Bit-Tech (Binary Digital - Technology)

Vol.6, No.1, August 2023
Available online at: http://jurnal.kdi.or.id/index.php/bt

Implementation of Data Mining Classification of People's Personalities Using Naïve Bayes Algorithm

Sylvain Revata [1)*], Rino [2)]

[1)2)] *Informatics Department, Buddhi Dharma University*
Jl. Imam Bonjol No.41, Tangerang, Indonesi

[1)] sylvainasty@gmail.com

[2)] mr. rino85@gmail.com

Article history:

Received 05 May 2023;
Revised 06 May 2023;
Accepted 08 August 2023;
Available online 25 August 2023

Keywords:

Algorithm
Data Mining
Information
Naïve Bayes algorithm
Personality

Abstract

Personality or personality comes from the word Persona, a word that refers to a guise or mask. This persona has the meaning of how a person looks before others. In general, a person's personality can be seen how they appear and give an impression on the people around them. Therefore there are many personality types in a person, from these types it is difficult to guess in a person. Because it must be able to know the characteristics that exist in its personality types. So with that made a data mining application to predict a person's personality type. Currently the naïve bayes classification algorithm is one method that is quite effective to be used in predicting an opportunity in the future using previously available data. This is because the naïve bayes algorithm is a classification method that is easily understood by system designers and system users. The purpose of making this data mining application is to help people find out what personality type they have. Also information is a data that has been processed and processed into knowledge that can be used for decision making for the recipients. Algorithm is a computation that is an interesting technique that is communicated as a finite series. Calculation is also a collection of commands to solve a problem. This command can be interpreted piecemeal from start to finish and the problem can be anything, considering that each problem has a model for the introductory state that must be satisfied before running the computation.

I. Introduction

Prediction (forecasting) is a calculation to forecast future conditions through testing conditions in the past. Forecasting future sales means determining the estimated amount of sales volume, even determining the sales potential and market area controlled in the future. The purpose of this prediction is to help a company or organization in making decisions to determine the amount of goods that must be provided by the company. Then this prediction can help and provide the best output so that it is expected that the risk of errors caused by the company in planning can be minimized to a minimum. [1]

Data mining is the process of extracting useful information and patterns from very big data. *Data mining* includes data collection, data extraction, data analysis, and data statistics. In other words, *data mining* is a process of analyzing data patterns which are then used as useful information[2].

Personality or personality comes from the word Persona, a word that refers to a guise or mask. This charm has a meaning about how a person looks before others. In general, a person's personality can be seen how they appear and give an impression on the people around them. Then knowing the personality traits can help us in understanding what our strengths and weaknesses are. And in the world of work, personality characteristics are very necessary because if something goes wrong in choosing / selecting people in a job can lead to disappointing results.

[*] Corresponding author

ISSN 2622-2728 (online) 2622-271X (print) © 2018 The Authors. Published by Komunitas Dosen Indonesia.
This is an open access article under the CC BY-SA 4.0 license (http://creativecommons.org/licenses/by/4.0/)
doi: 10.32877/bt.v6i1.741

II. Related Works/Literature Review (Optional)

According to [3] the Naive Bayes algorithm, it is a simple probability-based prediction technique based on the application of Bayes' theorem (or Bayes' rule) assuming strong (naïve) independence of features. Also [4] The Naive Bayes algorithm is a classification algorithm that can be used for text documents. This classification algorithm uses the Bayes theorem which exists in probability theory by assuming all attributes are unrelated to each other.

In Naive Bayes, strong feature independence means that one feature in the data does not depend on the presence or absence of other features in the same data. For example, if the fruit is red, round, and about 10 centimeters in diameter, it can be considered an apple. Naïve Bayes classifiers consider each of these features to contribute independently to the likelihood that this fruit is an apple, regardless of the possible correlation between features of color, roundness, and diameter.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} = P(X|H) \times P(H) \div P(X) \qquad (1)$$

X = Data with unknown class
H = Data hypothesis X which is a more specific class
P(H|X) = Probability hypothesis H based on XXX condition (posteriori probability)
P(H) = Probability of Hypothesis H (prior probability)
P(X}H) = Probability X based on conditions on hypothesis H
P(X) = Probability X

According to [5] Classification is a type of data analysis that can help people determine the label class of the sample they want to classify. Classification is a supervised learning method, a method that tries to find the relationship between input attributes and target attributes. The purpose of classification is to increase the reliability of the results obtained from the data.

