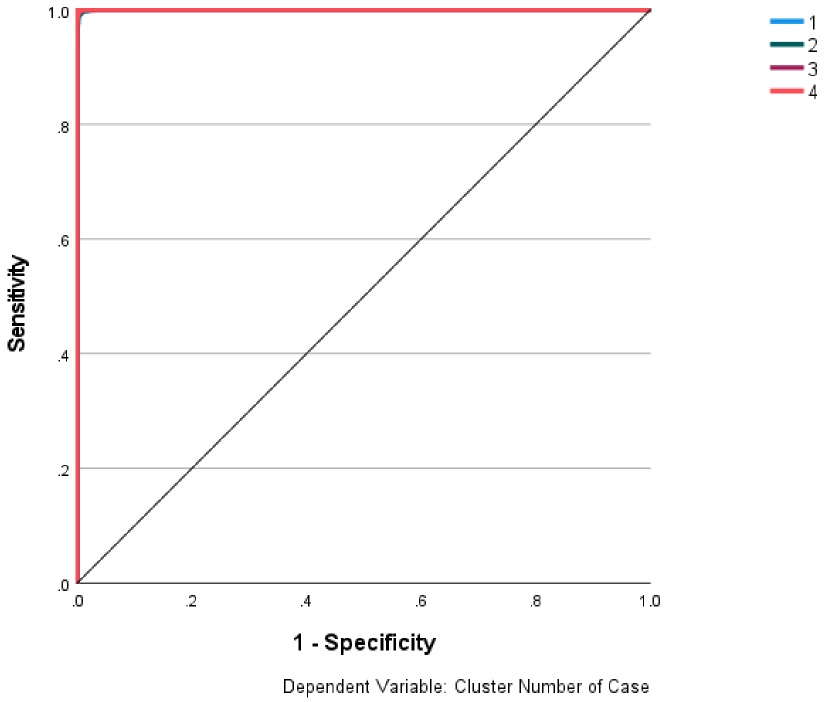

Figure 3. ROC Curve

The importance of the individual independent variables (factor influencing the personality type) is a measure of how much the network model predicted value changes for different independent variables [21]. The input parameters – the big five personality traits which influenced the four identified personality types have been ranked by the neural network model were given in the following Table 8. Hence, independent variable importance analysis provides the sensitivity analysis, by computing the importance of each independent variable, which in turn determines the structure of the neural network.

The first significant dominant factors that has been found was "Neuroticism" (100%), contributed the most in the neural network model construction, followed by "Conscientiousness" (88.9%), and "Openness to Experience" (86.2%), had the greatest effect on how the participants' perceptions, in terms of the big five personality traits. The next important factor was "Agreeableness" (79.9%), and the least important factor which was identified as "Extraversion" (64.1%).

Table 8. Independent Variable Importance Analysis

	Importance	Normalized Importance	Rank
Extraversion	0.153	64.1%	5
Neuroticism	0.239	100.0%	1
Agreeableness	0.191	79.9%	4
Conscientiousness	0.212	88.9%	2
Openness to Experience	0.206	86.2%	3

5. Conclusions

Understanding human personality can help us to recognize how people will respond to certain situations and their preferences and values, in terms of individual differences. There are many approaches that can be used to identify one's personality type (i.e., the Big Five Personality Traits). In business, for example, identifying human personality types could be useful for recognizing how we lead, influence, communicate, collaborate, negotiate business and manage stress.

In this study, a mixed model was introduced, *k*-means clustering analysis for data examination, discriminant analysis for classification, and multilayer perceptron neural network for prediction. Overall, this study adopted *k*-means clustering analysis to identify four personality types, named *Role Models* personality type (37.1% of 1,009,998 respondents), *Average* personality type (33.4%), *Reserved* personality type (23.5%), and *Self-Centered* personality type (6.0%).

Role Models have high levels of extraversion, agreeableness, conscientiousness and openness to experience, and comparably low levels of neuroticism. *Average* people are high on neuroticism, and below average on extraversion, agreeableness, conscientiousness and openness to experience. *Reserved* individuals are above average on all five traits, particularly high on neuroticism. *Self-centered* people are below average on all five traits, particularly low on neuroticism.

Theoretically, a cluster is a collection of items that are similar among themselves and are dissimilar to the items belonging to other clusters. It can be shown that there is no absolute best criterion, which would be independent of the final aim of the clustering. Hence, the structure of the clusters should be finalized by the user depending on the physical requirements. Thus, there is no unique approach to correctly classify the participants who provided the information of the Big Five Personality Traits.

The classification results based on discriminant analysis showed that 95.8% of original grouped cases are correctly classified. After the formation of the four identified clusters, an MLP neural network model was employed as a predictive model in deciding the classification of the respondents associated with their Big Five Personality Traits. As a result, 99.5% of the training cases were correctly classified, revealing that the predictive model developed had excellent classification accuracy.

The MLP neural network is widely considered as an efficient approach to adaptively classify patterns. In this work, an attempt was made to improve the learning capabilities of an MLP neural network and reduced the amount of time and resource required by the learning process. The multilayer perceptron neural network model was utilized as a predictive model in deciding the classification of the respondents based on their Big Five Personality Traits. The results show a 5-5-4 neural network from an architectural perspective, and also revealed that neuroticism and conscientiousness were the greatest effect on how the respondents perceives in terms of the Big Five Personality Traits.

Due to the nature of the data set, it only contained the information related to the Big Five Personality Traits individually. Technically, the results of this study can be the reference of the human personality classification. Thus, the main limitations of the study for the further research should consider including not only participants' socio-economic characteristics, i.e., age, cohort, gender, but also major business applications, i.e., business leadership.

