Paradigma, Vol. 25, No. 1, Maret 2023, E-ISSN: 2579-3500, Page. 1-5
Published: LPPM Universitas Bina Sarana Informatika

PARADIGMA

website: http://jurnal.bsi.ac.id/index.php/paradigma/

Application of Naïve Bayes for Classification of Criteria for Potable Water with the CRISP-DM Method

Ibnu Alfitra Salam[1], Katon Wahyudi Putra[2], Sisca Yuliatina[3], Betha Nurina Sari[4]

[1,2,3,4]Universitas Singaperbangsa Karawang, Indonesia

ARTICLE INFORMATION

Article History:
Received: January 13, 2023
Revised: January 28, 2023
Accepted: February 11, 2023

Keyword:
Water Quality
Naïve Bayes
CRISP-DM
Rapidminer
Google Collab

ABSTRACT

With water, living things can do various things easily. The adequacy of water is also important in maintaining human health. Water can be said to be feasible if its content is in accordance with the feasible criteria. From the dataset obtained regarding the feasibility of water for this study, it will calculate the accuracy value obtained using the Naive Bayes algorithm. To simplify the process of processing research data this time using the CRISP-DM methodology which is a stage for data mining. The study uses two tools, namely Rapidminer and Google Collab to compare their accuracy values. By using the two tools in implementing the Naive Bayes algorithm on a potable water quality dataset, an accuracy of 62.8% is obtained. This value is accurate enough to predict the quality of drinking water.

Corresponding Author:

Ibnu Alfitra Salam
Informatics Engineering, Faculty of Computer Science, Universitas Singaperbangsa Karawang Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat 41361, Indonesia
Email: 1910631170085@student.unsika.ac.id

INTRODUCTION

Water is a necessity that cannot be separated from everyday life. In guaranteeing potable water, the water must be clean and meet the indicators that are truly feasible (Irnawan et al., 2021). In accordance with the target of the sixth Sustainable Development Goals (SDGs), namely ensuring the availability and sustainable management of clean water and sanitation for all. It is hoped that by 2030, achieve universal and equitable access to safe and affordable drinking water for all (European Union, 2017).

The research on the quality of drinking water will be carried out using the CRISP-DM methodology and using the water potability dataset from the Kaggle dataset. This dataset will be tested using the Naïve Bayes algorithm where the accuracy of the Naïve Bayes algorithm is higher than the KNN algorithm (Saragi et al., 2022). The dataset used has a missing value, so it is necessary to apply pre-processing procedures. Application of missing data imputation procedures so that anomaly, bias, and noise data can be minimized and produce high values in the data processing process (Tempola et al., 2018). In the data preparation stage in the CRISP-DM methodology, a transformation will be carried out on the missing value, the transformation that will be used is interpolation which will create a new data within the range of existing data with the help of Google Collab (Primajaya et al., 2020).

To determine the classification criteria for potable water, there are many classification methods that can be used, one of which is the Naïve Bayes algorithm. In a previous study conducted by Yunita Sartika Sari (2021) regarding the application of the Naïve Bayes algorithm to determine water quality in the Jakarta area, the results showed a fairly high accuracy of 50.6% (Nurlia et al., 2021). Elin Nurlia (2021) conducted research on the application of Naïve Bayes to classify the level of dental diagnosis ratios at the Cingambul Health Center UPTD. The results of the tests performed show that the best performance is in 3-fold cross-validation by obtaining an accuracy value of 93.33% and an AUC value of 0.855 which is included in the Very Good category.

DOI: https://doi.org/10.31294/paradigma.v25i1.1754

In addition, the average of the nine test scenarios also obtained an accuracy value of 92.32% and an AUC value of 0.833 which is included in the Very Good category (Sari, 2021). A comparative study of the FUZZY C-MEANS algorithm and the Naïve Bayes algorithm in determining beneficiary families (KPM) based on the lowest socioeconomic status (SEE) by Putu, Gede, and I Gede Aris (2021) resulted in the fact that the Naïve Bayes accuracy value was obtained at 74% higher than the Fuzzy C-means accuracy value of 67%. This shows that by using the calculation of the confusion matrix, the results of an effective algorithm used in determining the beneficiary family are the Naïve Bayes algorithm (Saputra et al., 2021).

