CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers

by Dr.ayi Purbasari.,st.,mt -

Submission date: 18-Jun-2022 11:44AM (UTC+0700)
Submission ID: 1858878431
File name: CRISP-DM_for_Data_Quality_Improvement.pdf (600.42K)
Word count: 4514
Character count: 24020

CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers

[1]Ayi Purbasari
Informatics Engineering dept.
Universitas Pasundan
Bandung, Indonesia
pbasari@unpas.ac.id

[2]Fedri Ruluwedrata Rinawan
Public Health dept.
Universitas Padjadjaran
Sumedang, Indonesia
f.rinawan@unpad.ac.id

[3]Arief Zulianto
Master of Informatics
Universitas Langlangbuana
Bandung, Indonesia
madzul@unla.ac.id

[4]Ari Indra Susanti
Public Health dept.
Universitas Padjadjaran
Sumedang, Indonesia
ari.indra@unpad.ac.id

[5]Hendra Komara
Informatics Engineering dept.
Universitas Pasundan
Bandung, Indonesia
hendra@unpas.ac.id

***Abstract*—Many Machine Learning (ML) projects ended up only as proof concept and failed to be produced. Therefore, this research focused on well-defined processes that must be followed, adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) with the specifications and requirements of supervised and unsupervised learning which include a methodology for Classification/grouping. The Data Understanding and Data Preparation phases, used transactional data on examination of infants and toddlers in 2018-2021 on the iPosyandu application. At the Business Understanding stage, the ML was intended to predict stunting, so that data quality of iPosyandu can be informed and then recommendations and feature improvements and assistance for end-users can be made. The output of Data Understanding and Data Preparation was in the form of baby & toddler examination dataset, which was used in the Machine Learning modeling stage, especially to classify and predict nutritional/stunting status. Of the 192 tables contained in the iPosyandu application, there were 5 main tables that were needed to define the dataset. 75,652 data on infants and toddlers were checked with 49,615 data of examinations in 3173 Posyandu, which resulted in clean data of 39,411 rows of datasets for all examinations and 13,868 rows of datasets for the last examination of infants and toddlers. The dataset was combined with the nutritional status of infants and toddlers resulting from the calculation of the baby's weight, length of the baby's body, and the comparison of the baby's height and weight. The dataset was tested into the ML using the Orange Application and produced Classification model that can be used for prediction. From the results of the modeling evaluation, it can be seen that the Naïve Bayes Algorithm had an advantage with a predictive value of 0.851 while the Tree algorithm was 0.848 and the Neural Net was 0.845. From the overall evaluation, it can be concluded that there is a need to improve data quality by improving the application and improving the literacy of the end-users, so that the data has better quality and ready to be used as a ML dataset. The selected features can be aggregated to simplify the modeling process so as to obtain the expected model.**

***Keywords*—CRISP-DM, Machine Learning, Data, Prediction, Stunting, iPosyandu, Infant, Toddler, Orange**

I. INTRODUCTION

Analysis for large data sets using Machine Learning (ML) gives organizations a competitive advantage by gaining insight into customer behavior, process efficiency, business impact, and includes the classification and prediction of malnutrition and stunting. However, many ML projects ended up only as proof of concept and failed to be produced. One reason is that they lacked a well-defined process to follow. To run a ML project efficiently, it is important to define the tasks to be completed and the roles involved. These define a structured process that drives the project team towards well-defined goals and ensures a common understanding of business requirements. While many process models can be used for a Data Mining project, they cannot be applied effectively to a ML project without adapting and adding tasks.

This is also true for the application of ML in stunting prediction that utilizes data from iPosyandu. iPosyandu is an Android-based application created for monthly and annual recording and reporting of the Posyandu (Integrated Health Service Post) Information System, which was originally a manual written ledger [1]. The application is built with the work pattern approach of the cadres [2] who later become users of this application [3]. The application is accompanied by a manual and monitoring feedback from cadres who are users [4]. This application is currently used by cadres at Posyandu in Pasawahan sub-district, Purwakarta Regency and will be expanded to Posyandu assisted by PT Astra International, Tbk., which are spread throughout Indonesia. Thus, the iPosyandu application has large data with end users of cadres in various regions in Indonesia. The data generated are important and need a deeper understanding.

