HUMAN PERSONALITY CLASSIFICATION USING SUPERVISED MACHINE LEARNING ALGORITHMS

K.M.G.S. Karunarathna[1], M.P.R.I.R. Silva[2], R.A.H.M. Rupasingha[3*]
[1,2]Department of Information Technology, Sabaragamuwa University of Sri Lanka
[3]Deparment of Economics and Statistics, Sabaragamuwa University of Sri Lanka

Abstract

According to certain definitions, "personality" refers to a person's distinctive ways of thinking, feeling, and behaving in a variety of situations. The personality can be used to identify the behavior patterns of a human. The goal of this research is to categorize human personalities based on their behaviors. This study used data that was collected as secondary data targeting the main five personality types considering behavioral tendencies, namely the supervisor, the commander, the inspector, the doer, and the idealist. After the pre-processing, six classification algorithms were used: Support Vector Machine (SVM), Random Forest, Naïve Bayes, Logistic regression, Multilayer Perception (MLP), and Decision Tree. The result was validated using 10-fold cross-validation. Based on the result, the highest accuracy is obtained in SVM with an accuracy value of 88.5%. The highest precision, recall, f-measure, and lower error rates are obtained by comparing the above six supervised machine learning algorithms.

Keywords: *Personality, Classification, Machine Learning*

**Corresponding author: Tel.: +94 (71) 832 4740; Email: hmrupasingha@gmail.com; ORCID: https://orcid.org/0000-0003-3922-4290*

Introduction

People's consistent traits, attitudes, feelings, and behaviors are referred to as their personalities (Song et al., 2021). A person's personality will naturally develop based on their activities, thoughts, feelings, and overall behavior under different conditions (R, 2021). From a computational perspective, developments in the Natural Language Processing and Machine Learning (ML) domains have produced a wide variety of automated methods for the identification of personality traits in the text (dos Santos et al., 2019).

It may be difficult for a person to accurately classify a personality from a text, as there are so many different factors at play. Therefore, our research aims toclassify personality traits automatically based on their behaviors, using five personality types, namely the supervisor, the commander, the inspector, the doer, and the idealist which is possible because the behaviors of the people are different. The main purpose of this research approach is creating a model to automatically identify the conventional personality type of the person based on their behaviors.

Although there has been recent research on a range of personality classification, the same classification algorithms are not used in earlier studies. Some are compared only the two or three classification algorithms or used very little data and the majority of personality study relied on observations like social media usage and handwriting patterns. Our data set was gathered from the questionnaire's replies (Ong et al., 2017). But our study mainly increased the number of data and increased the number of different ML algorithms until six to achieve better comparison and better results.

Material and Methods

Following Figure 1 show the proposed approach

Data collection

The secondary data set was used for this and the data set was obtained from Kaggle website (Kaggle, 2022). In this research we selected 1000 data from the secondary data set. Following Figure 2 show the five personality types we used and Figure 3 show the some of the attributes we used.

Data Preprocessing

We used the Waikato Environment for Knowledge Analysis (WEKA) data mining tool for data pre-processing. The data set contained 62 attributes and 16 personality types. After ranking the attributes by the information gain ranking algorithm, the top 34 attributes are selected. Furthermore, the types ofpersonality were reduced to 5 types based on their personality qualities.

Figure 1:

Proposed Approach

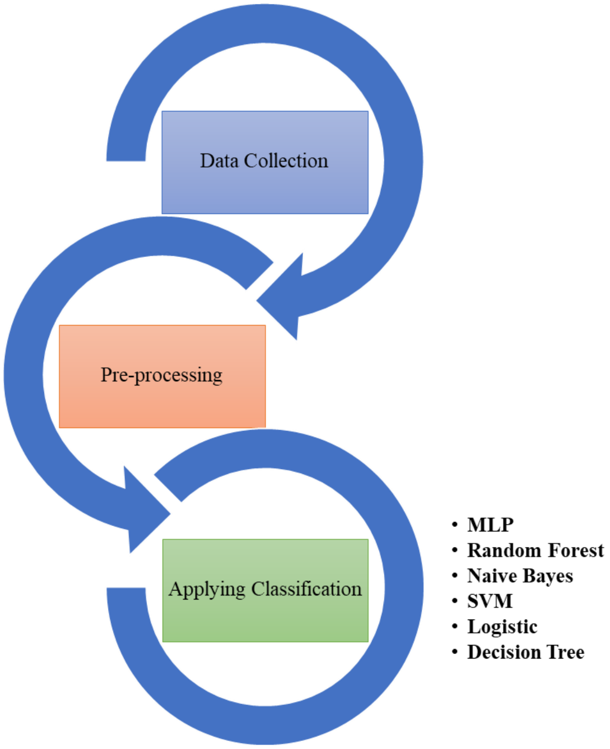


Figure 2:

Types of Personality

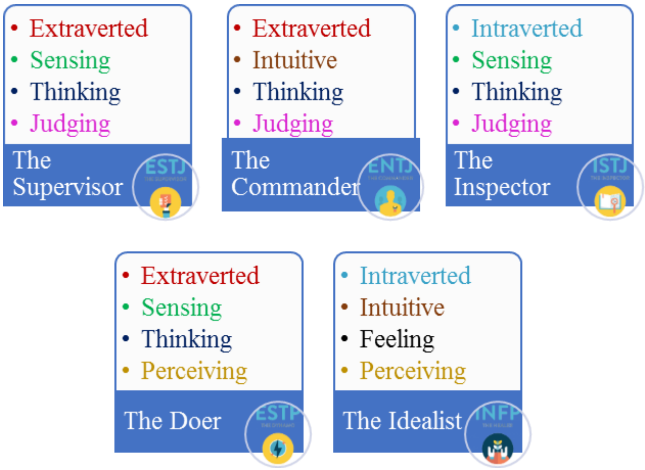
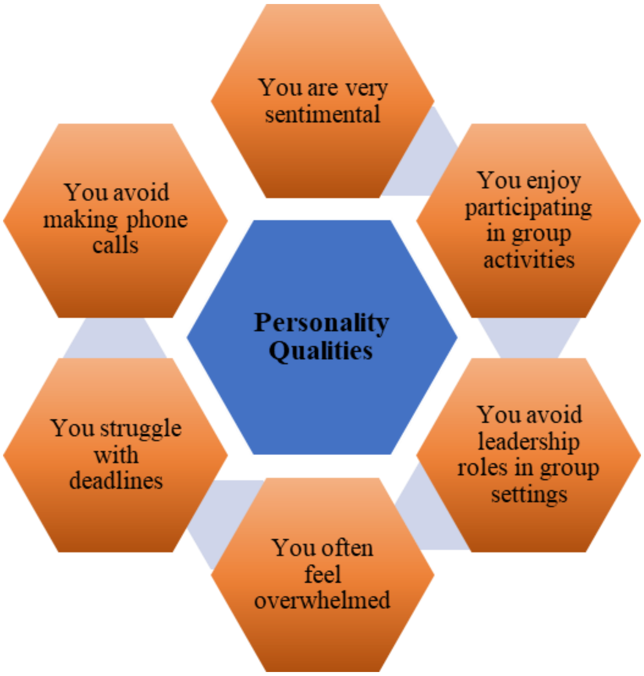


Figure 3:

Sample of Attributes


You are very sentimental
You avoid making phone calls
You enjoy participating in group activities
Personality Qualities
You struggle with deadlines
You avoid leadership roles in group settings
You often feel overwhelmed

Classification

The preprocessed data set is used for applying the classification process using the WEKA data mining tool. The prediction model is built up to identify the personality qualities. The data set was applied to the Random Forest, Naïve Bayes, Decision Tree (J48), Logistic, SVM, and MLP algorithms.