According to [6] "Character is a strong psycho and actual association of each and every individual that determines the extraordinary variation for his current situation."

As a general rule, character implies how an individual appears and attracts others and the character for each individual is unique and has its own uniqueness.

Each individual has an alternate type of character. Some have a gentle, happy, and kind character. There are also people who have different characters such as simple, difficult, and others.

According to[7] Algorithm is a logical sequence of problem solving steps that are systematically arranged. The flow of thought in completing a job as outlined in writing. What is emphasized first is the flow of thoughts, so that one person's algorithm can also be different from another person's algorithm. While the second emphasis is written, which means it can be in the form of certain sentences, pictures or tables.

According to [8]An application can be interpreted as a program in the form of software that runs on a certain system that is useful for helping various activities carried out by humans.

According to [9] "Information mining is an iterative and intelligent cycle of finding new, powerful, useful, and justifiable examples or models in massive data sets." Information mining consists of searching for desired patterns or examples in large data sets to help future leaders, these examples are perceived by specific devices that can provide valuable and intelligent examination of information that can then be concentrated more fully, which may use the help of other options apparatus.

According to [10] "Data is a reality that describes an event and is a rough structure that cannot be told much so it must also be handled through a model to convey data. Data is something that is not important to the recipient even though it requires handling. Data can be states, images, sounds, letters, numbers, mathematics, language, or other images that can be used as material to see climate, objects, events, or ideas.

According to [11] "A database pool is a capacity framework that stores a wide variety of organized data so that it is not difficult to access."

According to [12] website means a collection of pages containing data stored on the web that can be accessed or viewed via the web on gadgets that can actually access the web, such as computers.

According to [13] A computer is a device that has many uses suitable for customization, getting info or information (reality and images) and handling information and controlling it so that it becomes useful data.

[14]Information is data that has been processed and transformed into things that are easier to understand and useful for the recipient. Then according to [15] Information is data that is processed to be more useful and meaningful to the recipient, as well as to reduce uncertainty in the decision-making process about a situation. And it can be concluded that information is data that has been processed and processed into knowledge that can be used for decision making for recipients.

According to [16] UML is a graphic/image based language for visualizing, specifying, building and documenting an OO (Object-Oriented) based software development system. According to [17]

According to [18] Data retention is one way to improve database performance and even application performance, namely by forming a system that separates active and inactive data by the database administrator (dba) periodically transferring transaction data that has not been carried out change again to another database server and delete the data.

III. Methods

According to "Computation is an interesting technique that is communicated as a finite circuit. A calculation is also a collection of commands to solve a problem. This order can be interpreted bit by bit from beginning to end." The problem can be anything, given that each problem has a model for introductory states that must be met before running the calculation. The calculation also has a redundancy cycle (emphasis), and then has a choice until the choice is complete.[19]

[20]The Naïve Bayes Classifier algorithm is one of the algorithms that says that whether or not there are certain characteristics of one class there is no other relationship with the characteristics of other classes. Here is the general formula used:

$$P(X|Y) = \frac{P(X|Y)*(P(Y)}{P(X)} \qquad (2)$$

Then the formulation of Naïve Bayes Classifier to perform classification is as follows:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \qquad (3)$$

A Gaussian distribution is chosen to represent an opportunity class that has conditions for the continue attribute. This distribution has the characteristics of two parameters, namely the mean μ and the α2 variant for each class Y, while the conditional probability for attribute X is contained in the following formula:

$$P(= X_i|Y = Y_i) = \frac{1}{\sqrt{2\pi\sigma ij}} exp \frac{(x_i-\mu ij)^2}{2\sigma_{ij}^2} \qquad (4)$$

Where:
MIJ : mean sample Xi (x) from the overall training data
$2\sigma_{ij}^2$: sample (s^2) from the training data