Based on the presentation of the results of previous studies regarding the performance of classification algorithms, it can be concluded that the Naïve Bayes algorithm has good data classification performance. This is based on the high accuracy value obtained using the Naïve Bayes algorithm. Therefore, researchers will apply the Naïve Bayes algorithm in classifying the criteria for potable water. Based on the description of this background, the title taken is "Application of Naïve Bayes for Classification of Criteria for Potable Water with the CRISP-DM Method ".

Water

Drinking water is water that goes through a processing process or without a processing process that meets health requirements and can be drunk directly. Quality drinking water is drinking water that meets biological, physical and chemical requirements. There are parameters used to determine the feasibility of drinking water consisting of maximum levels of physico-chemical parameters, and instrumentation for the amount of water-soluble solids (Total Dissolved Solids/TDS), manganese, iron, degree of acidity (pH), hardness, sulfate, chloride and nitrite (Sari, 2021).

Data Mining

Data Mining is a process of looking for patterns or information in selected data sets using certain techniques or methods. Data mining is divided into 5 parts according to their main roles, namely estimation, prediction, classification, clustering, and association (Hasanah et al., 2021). Data mining is a term used to find hidden knowledge in databases. Data mining is a semi-automated process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify potentially useful knowledge and information stored in large databases (Hasanah et al., 2021).

Classification

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts that aim to be used to predict the class of objects whose class labels are unknown. Classification algorithms that are widely used, namely decision/classification trees, bayesian classifiers/naïve Bayes classifiers, neural networks, statistical analysis, genetic algorithms, rough sets, k-nearest neighbors, rule based methods, memory based reasoning, and support vector machines (Khakim, 2022).

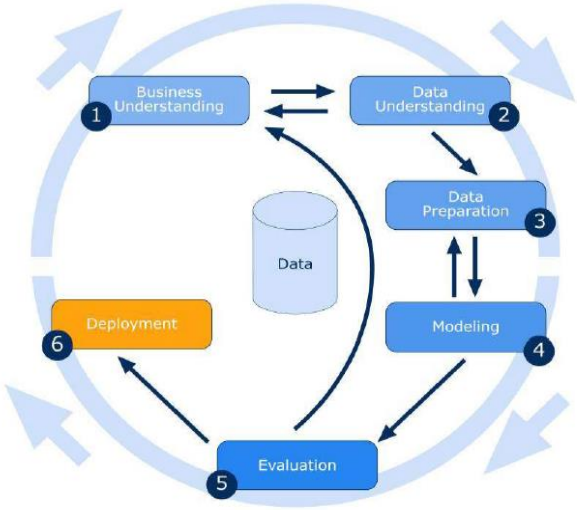
Naïve Bayes

Naïve Bayes Classifier is a classification method that is rooted in Bayes' theorem. The classification method uses probability and statistical methods put forward by the British scientist Thomas Bayes, namely predicting future opportunities based on previous experience so that it is known as Bayes' Theorem. The main feature of this Naïve Bayes Classifier is the very strong (naïve) assumption of the independence of each condition/event (Nurlia et al., 2021).

CRISP-DM

The CRISP-DM (Cross Industry Standard Process for Data Mining) approach is one of many procedural models for standardizing the data mining process, where the existing data will pass through each structured and clearly defined and efficient phase. This methodology consists of six stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. (Hasanah et al., 2021).