This research adopted the well-known and widely accepted Cross-Industry Standard Process for Data Mining (CRISP-DM) to reflect ML specifications on processes and ensure successful implementation of ML projects. Since ML covers a wide spectrum of methods, this research focuses on adapting CRISP-DM to the specifications and requirements of supervised and unsupervised learning that includes specific methodologies such as Artificial Neural Networks or Decision Tree.

There are six phases to CRISP-DM that cover further general tasks. Common tasks describe what needs to be done one phase before moving on to the next. The six phases are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluating, and Deployment. The phases of the CRISP-DM, especially the Data Understanding

978-1-6654-1743-3/21/$31.00 ©2021 IEEE of ICAICTA2021

and Data Preparation phases can be used in the ML stage for predicting stunting in babies and toddlers.

The initial stages of Data Understanding and Data Preparation in the CRISP-DM were implemented in the ML stages. The work referred to the CRISP-DM but with a focus on Data Understanding and Data Preparation. The data used were transaction data of examination of infants and toddlers in 2018-2021 on the iPosyandu application. The output of Data Preparation was in the Machine Learning modeling stage, especially for classifying and predicting nutritional/stunting status.

This research used the CRISP-DM method as in Figure 1.

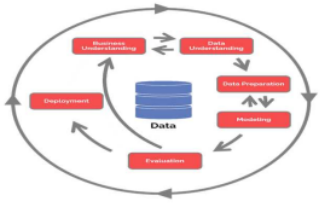

Figure 1 CRISP-DM [5]

The method has the following stages:

1. *Business understanding – What does the business need?*
2. *Data understanding – What data do we have / need? Is it clean?*
3. *Data preparation – How do we organize the data for modelling?*
4. *Modelling – What modelling techniques should we apply?*
5. *Evaluation – Which model best meets the business objectives?*
6. *Deployment – How do stakeholders access the results?*

The research focused on Data Understanding and Data Preparation. The output of data preparation activities was the iPosyandu dataset used in the modeling stage.

II. CRISP-DM

Research related to the use of CRISP-DM dates back to the early 2000s. CRISP-DM, which has existed for 20 years, continues to be used by various projects and research. Martinez et al [5] conducted research on CRISP-DM in the second half of the nineties, which is about two decades ago. According to many user surveys and polls, it is still the *de facto* standard for developing data mining and knowledge discovery projects. However, there is no doubt that this field has grown rapidly in the past twenty years, with data science now becoming the preferred term over data mining. Martinez et al investigated whether, and in what context, CRISP-DM is still suitable for data science project objectives. Research results show that if projects are goal-directed and process-driven, the process model view is still largely applicable. On the other hand, as data science projects become more exploratory, the paths that projects can take become more varied, and more flexible models are needed. Martinez et al suggested a trajectory-based model outlines and how it can be used to categorize data science projects (goal-directed, exploratory, or data management). From this research, it can be inferred that CRISP-DM is still relevant for research related to data science and therefore can also be used for ML research.

R. Wirth and J. Hipp stated that the CRISP-DM Project proposes a comprehensive process model for implementing a Data Mining project. The process model does not depend on the industry sector and the technology used. In their research, R. Wirth and J. Hipp applied a standard process model to data mining and reported experience with the CRISP-DM process model in practice [6].

Meanwhile, H. Wiemer et al [7], conducted research on an extension of the CRISP-DM methodology, named Data Mining Methodology for Engineering Applications (DMME) aimed at engineering applications. The research was based on the fact that CRISP-DM does not determine the data acquisition phase in the production scenario. DMME provides a communication and planning foundation for data analysis in the manufacturing domain as well as the design and evaluation of the infrastructure for data acquisition that is integrated into the process. In addition, the methodology includes the design function of the experimenter's ability to systematically and efficiently identify relevant interactions. Wiemer et al, in their research, showed the DMME methodological procedures that are presented in detail and project examples that illustrate the workflow. The research was complemented by a case study that was part of a collaborative project with an industrial partner who wanted an application to detect marginal lubrication in the linear guide of a servo driven shaft based solely on data from a drive controller. The case study showed that the DMME methodology was used as expected.

