Comparison between Naïve Bayes and certainty factor to predict big five personality

U Syaripudin1,*, R Zaenal1, M F A Duri1, E Firmansyah2 and A Rahman1

1 Department of Informatics, UIN Sunan Gunung Djati Bandung, Bandung, Indonesia
2 Department of Informatics, STMIK Sumedang, Sumedang, Indonesia

*undang_if@uinsgd.ac.id

Abstract. The Big Five is one of the methods known in the world of psychology to interpret persons personality, especially to find a relationship between personality and work environment. The Big Five consists of openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). This study aims to predict the level of agreeableness in the classification of The Big Five. The method used in detecting this level of agreeableness is using the Certainty Factor and Naive Bayes algorithms. The results obtained in a comparison of the level of accuracy in personality prediction, the Certainty Factor algorithm get results of 80% and Naive Bayes get 90%. From these results it can be concluded that the Naive Bayes algorithm is better at predicting personality.

1. Introduction

Personality is the whole sika p, expressions, feelings, temperaments, characteristics and also a person's behaviour. The feeling of expression and temperament will be manifested in one's actions if faced with certain situations. Every person has a tendency to behave in a standard / consistent manner continuously in facing the situation at hand, so that it becomes his personal characteristics. At this time personality has become one of the factors in accepting or not in some companies.

Identifying one's personality can be done in various ways, one of which is the Big 5. Personality Big 5 is one of the methods known in the world of psychology to interpret a person's personality, especially to find the relationship of personality to the work environment. Big 5 personality consists of openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N).

In today's social life where people are needed for young people to work together in their lives in learning, work and other things, agreeableness in the Big 5 classification is a good indicator of measuring cooperative levels and the ease of working together.

This study uses the question answer or questionnaire method to predict someone's personality, to predict a person's personality using the Naïve Bayes algorithm and certainty factor. Both algorithms are used to determine a person's personality and also determine the accuracy of the results of predictions and then compare the predicted values between the two algorithms.

This research will utilize Android technology. Android is an operating system whose development is very massive and fast. Currently, android has been synonymous a smartphone. The development of applications on Android is also very fast, even in 2010, every month there were more than 10 thousand applications added for Android. Naive Bayes Classifier is a method statistical classification that can be
used to predict the probability of membership of a class. Naïve Bayes Classifier (NBC) has proven to have high accuracy and speed when applied to databases with large data [1].

Naïve Bayes (NB) is a machine-learning method that has been used for over 50 years in biomedical informatics [2,3]. It is very efficient computationally and has often been shown to perform classification surprisingly well, even when compared to much more complex methods. However, NB is known to be miscalibrated and this problem is generally accentuated when there are large numbers of features; it tends to make predictions with posterior probabilities that are too close to 0 and 1 [2,4,5].

Naïve bayesian classifier based on Bayes’ theorem is employed to derive a probabilistic classification model from previously assessed attributes. It is most commonly used for antispam mail filtering, which trains the filter to automatically separate spam mail and legitimate messages in a binary manner [6]. Bayesian networks specify dependence properties between variables by using a directed acyclic graph. They describe probabilistic models useful for nondirected classification. That is, one can predict (and compute likelihoods for) one subset of variables from any other [7].

Certainty Factor (CF) is a classification method that uses the value of confidence or weight for the symptoms experienced by an object. This is to overcome the uncertainty in decision making [8]. There are two types of values in CF, namely [7]:

- Assessment of an expert who sets the standard for the assessment of the symptoms of the disease
- The value of the user to show the degree of confidence of the premise of symptoms of the disease. Usually done by users who have had previous experience.

In expert systems, the certainty factor model is used for reasoning under uncertainty. It is a useful model as an alternative to a probabilistic one when probabilistic and risk information are inadequate. Certainty factors based on experts’ estimate are intended to capture the notions of both probability and utility [9].

With the use of the Naïve Bayes method and certainty factor, it will be possible to predict one's personality. The prediction process starts with a question and answer where the user will answer the question by selecting the available answers. Based on the research that has been done before, it is known that a system is built to determine one's personality and by using only one algorithm in the process of determining it. The differentiator of this study from previous studies is from the use of two algorithms in the process of determining which will be compared by looking at the level of accuracy of the processes of the two algorithms. The Naïve Bayes method is a simple classifier type that applies Bayes theorems with strong (naïve) independent assumptions [8]. In connection with this, the research takes the title "Comparison of Analysis of Naïve Bayes and Certainty Factor Algorithms to Predict the Level of Agreeableness of Big 5-Based Android".

2. Methodology
The method that used in this study is Comparison of Analysis of Naïve Bayes and Certainty Factor Algorithms to Predict the Level of Agreeableness of Big 5-Based Android. As for making applications using Prototype software development method, because this method is suitable for systems or applications that are built following user needs, this method is very suitable to be applied in the design process of software to be built which focuses on approaches to aspects of design, function, and user-interface [10].

2.1. Naïve bayes
Naïve Bayes Classifier is a method statistical classification that can be used to predict the probability of membership of a class [1]. Naïve Bayes provides speed and high accuracy when applied into a large data [11]. Bayes classifications are statistical classifications that can predict the class of a member of probability. For simple Bayes classifications better known as naïve Bayesian Classifiers it can be assumed that the effect of an attribute value of a given class is free from other attributes.
P(Ci|X) = \frac{P(X|Ci)}{P(X)} \tag{1}

X : Data with unknown classes
x : The data hypothesis is a specific class
P (Ci | X) : Probability of hypothesis based on condition (posteriori probability )
P (Ci) : Probability of hypothesis (prior probability )
P (X | Ci) : Probability is based on the definition of the hypothesis
P (X) : Probability X

Naïve Bayes method categorizes data into specific categories based on the highest posterior probability P(Ci | X). Classification of data will occur if and only if the posterior probability data H based on condition X {P (H | X)} is smaller than posterior probability data H based on \{ P (H | X)\} \cite{11}.