RESEARCH METHOD



Source: (Hasanah, at. al. 2021)

Figure 1 Stages of CRISP-DM

This research was conducted using a dataset from Kaggle with owner Aditya Kadiwal in 2020. The methodology used in this research is the Cross-Industry Standard Process for Data Mining Model (CRISP-DM). CRISP-DM was developed by an analysis of several industries conducted in 1996. CRISP-DM provides a standardized Data Mining process as a common problem-solving strategy for business research (Alfiah, 2021). This methodology consists of six stages, namely Business Understanding, Data Understanding,

Data Preparation, Modeling, Evaluation, and Deployment. This methodological process

consists of 6 stages which can be explained as follows (Hasanah et al., 2021). (1) Business Understanding, understanding the needs and goals from a business point of view then translates knowledge into the form of defining problems in data mining and then determines plans and strategies to achieve data mining goals. (2) Data Understanding, data collection is carried out, studying the data to be able to understand the data that will be used in research, identifying problems related to the data. (3) Data Preparation, namely repairing datasets that experience noise and missing data to facilitate data processing. (4) Modelling, determines the data mining techniques used, determines data mining tools, data mining algorithms, and determines parameters with optimal values. (5) Evaluation, is the interpretation phase of the data mining results shown in the modeling process contained in the previous phase. Evaluation is carried out in depth with the aim of adjusting the model obtained to match the objectives to be achieved in the first phase. (6) Deployment, a report or presentation is prepared from the knowledge obtained from the evaluation of the data mining process.

Stage 1 Business Understanding

The application of data mining in this study refers to data on criteria for drinking water to gain knowledge about a pattern of clean water quality standards that are suitable for drinking which has the potential to be an indicator of human health. This is based on the parameters that affect the standard quality of clean water that is suitable for drinking.

Stage 2 Data Understanding

In this study using dataset from Kaggle. The data contains water quality metrics for 3,276 different bodies of water. The parameters used are like 10 attributes, namely ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity, and Potability (as labels). Verification of the classification criteria for potable water which has the potential to be an indicator of the level of human health is categorized by the nature of the water that is potable and the nature of the water that is not potable. This indicates whether the water is safe for human consumption where 1 means drinkable and 0 means undrinkable.

Stage 3 Data Preparation

In this phase there are three processes, namely: (1) Data selection is selecting data to be used in the data mining process. In this process, the selection of attributes that are adapted to the data mining process is also carried out. (2) Data preprocessing, namely ensuring the quality of the data that has been selected at the data selection stage, at this stage the problem that must be faced is if there is noisy data and missing values. The process of cleaning data or cleansing is carried out to find data anomalies that may still be present in the data. (3) Transformation is grouping the attributes or fields that have been selected into a new database for data mining materials.

Stage 4 Modeling

In this stage, the process of modeling the selected classification technique is carried out to produce information patterns that can facilitate interested parties. The classification pattern produced by this data mining technique is used to predict the criteria for potable water with indicators that influence it. The tools used for the data mining exploration process are Rapidminer and Google Collaboratory assistance to run Python code. This step displays and informs about the performance of the classification method using Naïve Bayes. This study also uses a classification technique with the selected algorithm, namely Naïve Bayes. This is because the Naïve Bayes method only requires a small amount of training data (Training Data) to determine the parameter estimates needed in the classification process, and this is an advantage of using the Naïve Bayes method. Most situations in the real world of Naive Bayes often work out much better than expected (Alfiah, 2021).

Stage 5 Evaluation

At this stage there are two processes, namely: (1) Evaluate results, namely summarizing the results of the assessment in terms of business success criteria, including a final statement regarding whether the research has fulfilled the business objectives. (2) Determine the next step, namely giving a decision whether the modeling technique used can be used as a standard in determining research objectives.

Stage 6 Deployment

At this stage there are two processes, namely: (1) Deployment plan, which describes an overview of the plans for making the report that will be made. (2) Produce final report, which provides a visualization of the reports that have been made based on the deployment plan.