IV. Results

The dataset to be used has as many as 20 records, then from these data have been grouped into 4 classes of personality characteristics, namely Sanguine, Choleric, Melancholy, and Plegmatic. The 20 data records are divided into 6 data records in the Sanguine class, 5 data records in the Koleris class, 4 data records in the Plegmatic class, and 5 data records in the melancholy class. Then from the existing training data, the probability value of the class is calculated as follows:

$$P(Sanguin) = \sum \frac{Sanguin}{Jumlah\ Data} = \frac{6}{20} = 0{,}3 \ (4)$$

$$P(Koleris) = \sum \frac{Koleris}{Jumlah\ Data} = \frac{5}{20} = 0{,}25 \ (5)$$

$$P(Koleris) = \sum \frac{Koleris}{Jumlah\ Data} = \frac{5}{20} = 0{,}25 \ (6)$$

$$P(Melankolis) = \sum \frac{Melankolis}{Jumlah\ Data} = \frac{5}{20} = 0{,}25 \ (7)$$

TABLE 1
Number Of Gender Attributes

	Class				
Gender	**Sanguine**	**Choleric**	**Plegmatis**	**Melancholy**	**Total**
L	3	2	1	4	10
P	3	3	3	1	10
Total	6	5	4	5	20

TABLE 2
Number of Answer-Answer Attributes of each class

	Class				
Answer	**Sanguine**	**Choleric**	**Plegmatis**	**Melancholy**	**Total**
A	88	47	31	42	208
B	49	83	27	46	205
C	52	33	28	67	180
D	51	37	74	45	207
Total	240	200	160	200	800

Then calculated the probability of each sex with its classes, is as follows:

TABLE 3
Probability of each Attribute with its Class

	Sanguine	Choleric	Plegmatic	Melancholy
P(Gender = L)	$\frac{3}{6} = 0,5$	$2/5 = 0,4$	$1 / 4 = 0,25$	$4/5 = 0,8$
P(Gender = P)	$\frac{3}{6} = 0,5$	$3/5 = 0,6$	$¾ = 0,75$	$1/5 = 0,2$
P(Answer A)	$88/6 = 14,6667$	$47/5 = 9,4$	$31/4 = 7,75$	$42/5 = 8,4$
P(Answer B)	$49/6 = 8,16667$	$83/5 = 16,6$	$27/4 = 6,75$	$46/5 = 9,2$
P(Answer C)	$52/6 = 8,66667$	$33/5 = 6,6$	$28/4 = 7$	$67/5 = 13,4$
P(Answer D)	$51/6 = 8,5$	$37/5 = 7,4$	$74//4 = 18,5$	$45/5 = 9$

After that calculate the standard deviation value on all features, here is a table of the results of the standard deviation value:

TABLE 4
Standard Deviation Value

	Sanguine	Choleric	Plegmatic	Melancholy
S(Answer A)	$47,3335\backslash 5 = 9,4667$ $\sqrt{9,4667} = 3,0768$	$43,2\backslash 4 = 10,8000$ $\sqrt{10,8000} = 3,286335$	$32,7501\backslash 3 = 10,9167$ $\sqrt{10,9167} = 3,304043$	$29,2\backslash 4 = 7,3$ $\sqrt{7,3} = 2,70185$
S(Answer B)	$16,8335\backslash 5 = 3,3667$ $\sqrt{3,3667} = 1,8349$	$75,2\backslash 4 = 18,8000$ $\sqrt{18,8000} = 4,335897$	$32,7501\backslash 3 = 10,9167$ $\sqrt{10,9167} = 3,304043$	$26,8\backslash 4 = 6,7$ $\sqrt{6,7} = 2,588436$
S(Answer C)	$4,2667\backslash 5 = 21,3335$ $\sqrt{21,3335} = 2,065599$	$43,2\backslash 4 = 10,8000$ $\sqrt{10,8000} = 3,286335$	$26,0001\backslash 3 = 8,6667$ $\sqrt{8,6667} = 2,943926$	$5,2\backslash 4 = 1,3$ $\sqrt{1,3} = 1,140175$
S(Answer D)	$25,5\backslash 5 = 5,1$ $\sqrt{5,1} = 2,25832$	$33,2\backslash 4 = 8,3$ $\sqrt{8,3} = 2,88097$	$51\backslash 3 = 17$ $\sqrt{17} = 4,123106$	$40\backslash 4 = 10$ $\sqrt{10} = 3,162278$