References

[1] F. E. Ahmed, "Artificial neural networks for diagnosis and survival prediction in colon cancer," Molecular Cancer, 4:29, 1-12, 2005.

[2] H. Ahmad, M. Z. Asghar, A. S. Khan, and A. Habib, "A systematic literature review of personality trait classification from textual content," Open Computer Science, 10(1), 175-193, 2020.

[3] T. Beatley, "Protecting biodiversity in coastal environments: introduction and overview," Coastal Management, 19(1), 1–19, 1991.

[4] C. M. Bishop, "Pattern recognition and machine learning," New York, NY: Springer Science + Business Media, 2006.

[5] B. K. Bose, "Neural network applications in power electronics and motor drives - an introduction and perspective," IEEE Transactions on Industrial Electronics, 54(1), 14-33, 2007.

[6] D. Child, "The essentials of factor analysis (3rd ed.)," New York, NY: Continuum International Publishing Group, 2006.

[7] G. A. Churchill, Jr., and D. Iacobucci, "Marketing research: methodological foundations (9th ed.)," Mason, OH: Thomson/South-Western, 2005.

[8] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," Psychometrika, 16(3), 297-334, 1951.

[9] J. G. De Gooijer, and R. J. Hyndman, "25 years of time series forecasting," International Journal of Forecasting, 22(3), 443-473, 2006.

[10] L. N. N. Do, N. Taherifar, and H. L. Vu, "Survey of neural network-based models for short-term traffic state prediction," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(1), 1-24, 2019.

[11] H. El-Amir, and M. Hamdy, "Deep learning pipeline: building a deep learning model with TensorFlow," Berkerly, CA: Apress, 2020.

[12] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, 7, 179-188, 1936.

[13] J. P. Freudenstein, C. Strauch, P. Mussel, and M. Ziegler, "Four personality types may be neither robust nor exhaustive," Nature Human Behaviour, 3(10), 1045-1046, 2019.

[14] C. Gaisendrees, N. Kreuser, O. Lyros, J. Becker, J. Schumacher, I. Gockel, A. Kersting, and R. Thieme, "Classification of personality traits using the Big Five Inventory-10 in esophageal adenocarcinoma patients," Annals of Esophagus, 3:22, 1-8, 2020

[15] M. W. Gardner, and S. R. Dorling, "Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences," Atmospheric Environment, 32(14), 2627-2636, 1998.

[16] M. Gerlach, B. Farb, W. Revelle, and L. A. N. Amaral, "A robust data-driven approach identifies four personality types across four large data sets," Nature Human Behaviour, 2(10), 735-742, 2018.

[17] M. Gerlach, W. Revelle, and L. A. N. Amaral, "Reply to: Four personality types may be neither robust nor exhaustive," Nature Human Behaviour, 3(10), 1047-1048, 2019.

[18] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," The MIT Press, 2016.

[19] S. S. Haykin, "Neural networks and learning machines (3rd ed.)," Upper Saddle River, New Jersey: Pearson Education, Inc., 2009.

[20] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, 2, 359-366, 1989.

[21] IBM, "IBM SPSS Neural Networks 26," Armonk, NY: IBM Corporation, 2019.

[22] O. P. John, and S. Srivastava, "The big-five trait taxonomy: history, measurement, and theoretical perspectives," In L. A. Pervin, and O. P. John, (Eds.), "Handbook of personality: theory and research," Vol. 2, pp. 102-138, New York, NY: Guilford Press, 1999.

[23] C. M. Jones, and T. Athanasiou, "Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests," The Statistician's Page, 79(1), 16-20, 2005.

[24] K. Katahira, Y. Kunisato, Y. Yamashita, and S. Suzuki, "Commentary: A robust data-driven approach identifies four personality types across four large data sets," Frontiers in Big Data, 3(8), 1-3, 2020.

[25] A. S. Khan, H. Ahmad, M. Z. Asghar, F. K. Saddozai, A. Arif, and A. Kalid, "Personality classification from onlinr text using machine learning approach," International Journal of Advanced Computer Science and Applications, 2020, 11(3), 460-476, 2020.

[26] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," Journal of Thoracic Oncology, 5(9), 1315-1316, 2010.

[27] H. Ramchoun, M. A. Janati Idrissi, Y. Ghanou, and M. Ettaouil, "New modeling of multilayer perceptron architecture optimization with regularization: an application to pattern classification," IAENG International Journal of Computer Science, 44(3), 261-269, 2017.

[28] R. J. Rossberger, "National personality profiles and innovation: The role of cultural practices," Creativity and Innovation Management, 23(3), 331–348, 2014.

[29] K. G. Sheela, and S. N. Deepa, "Review on methods to fix number of hidden neurons in neural networks," Mathematical Problems in Engineering, Article ID 425740, 1-11, 2013.

[30] A. Souri, S. Hosseinpour, and A. M. Rahmani, "Personality classification based on profiles of social networks' users and the five-factor model of personality," Human-centric Computing and Information Sciences, 8(1), 8-24, 2018.

[31] B. G. Tabatchnick, and L. S. Fidell, "Using multivariate statistics (6th ed.)," Boston, MA: Pearson Education, Inc., 2013.

[32] A. Talasbek, A. Serek, M. Zhaparov, S. Moo-Yoo, Y. Kim, and G. Jeong, "Personality classification experiment by applying k-means clustering," International Journal of Emerging Technologies in Learning, 15(16), 162-177, 2020.

[33] N. Z. Zacharis, "Predicting student academic performance in blended learning using artificial neural networks," International Journal of Artificial Intelligence and Applications, 7(5), 17-29, 2016.

[34] K. H. Zou, A. J. O'Malley, and L. Mauri, "Receiver operating characteristic analysis for evaluating diagnostic tests and predictive models," Circulation, 115(5), 654-657, 2007.