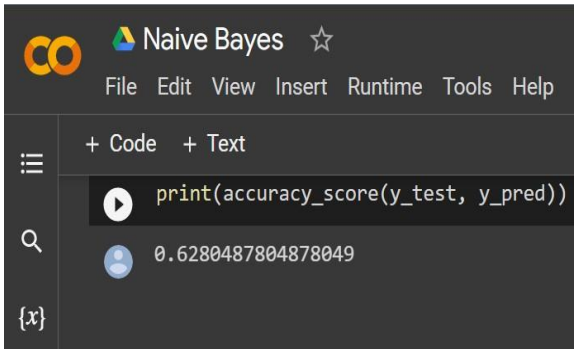
RESULTS AND DISCUSSION

The data used consists of 3,276 data lines consisting of 10 attributes, namely pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity, and Potability (as labels). In this data there are missing data for the attributes of pH, Sulfate, and Trihalomethanes. To deal with the absence of errors or constraints at the modeling stage, the data goes through a transformation stage by imputing the average value of each attribute that has missing data. The transformation process is assisted by Google Collab to make it easier in the process of imputing the average value of the missing data.

Table water quality dataset parameters

Parameter	Description
Ph	PH is an indicator of the acid or alkaline status of water. The optimum pH required is around 6.5-8.
Hardness	Hardness is the level of water's ability to precipitate soap due to the presence of valence

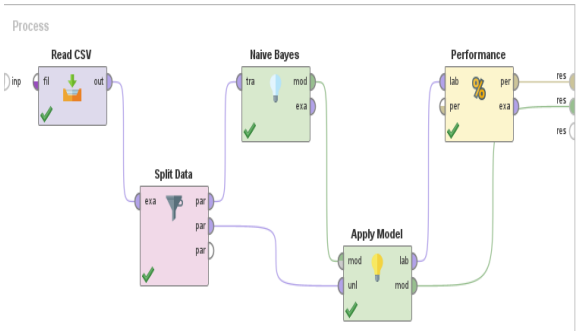
	metal ions such as calcium and magnesium.
Solids	Solid describes inorganic salts and the amount of organic matter that is in aqueous solution.
Chloramines	Chlorine and chloramine are the main disinfectants used in public water systems. Chloramines are formed when ammonia is added to chlorine to treat drinking water
Sulfate	Sulfates are found in minerals, soil, rocks. Sulfate can cause a high concentration of concentrated taste and can cause a laxative effect in consumers who are not used to it.
Conductivity	Concentration of the number of positively charged cations or ions and negatively charged anions or ions in the water. Increasing ion concentration increases the electrical conductivity in water.
Organic_carbon	Organic carbon comes from organic materials or compounds that decompose in water
Trihalomethanes	Trihalomethanes (THMs) are formed in drinking water mainly as a result of chlorination of organic matter
Turbidity	The degree of turbidity in the water
Potability	The degree of turbidity in the water



Source: (Salam, at. al. 2023)

Figure 2. Accuracy results with Google Colab

The tests carried out will use Rapidminer and Google Collab in order to be able to compare the higher accuracy results between the two software. The test was carried out by sharing the same data between Rapidminer and Google Collab, namely with 80% training data and 20% test data so that the data processed is no different.



Source: (Salam, at. al. 2023)

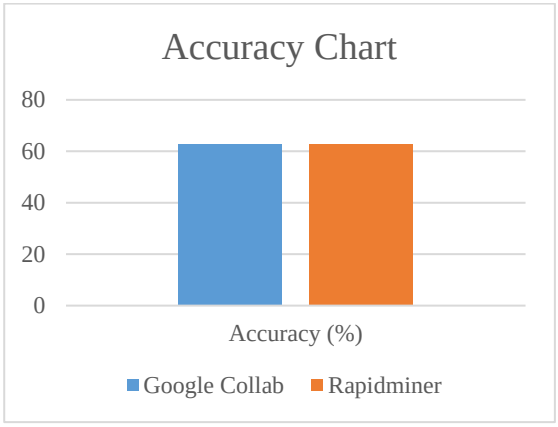
Figure 3. The process of finding Naïve Bayes accuracy using Rapidminer

accuracy: 62.80%

	true Tidak Layak	true Layak	class precision
pred. Tidak Layak	357	201	63.98%
pred. Layak	43	55	56.12%
class recall	89.25%	21.48%	

Source: (Salam, at. al. 2023)