DMME related research by [8] stated that DMME provides a holistic view to data analytics in the production department that supports the prerequisites necessary for successful implementation of data-driven processes and analytics. In addition, DMME was developed to be accessible to engineers implementing it to support startup operations as well as optimize production or maintenance processes across the value chain. The proposed expansion of CRISP-DM in DMME is to add a Technical Understanding and Technical Reference which bridge the Business Understanding and Data Understanding. In addition, the Technical Implementation stage was also added before the Deployment stage was carried out.

S. Studer et al [9] explained that 75 to 85 percent of current practical ML projects do not meet the expectations of their sponsors, in terms of data and software quality that are the main challenges in the ML life cycle. Another reason is the lack of guidance through standards and specific development process models for ML applications. Industry organizations, in particular, rely heavily on standards to ensure the consistent quality of their products or services. The Japanese Industrial Consortium (QA4AI) was established to meet this need. Due to the lack of process models for ML applications, many project organizations rely on alternative models closely related to ML, such as the CRISP-DM model. This model is based on the experience of industrial data mining and is considered the most suitable for industrial projects among related process models. However, two major drawbacks of CRISP-DM were identified:

First, CRISP-DM focuses on extracting data and does not cover ML model application scenarios that infer real-time decisions over long periods of time. Figure 2 shows the basic differences between Data Mining and Machine Learning.

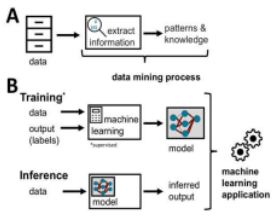

Figure 2 Differences between Data Mining and Machine Learning [9]

The ML model must be able to adapt to a changing environment, otherwise the model's performance will degrade over time, which requires permanent monitoring and maintenance of the ML model after deployment. Second and more concerning, CRISP-DM does not have guidance on the Quality Assurance (QA) methodology. This deficit is evident compared to standards in information technology, but is also evident in alternative process models for data mining. In the context of ML process model, quality is not only determined by the suitability of the product for the purpose, but the also the quality of task execution at each phase during the development of an ML application. This ensures that errors are detected as early as possible to minimize costs at later stages of the development process.

S. Studer et al proposed the CRISP-ML(Q) process model for ML application development, which includes six phases from defining the scope to maintaining the implemented machine learning application. Studer et al proposed that business understanding and data are run concurrently in the first stage, as they both have a considerable impact on the feasibility of the project. The next stage consists of data preparation, modeling, evaluation, and dissemination.

Meanwhile, M. Bohanec [10] conducted research on a new approach to support knowledge-based decision-making processes by applying a general explanatory methodology to ML models. The proposed framework, which formalizes the process of using a company's historical business data in the decision-making process and enhances them with explanations and the people involved. This framework used a transparent description of arbitrary predictive machine learning models, including the best performing black box models.

In the medical/health domain, O. Niakšu [11] saw the lack of a specific and detailed framework to perform data mining analysis. Therefore, an extension of CRISP-DM, called CRISP-MED-DM, was proposed, which addressed the specific challenges of data mining in medicine. The medical application domain with its distinctive challenges was mapped with the CRISP-DM reference model, proposing improvements in the CRISP-DM reference model. Next, a model to evaluate adherence to CRISP-MED-DM was proposed. This model allowed evaluation and comparison of the extent to which various data mining projects followed the CRISP-MED-DM process model. O. Niaksu conveyed the uniqueness of data processing in medicine. R. Belazzi and B. Zupan, K.J. Cios and G.W. Moore, Canlas Jr., R. D., and others stated that the practical application of data processing/data mining in medicine has a number of barriers: technology, interdisciplinary communication, ethics, and patient data protection. In addition, there are several well-known biomedical data problems, such as inaccurate and fragmented information. The challenges of medical data processing/data mining include: Variation of medical data and data formats (multi-relational structured data, video and image files, text files, etc.), Heterogeneous data, Patient data privacy and Clinical data quality and completeness.

Based on previous research, it can be concluded that CRISP-DM is still used as a basic reference for the implementation of Data Mining and is extended to the use of ML, including specific health domains. This is a follow-up research on how to use CRISP-DM that is more suitable for ML in the medical field.