Results and Discussion

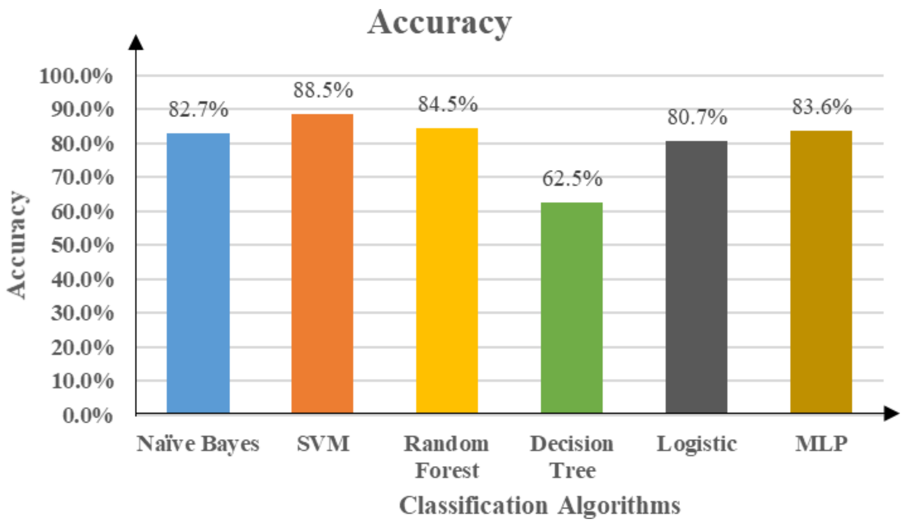
Microsoft Windows 10 on PC with Processor Intel@ Core i5, RAM 4.0GB has been used to obtain these comparative results. And the WEKA 3.8.5 tool is used for the data mining process.

Accuracy of the classification algorithm

The result with respect to accuracy was compared for the six machine learning algorithms. Figure 4 shows that the SVM has the highest accuracy among these six algorithms.

Figure 4:

Accuracy of algorithms



Results of precision, recall and f-measure

Following are the results of recall, precision, and f-measure obtained according to the equations given in (1), (2), and (3). Here, Ps, Px, Pstx denotes the number of all relevant members included in a specific cluster, the number of all members in a specific cluster, and the number of all relevant specified-cluster members in the corpus respectively.

$$Precision = \frac{P_s}{P_x} \qquad (1)$$

$$Recall = \frac{P_s}{P_{stx}} \qquad (2)$$

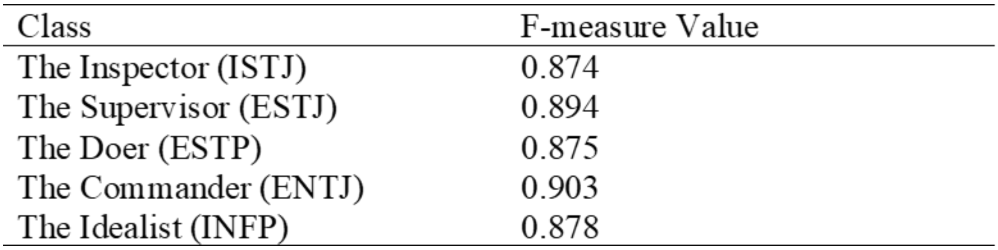
$$F\_measure = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (3)$$

According to the Table 1, the highest values were achieved by the SVM, which also achieved the highest accuracy.

Table 1:
Precision, recall & F-measure values

Algorithm	Precision	Recall	F-measure
SVM	0.885	0.885	0.885
Random Forest	0.846	0.845	0.845
MLP	0.836	0.836	0.836
Naïve Bayes	0.830	0.827	0.827
Logistics	0.807	0.807	0.807
Decision Tree	0.626	0.625	0.625

Table 2:
Class wise f-measure values in SVM algorithms



Class	F-measure Value
The Inspector (ISTJ)	0.874
The Supervisor (ESTJ)	0.894
The Doer (ESTP)	0.875
The Commander (ENTJ)	0.903
The Idealist (INFP)	0.878

Results of MAE and RMSE

We then calculated the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for all six algorithms, according to the following Eq. (4) and (5). Here T is the number of predicted values, Pvx represents the existing labeled value based on the outcomes, and Mvx represents the predicted result. Figure 5 gives the MAE and RMSE results.