The calculation of standard deviation values is only done in answer A, answer B, answer C, and answer D. Because each data in the feature has a value (numeric). While the gender attribute has no value, therefore the gender attribute cannot be calculated standard deviation value. Next is the calculation process with existing test data, there are 4 test data samples, the following table of test data will be carried out for calculation:

TABLE 5
Test Data

No	Name	Gender	Answer A	Answer B	Answer C	Answer D	Original Class
1	Kipli Scissors	L	10	10	10	10	Plegmatis
2	Udin Idin	L	18	5	7	10	Blood
3	Sacred Minaso	P	2	12	16	10	Melancholy
4	Ida Pidang	P	5	10	20	5	Choleric

In the calculation of this test data, first of all is to calculate the feature values of answers A, B, C, D. in this example the author took a sample from the first test data, namely with the name Kipli Scissors, Gender L (male), Answer A 10, Answer B 10, Answer C 10, Answer D 10 with the original class Plegmatis. After that the calculation was carried out with the formula of Naïve Bayes Hippocrates above:

TABLE 6
Calculation of Answer Features of Each Class

	Blood	Choleric	Plegmatis	Melancholy
Answer A	0,11485	0,12123781	0,11823611	0,14418736
Answer B	0,18751	0,0865324	0,115539	0,15306888
Answer C	0,18398	0,11555459	0,12766574	0,01144748
Answer D	0,16921	0,1318782	0,08541004	0,12555925

After calculating the feature value, the next step is to calculate the final probability value in each test data in table 6 above:

a) *Sanguine Class*

$P(X|Sanguin) = P(Sanguin) * \ P(Gender = L \mid Sanguin) * P(Answer\ A = 10 \mid Sanguin) * P(Answer\ B = 10 \ |Sanguin) * P(Answer\ C = 10 \mid \ Sanguin) * P(Answer\ D = 10 \mid Sanguin)(8)$
$= 0{,}3 * 0{,}5 * 0{,}11485 * 0{,}18751 * 0{,}18398 * 0{,}16921 = 0{,}00010057(9)$

b) *Choleric Class*

$P(X|Koleris) = P(Koleris) * \ P(Gender = L \mid Koleris) * P(Answer\ A = 10 \mid Koleris) * P(Answer\ B = 10 \ |Koleris) * P(Answer\ C = 10 \mid Koleris) * P(Answer\ D = 10 \ |Koleris)(10)$
$= 0{,}25 * 0{,}4 * 0{,}12123781 * 0{,}0865324 * 0{,}11555459 * 0{,}1318782 = 0{,}00001599(11)$

c) *Plegmatic Class*

$P(X|Plegmatis) = P(Plegmatis) * \ P(Gender = L \mid Plegmatis) * P(Answer\ A = 10 \mid Plegmatis) * P(Answer\ B = 10 \mid Plegmatis) * P(Answer\ C = 10 \mid Plegmatis) * P(Answer\ D = 10 \ |Plegmatis)(12)$
$= 0{,}2 * 0{,}25 * 0{,}11823611 * 0{,}115539 * 0{,}12766574 * 0{,}08541004 = 0{,}00000745(13)$

d) *Melancholic Class*

$P(X|Melankolis) = P(Melankolis) * \ P(Gender = L \mid Melankolis) * P(Answer\ A = 10 \mid Melankolis) * P(Answer\ B = 10 \ |Melankolis) * P(Answer\ C = 10 \mid Melankolis) * P(Answer\ D = 10 \ |Melankolis)(14)$
$= 0{,}25 * 0{,}8 * 0{,}14418736 * 0{,}15306888 * 0{,}01144748 * 0{,}12555925 = 0{,}00000634(15)$

After that continue the calculation with all existing test data, then from the results that have been obtained calculate the accuracy and error rate, as follows:

TABLE 7
Calculation of All Test Data

No	Name	Sanguine Value	Choleric Value	Plegmatic Value	Melancholic Value	Original Class	Results Class	Information
1	Kipli Scissors	0,00010057	0,00001599	0,00000745	0,00000634	Plegmatis	Blood	**Wrong**
2	Udin Idin	0,00007739	0,00001077	0,00000000	0,00000536	Blood	Blood	**True**
3	Sacred Minaso	0,00000675	0,00000940	0,00000382	0,00001690	Melancholy	Melancholy	**True**
4	Ida Pidang	0,00000172	0,00000722	0,00000210	0,00000000	Choleric	Choleric	**True**

From the results above, how to calculate accuracy is by the amount of data with the correct information divided by the amount of data that exists then multiplied by 100, for example as follows:

$Accuracy = (correct\ total \div total\ data) * 100\ (16)$
$Accuracy = (3 \div 4) * 100 = 75\ (17)$

After that, the opposite is by calculating the error rate, namely by the number of incorrect data information divided by the number of data multiplied by 100, which is as follows:

$Error\ Rate = (wrong\ total \div total\ data) * 100\ (18)$
$Error\ Rate = (1 \div 4) * 100 = 25\ (19)$

That way the accuracy and error rate results obtained with the 4 test data above are accuracy with 75% and error rate with 25%.

II. DISCUSSION

TABLE 1
BlackBox Testing

No	Test Scene	Scenario	Expected results	Test Results
1	Login Page	User and admin input Username and Password and press the login button to login	If the Username and Password are correct, it will enter the user or admin	Valid

			dashboard page depending on the username and password used	
2	Login Page	User and admin input Username and Password and press the login button to login	If the Username or Password is incorrect, a failed login error message will appear	Valid
3	Home User Page	Users enter the home page and press the classification button to answer questions or log out to exit the application	Users can enter the question page from home and can also log out of the application	Valid
4	Question Page	The user will be asked questions and will be required to answer everything	Users can answer questions and then the results will come out based on the answers received	Valid
5	Question Page	The user will be asked questions and will be required to answer everything	Users skip multiple questions or haven't answered any of them and an error message says there are unanswered questions	Valid
6	Admin Home Page	The admin enters the admin home page and is given various menus intended for admins	Admin can access all menus that have been given	Valid
7	User Data Page	The admin fills in the data and presses the save button and presses the delete all data button	By pressing the save button the admin can save user data and the user will be able to access this application, if the admin presses the delete all data button then the user data will disappear from the application and database	Valid
8	Question Data Page	Admin presses the upload question button and delete the question data	By pressing the upload question button, the admin will upload a new question and if the delete button is pressed, the existing question data will be deleted	Valid
9	Accuracy Test Page	Admin presses clear data and Accuracy Test button	By pressing the delete button, the existing test data will be deleted and if the accuracy test button is pressed, the existing test data will be calculated for accuracy	Valid

10	Results Page	Admins can check the results	Here the admin can see the results of the user's answer	Valid

III. Conclusions

Based on the results of design, manufacture and testing using the Naïve Bayes algorithm, it can be concluded that this application can help someone to find out the personality owned by him, then this application has an accurate rate of around 75%-85% in classifying and determining a person's personality, and by using the Naïve Bayes algorithm is able to classify well and accurately.

References

[1] O. Eriyanto, ‘Analisis Peramalan Penjualan Handphone Blackberry Pada PT. Selular Shop Mall’, 2012.

[2] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*. Penerbit Andi, 2020.

[3] M. M. Huda and R. D. R. Yusron, ‘Kombinasi Naive Bayes dan Metode Time Series Sebagai Peramalan Pergerakan Harga pada Perdagangan Valuta Asing’, *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 2, no. 2, pp. 151–155, Aug. 2020, doi: 10.28926/ilkomnika.v2i2.186.

[4] A. Dwi Herlambang and S. Hadi Wijoyo, ‘ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI SUMBER BELAJAR BERBASIS TEKS PADA MATA PELAJARAN PRODUKTIF DI SMK RUMPUN TEKNOLOGI INFORMASI DAN KOMUNIKASI’, vol. 6, no. 4, pp. 431–436, 2019, doi: 10.25126/jtiik.201961323.