III. CRISP-DM in Supporting Data Collection for Machine Learning to Predict Stunting

A. Business understanding

At this stage, an understanding is needed – What does the business need? It was agreed that ML was intended to predict stunting. However, what is more important is that with ML, the quality of data from the iPosyandu application can be informed, so that recommendations and feature improvements can be made as well as assistance for end-users.

B. Data Understanding

At this stage, an understanding is needed – *What data do we have / need? Is it clean?*

iPosyandu has master data and also transactional data in the form of baby check-ups, as well as data for reporting. The total table created for the iPosyandu application increased, along with the continuous development of iPosyandu. There were a total of 192 tables contained in the iPosyandu application. Some of the data that focused on are:

1) Babies data: baby master data, for all babies registered at Posyandus
2) Posyandu data: the Posyandu data itself
3) Babies checkup data: monthly baby checkup data by cadres in each Posyandu
4) bgo_all data: data from check-up results that show the baby is below the orange line, which indicates stunting.
5) bgm_all data: data from check-up results that show the baby is below the red line, which indicates stunting.

The summary of the data used is presented in Table 1.

The table data structure used is summarized in Figure 3 (iPosyandu data structure).

C. Data Preparation

At this stage, an understanding is needed – How do we organize the data for modeling? The target of this stage is a data set that can be used for classification. Hence, it was necessary to set features and targets. In order to form a dataset, data cleansing is required. Data Preparation is presented in Table 1 below. The data that had been prepared produced the Infant and Toddler Examination dataset from iPosyandu. Here is the resulting dataset, in .csv format.


```
"id","posyandus_id","babies_id","checkup_date","height","height_status","weight","weight_status","measure","exclusive_breast_milk","imd","feb_vit_a","aug_vit_a","hbo","dpt_hb","dpthb_2","dpthb_3","mr","bcg","polio","polio2","polio3","polio4","ipv","kia","pmt","pmt_amount","diare","oralit_amount","created_at","updated_at","deleted_at","indexed","f2_indexed","f6_indexed","f3v2_indexed","f2v2_indexed","child_no","dob","gender","birth_weight","expectant_labor_id","cried_at_birth","bbu","pbu","bbpb","age"

"11","5","1952","11/12/2018","0","Naik","8.5","Naik","Terlentang","0","0","0","0","0","0","NULL","NULL","0","0","0",,,,"0","0",,,,,"11/12/2018 08:17:36","28/4/2021 22:18:36",,,"y","n","n","n","n","2","18/2/2018","perempuan","2.6",,,"0","0","0","10"
```

Figure 3 Dataset of infant check-up

Figure 4 iPosyandu data structure

D. Checking Dataset Using Orange

Before the modeling, the dataset was checked. The following is the procedure for using the d15cu.csv dataset in the Orange application, summarized in Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, and Figure 11.

1) File component was used, and connected with dataset

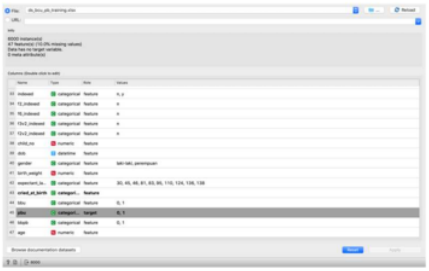

Figure 5 Feature and target setting

2) The infant examination dataset was used on the date of 24last examination. The data were divided into 3 parts: training data, testing data, and evaluation data with a portion of 70:20:10. The pbu field feature was set and was ensured as the target, with a value of 1 for stunting and 0 for not stunting.
3) The data training is as follows

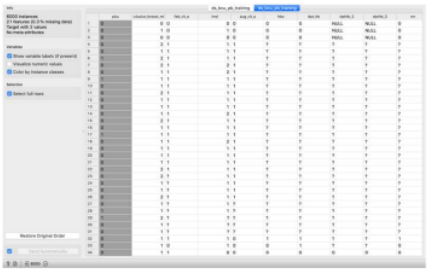

Figure 6 Data training

Features are selected as follows:

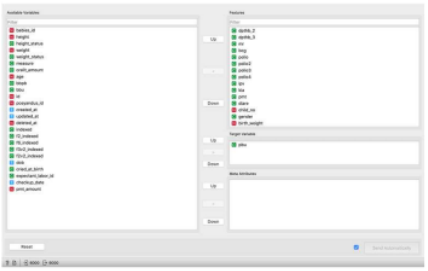

Figure 7 Feature selection

The prepared dataset had the statistics summary as follows:

Figure 8 Statistics of data training

Figure 9 Statistics of data training (continued)

It was found that there are 8.5% stunting data or 510 instant data:

E. Modeling Using Orange

This section discusses a complete modeling with 26the Orange application. The Orange application, or Orange Data Mining, is open source software to perform data mining or data analytic processes through visual programming concepts. Orange provides many widgets placed on the canvas/drawing board and linked with other widgets [12]. Amala [13] used Orange for Classification. D. Vaishnav and B. R. Rao [14] used Orange to compare machine learning algorithms for Fruit Classification. Meanwhile Naik and Sama14,15] conducted a review for classification algorithms in data mining using WEKA, Rapidminer, Tanagra, Orange and Knime. Dems˘ar et al explained that Orange can be used with additional scripts in Python [16].

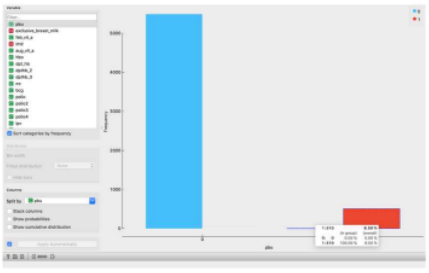

Figure 10 Stunting data based on gender

In this research, the Orange application was connected to 2 datasets that have been prepared, namely training data and test data. Feature selection was carried out, where 22 features were selected to be used. By using the Tree widget for prediction, an evaluation was carried out using testing data. Figure 12 shows the model used.

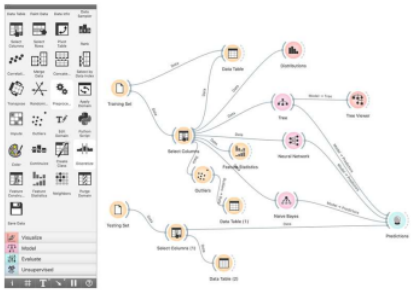

Figure 11 Stunting prediction modeling

The tree results show that the root chosen is vitamin A, in Figure 13 below.

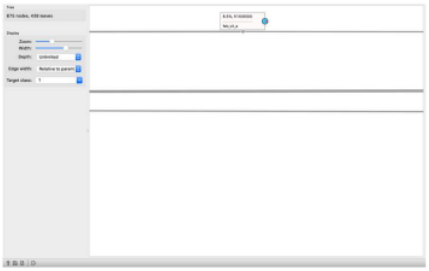

Figure 12 Produced tree

F. Evaluation

At this stage, an understanding is needed – Which model best meets the business objectives? Because prediction is the goal, in this evaluation, prediction was tested using the Prediction widget and a dataset in the form of a testing set. From the prediction results, it can be inferred that the comparison of the three algorithms used is summarized in Figure 14.

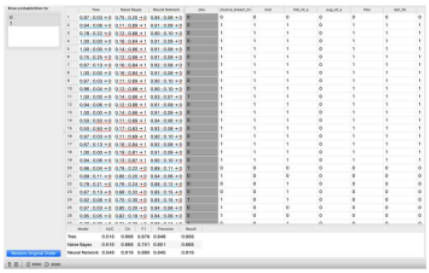

Figure 13 Model evaluation

It can be inferred that Naïve Bayes has a higher level of precision than Tree and Neural Net.

G. Deployment

At this stage, an understanding is needed – How do stakeholders access the results? Since this ML research is ongoing, at the deployment stage, stakeholders are given access in the form of reports on the results of data cleansing and initial ML modeling with 3 algorithms.

IV. Conclusions and Future Works

Based on the research results, there are several conclusions and recommendations for development.

1. Data quality aspect. Handling is needed for the application and end-users' literacy is also needed, so that the inputted data is better and ready to be used as ML datasets;
2. It is necessary to improve the model and improve the data for Classification;
3. From the results of the modeling evaluation, it can be inferred that the Naïve Bayes Algorithm is slightly superior to the Tree and Neural Net algorithms;
4. The data used for ML is the last baby examination data, with 22 features that should be summarized and regrouped;
5. In terms of tools utilization, the Orange application is easier to use because it is in a visual form with drag-and-drop widgets. However, a basic understanding of the datasets, features, and algorithms used is still required;
6. The use of the CRISP-DM method can be studied further specifically to support ML or for the medical domain.