2.2. Certainty factor
Certainty Factor (CF) is a method proposed by Shortliffe and Buchanan in 1975 to accommodate an expert's uncertainty and inexact reasoning. Certainty Factor (CF) is a method proposed by Shortliffe and Buchanan in 1975 to accommodate inexact reasoning by an expert. An expert (for example: a doctor) often analyses information with the phrase "maybe", "most likely", "almost certain". So that the existence of the Certainty Factor method can describe the level of confidence of an expert on the problem at hand. \cite{5}.

\[ CF [H,E]=MB [H,E] MD [H,E] \tag{CF1} \]

CF [H, E] : The certainty factor hypothesis that is influenced by evidence E is known with certainty.
MB [H, E] : measure of belief in hypothesis H, if given evidence E (between 0 and 1)
MD : measure of disbelief (value of unbelief)
P : probability
E : evidence (event / fact)

2.3. Architecture system

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{architecture_system.png}
\caption{Architecture system.}
\end{figure}

2.1. Unified Modelling Language (UML) diagram
The UML Diagram on this system is as follows:
2.1.1. Use case diagram. Use case diagram illustrates what can be done by the software to be developed.

![Use case diagram system]

**Figure 2.** Use case diagram system.

2.1.2. Class diagram. Class diagrams are descriptions of software with the relationships that are in it. Using class diagrams aims to describe the classes, functions, and attributes that exist in the software.

![Class diagram]

**Figure 3.** Class diagram.

2.1.3. Sequence diagram. Sequence diagram is an illustration of the interaction between objects and shows a series of message exchanges that occur on each object in the software.
3. Result and discussion
In testing the method, testing of each of the Certainty Factor and Naïve Bayes methods will be conducted using predetermined data and parameters and several trial scenarios to be carried out. After the trial scenario is done, there will be experimental results in each scenario, then the results will be analysed for conclusions. The parameters taken in the test are in the form of search results from the process predictions from each algorithm. A data is tested as many as 40 data.

Table 1. Algorithm test results.

| No | Algorithm          | Amount of Correct Results | Total Data Test |
|----|--------------------|---------------------------|-----------------|
| 1  | Naïve Bayes        | 40                        | 40              |
| 2  | Certainty Factor   | 38                        | 40              |

Based on testing the validity by comparing the actual data with the prediction results of the system, it is obtained for the results of predictions with the Certainty Factor algorithm of 40 test data 38 data right so that with the test data the accuracy rate obtained is 95%. Whereas for the Naïve Bayes algorithm from 40 test data as many as 40 prediction data right according to the existing test data so that the level of accuracy for the Naïve Bayes algorithm is 100%.

4. Conclusion
Through testing predicting the algorithmic personality naïve bayes and certainty factor can be implemented in knowing the level of agreeableness of a person.

In this case, personality predictions with the Naïve Bayes method and the Certainty Factor algorithm and with the same questions and input, the Naïve Bayes algorithm is superior to the Certainty Factor algorithm. Comparing the results of system predictions with the existing test data, it is found that the level of constellation for the Certainty Factor algorithm is 95%, while the Naïve Bayes algorithm has an accuracy rate of 100% of the 40-test data available. From the implementation of the experiments above we can make a conclusion that the implementation of naive bayes in this case study yields better value than certainty factor for its accuracy value. It is expected that in the next study use certainty factor by adding other variables, as well as naive bayes compared to other algorithms such as support vector machine (SVM).

Acknowledgments
We would like to say thank you for Research and Publication Centre of UIN Sunan Gunung Djati Bandung that give the full support and fund for this publication research.
References

[1] Zulfikar W B and Lukman N 2017 Perbandingan Naive Bayes Classifier Dengan Nearest Neighbor Untuk Identifikasi Penyakit Mata *J. Online Inform.* 1 82–86

[2] Wei W, Visweswaran S and Cooper G F 2011 The application of naive Bayes model averaging to predict Alzheimer’s disease from genome-wide data *J. Am. Med. Informatics Assoc.* 18 370–375

[3] Warner H R 2011 A Mathematical Approach to Medical Diagnosis *Jama* 177 177

[4] Bennett P N 2000 Assessing the Calibration of Naive Bayes’ Posterior Estimates *Cmu-Cs-00-155*

[5] Zadrozny B and Elkan C 2001 Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers *18th Int. Conf. Mach. Learn.* 609–616

[6] Jang W, Lee J K, Lee J and Han S H 2015 Naive Bayesian Classifier for Selecting Good/Bad Projects during the Early Stage of International Construction Bidding Decisions *Math. Probl. Eng*

[7] Buntine W 2014 Theory Refinement on Bayesian Networks *Uncertain. Proc.* 1991 52–60

[8] Syarief M, Prastiti N and Setiawan W 2017 Comparison of Naïve Bayes and Certainty Factor Method for Corn Disease Expert System : Case in Bangkalan, Indonesia *J. Eng. Res. Appl.* 7 30–34

[9] Fu L M and Shortliffe E H 2000 The application of certainty factors to neural computing for rule discovery *IEEE Trans. Neural Networks* 11 647–657

[10] Suryani D, Irfan M, Uriawan W and Zulfikar W B 2018 Implementasi Algoritma Divide And Conquer Pada Aplikasi Belajar Ilmu Tajwid *J. Online Inform.* 1 13

[11] Stiawan D, Sandra S, Alzahrani E and Budiarto R 2017 Comparative analysis of K-Means method and Naïve Bayes method for brute force attack visualization *2017 2nd Int. Conf. Anti-Cyber Crimes, ICACC 2017* 177–182