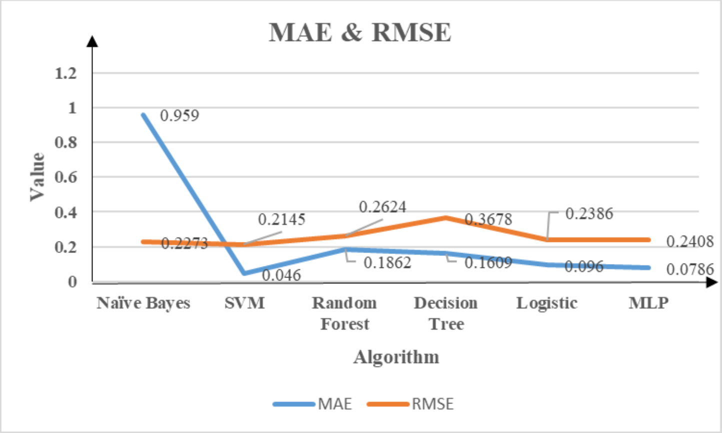
$$MAE = \frac{1}{T} \sum_{x=1}^{T} | \, p_{vx} - M_{vx} \, | \qquad (7)$$

$$RMSE = \sqrt{\frac{1}{T} \sum_{x=1}^{T} \left(p_{vx} - M_{vx,}\right)^2} \qquad (8)$$

Figure 5, the MAE and RMSE results were lowest for the SVM algorithm.

Figure 5:

MAE & RMSE results



5-fold cross-validation vs 10-fold cross-validation

Our research experiments used both 5-fold cross-validation and 10-fold cross-validation, with 10-fold cross-validation it obtaining better accuracy as shown in Table 2.

Table 3:

5-fold cross validation vs 10-fold cross validation

Fold	SVM		Accuracy
	Test Data	Training Data	
5	20%	80%	83.5%
10	10%	90%	88.5%

Conclusion and Recommendations

The main purpose of this is to identify the personality qualities of people in various fields and classify them. This enabled to build a system that can classify personality types considering characteristics of human behavior. After pre-processing the collected data, compared six classification algorithms. Among these algorithms, the SVM algorithm gave the most accurate result obtaining 88.5%. And also Random Forest 84.5%, MLP 83.6%, Naïve Bayes82.7%, Logistics 80.7%, and Decision Tree 62.5% were obtained as the accuracy results. In summary, the SVM algorithm was found to be the best ofsix individual algorithms with the highest accuracy, highest values for precision, recall, and f-measure and it also obtained the lowest error rate. In the comparison of 5-fold with 10-fold cross-validation, 10-fold obtained the highest accuracy.

In future work, we intend to increase the size of the data and plan to investigate the Ensemble Learning algorithm (Vote) combining above six classification algorithms to further increase the accuracy of the classification process.

References

dos Santos, W. R., Ramos, R. M. S., & Paraboni, I. (2019). Computational personality recognition from Facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, *25*(4),268–287

Kaggle, https://www.kaggle.com/ (accessed: May. 27, 2022)

Ong, V., Rahmanto, A. D. S., Williem, W., Suhartono, D., Nugroho, A. E., Andangsari, E. W., & Suprayogi, M. N. (2017). Personality prediction based on Twitter information in Bahasa Indonesia. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, *11*, 367–372

R, V. (2021). Comparative Analysis for Personality Prediction by Digital Footprints in Social Media. *Journal of Information Technology andDigital World*, *3*(2), 77–91

Song, M. R., Chu, W., & Im, M. (2021). The effect of cultural and psychological characteristics on the purchase behavior and satisfaction of electric vehicles: A comparative study of US and China.
International Journal of Consumer Studies, *April 2020*, 1–20