[5] S. Hendrian, ‘Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan’, *Faktor Exacta*, vol. 11, no. 3, Oct. 2018, doi: 10.30998/faktorexacta.v11i3.2777.

[6] B. A. Karim, ‘Teori Kepribadian dan Perbedaan Individu’, *Education and Learning Journal*, vol. 1, no. 1, pp. 40–49, 2020.

[7] R. Novianti and R. A. Krisdiawan, ‘IMPLEMENTASI ALGORITMA FLOYD WARSHALL PADA APLIKASI PENGADUAN MASYARAKAT BERBASIS ANDROID’, vol. 13, no. 1, 2019, [Online]. Available: https://journal.uniku.ac.id/index.php/ilkom

[8] B. Huda and B. Priyatna, ‘Penggunaan Aplikasi Content Manajement System (CMS) Untuk’, 2019.

[9] E. D. Sikumbang, ‘Penerapan Data Mining Penjualan SepatuMenggunakan Metode Algoritma Apriori’, *Jurnal Teknik Komputer*, vol. 4, no. 1, pp. 156–161, 2018.

[10] Nawassyarif, M. Julkarnain, and K. Ananda, ‘SISTEM INFORMASI PENGOLAHAN DATA TERNAK UNIT PELAKSANA TEKNIS PRODUKSI DAN KESEHATAN HEWAN BERBASIS WEB’, *Jurnal Informatika Teknologi dan Sains*, vol. 2, no. 1, pp. 32–39, Feb. 2020.

[11] D. Sitinjak, Maman. D, and J. Suwita, ‘ANALISA DAN PERANCANGAN SISTEM INFORMASI ADMINISTRASI KURSUS BAHASA INGGRIS PADA INTENSIVE ENGLISH COURSE DI CILEDUG TANGERANG’, *JURNAL IPSIKOM*, vol. 8, no. 1, pp. 1–10, 2020.

[12] M. Destiningrum and Q. J. Adrian, ‘SISTEM INFORMASI PENJADWALAN DOKTER BERBASSIS WEBDENGAN MENGGUNAKAN FRAMEWORK CODEIGNITER’, *Jurnal Teknoinfo*, vol. 11, no. 2, pp. 30–37, 2017.

[13] Rada, ‘Pengertian Komputer’, Mar. 20, 2022.

[14] J. Hutahean, *Konsep sistem informasi*. Deepublish, 2015.

[15] E. Anggraeni, *Pengantar Sistem Informasi*. Penerbit Andi, 2017.

[16] R. Maharani, M. Aman, and J. Sistem Informasi Akuntansi STMIK INSAN PEMBANGUNAN Jl Raya Serang Km, ‘SISTEM INFORMASI NILAI SISWA BERRBASIS WEB PADA SMA NEGERI 19 KAB. TANGERANG’, vol. 5, no. DESEMBER, 2017.

[17] F. C. Ningrum, D. Suherman, S. Aryanti, H. A. Prasetya, and A. Saifudin, ‘Pengujian Black Box pada Aplikasi Sistem Seleksi Sales Terbaik Menggunakan Teknik Equivalence Partitions’, vol. 4, no. 4, 2019, [Online]. Available: http://openjournal.unpam.ac.id/index.php/informatika

[18] N. Eyni Alfia and B. Waseso, ‘Perancangan Aplikasi Retensi Data Pada Database MySQL (Studi Kasus: PT. Telkomsigma)’, 2020. [Online]. Available: https://jurnal.ikhafi.or.id/index.php/jusibi/364

[19] G. G. Maulana, ‘PEMBELAJARAN DASAR ALGORITMA DAN PEMROGRAMAN MENGGUNAKAN EL-GORITMA BERBASIS WEB’, 2017.

[20] KANTINIT, ‘Belajar Naive Bayes: Alur Algoritma, Rumus dan Contoh Perhitungan Naive Bayes’, *KANTIN IT*, Dec. 14, 2022.