Figure 4. Accuracy results with Rapidminer



Source: (Salam, at. al. 2023)

Figure 5. Accuracy results with two tools

Based on the two tests that have been carried out, the accuracy results are the same as Google Collab and Rapidminer, which is 62.8%. From these results it can be concluded that the dataset used as research material has an accuracy of approximately 63%. With the accuracy figures that have been obtained, it can be said that the dataset used is accurate enough to predict water feasibility. For the dataset itself, the predictions produced are more criteria that are not suitable for drinking, which can be seen in Figure 4. Unfit predictions are 63.98%, while Feasible predictions are 56.12%.

CONCLUSION

Based on the research that has been done regarding the application of the Naïve Bayes Classification algorithm to the criteria for drinkable water using Rapidminer and Google Collab. The test uses transformation data imputation process average value for missing data. The results of the same

accuracy were obtained using Google Collab and Rapidminer, namely 62.8% with a lot of 80% training data and 20% test data. It can be concluded that the Naïve Bayes algorithm uses Google Collab and Rapidminer to process data to produce accurate accuracy.

REFERENCE

Alfiah, N. (2021). Klasifikasi Penerima Bantuan Sosial Program Keluarga Harapan Menggunakan Metode Naive Bayes. *Jurnal Teknologi Informasi*, *16*(1), 32–40. https://doi.org/10.35842/jtir.v16i1.386

European Union. (2017). *Sustainable Development Goals*. Retrieved from https://www.sdg2030indonesia.org

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Irnawan, F. D., Hidayah, I., & Nugroho, L. E. (2021). Metode Imputasi pada Data Debit Daerah Aliran Sungai Opak, Provinsi DI Yogyakarta. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *10*(4), 301–310. https://doi.org/10.22146/jnteti.v10i4.2430

Khakim, E. N. R. (2022). The Comparison of the Social Welfare Data Classification Algorithm for Bantul Regency, *17*(2), 91–100. DOI: https://doi.org/10.33998/processor.2022.17.2.1222

Nurlia, E., Jajuli, M., & Purnamasari, I. (2021). Penerapan Naïve Bayes Untuk Klasifikasi Tingkat Risiko Diagnosis Gigi Di Uptd Puskesmas Cingambul. *JIKO (Jurnal Informatika Dan Komputer)*, *4*(2), 127–132. https://doi.org/10.33387/jiko.v4i2.3190

Primajaya, A., Sari, B. N., & Khusaeri, A. (2020). Prediksi Potensi Kebakaran Hutan dengan Algoritma Klasifikasi C4.5 Studi Kasus Provinsi Kalimantan Barat. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *6*(2), 188–192. https://doi.org/10.26418/jp.v6i2.37834

Saputra, P. S., Dantes, G. R., & Gunadi, I. G. A. (2021). Perbandingan Algoritma Fuzzy C-Means Dan Algoritma Naive Bayes Dalam Menentukan Keluarga Penerima Manfaat (Kpm) Berdasarkan Status Sosial Ekonomi (Sse) Terendah. *JST (Jurnal Sains Dan Teknologi)*, *10*(1), 1–8. https://doi.org/10.23887/jstundiksha.v10i1.23340

Saragi, N. R., Sembiring, A., & Nurhayati. (2022). Sistem Pakar Mendiagnosa Kelayakan Air Minum untuk Dikonsumsi menggunakan Metode Certainty Factor pada PDAM Tirta Sari Kota Binjai. *Jurnal Citra Sains Teknologi*, *2*(1), 23–26. DOI: https://doi.org/10.2421/cisat.v2i1.63

Sari, Y. S. (2021). Penerapan Metode Naïve Bayes Untuk Mengetahui Kualitas Air Di Jakarta. *Jurnal Ilmiah FIFO*, *13*(2), 222–228. https://doi.org/10.22441/fifo.2021.v13i2.010

Tempola, F., Muhammad, M., & Khairan, A. (2018). Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(5), 577–584. https://doi.org/10.25126/jtiik.201855983