Acknowledgment

This research was funded by RISPRO Innovation Research – LPDP in collaboration with Padjadjaran University, Pasundan University, Langlangbuana University, and Purwakarta District Health Office.

Table 1 iPosyandu data understanding and preparation

Data	Quantity	Condition	Follow-up	Number of clean data	Feature candidate
Babies data	75,652	Fields that must be considered: Date-of-birth, null or empty conditions and Gender, birthweight, and child_no, each of which describes gender, birth weight, and children older. There were 25,758 rows of problematic data	1. The dobs containing null or '00-00-000 were cleaned 2. Incorrect weight data due to unit differences were cleaned 3. Data with null or empty gender and child_no were deleted	48,882	Gender, birth_weight, and child_no

Posyandu data	3,173 Posyandu	No problems were found in this table	The data were used for data row selection based on the area of the Posyandu code	3,173	None
Babies_checkup data	49,615	As of 2021, there are 49,615 data rows in the babies_checkup table. The problem is that babies data is bigger than check_up data. Considering that babies will do check-ups repeatedly, this check_up data should be more than babies. This will be a consideration for the data preparation phase	1. Null on check_up date was checked 2. Height and weight data, height and weight of infants & toddlers were examined. 9,441 data contained null 3. 194 data rows for infants and toddlers with a height above 120 cm were found. According to statistics, no child under five is taller than 120 cm. The data were removed from the system. 4. 968 data rows for infants and toddlers weighing more than 50 kg. According to statistics, no child under five is over 50 kg. The data were removed from the system. 5. Data on babies and toddlers at the last check were searched	Overall checked: 39,411 Last check: 13,868	id, posyandus_id, babies_id, checkup_date, height, height_status, weight, weight_status, measure, exclusive_breast_milk, imd, feb_vit_a, aug_vit_a, hbo, dpt_hb, dpthb_2, dpthb_3, mr, bcg, polio, polio2, polio3, polio4, ipv, kia, pmt, pmt_amount, diare, oralit_amount
Bgo_all data	6,358 data rows in bgo_all table.	The attributes of bbu, pbu, bbpb are filled in by the application from the calculation of the baby's height and weight. bbu 1 means malnourished, pbu = 1 means short/stunting, bbpb = 1 means underweight. Stunting case if found pbu is valued 1 (true)	The bbu, pbu, bbpb and, age fields were combined into the dataset		Bbu, pbu, bbpb Pbu as target. 1 for stunting and 0 for not stunting
Bgm_all data	3,567 data row in bgm_all table.	Basically, this table is used for reporting stunting for the criteria below the red line and it is worse than the bgo/orange line. The decision of whether a baby under five is stunted is calculated by the application by looking at the baby's height and weight. bbu = 1 means poor nutrition, pbu = 1 short, means stunting, and bbpb = 1 means very underweight	1. The bbu, pbu, bbpb and age fields were combined into the dataset 2. For dataset version 1, the data from bgm had not been used		Bbu, pbu, bbpb Pbu as target. 1 for stunting dan 0 for not stunting.

REFERENCES

[1] F. R. Rinawan *dkk.*, "Understanding mobile application development and implementation for monitoring Posyandu data in Indonesia: a 3-year hybrid action study to build 'a bridge' from the community to the national scale," *BMC Public Health*, vol. 21, no. 1, hlm. 1024, Des 2021, doi: 10.1186/s12889-021-11035-w.

[2] A. I. Susanti, F. R. Rinawan, dan I. Amelia, "Mothers Knowledge and Perception of Toddler Growth Monitoring Using iPosyandu Application," *Glob. Med. Health Commun. GMHC*, vol. 7, no. 2, Agu 2019, doi: 10.29313/gmhc.v7i2.3892.

[3] A. I. Susanti, D. Didah, A. N. Sari, D. Ferdian, dan F. R. Rinawan, "PERSEPSI PETUGAS GIZI DALAM PEMANTAUAN STATUS GIZI BALITA DENGAN MENGGUNAKAN WEBSITE IPOSYANDU," *J. Kebidanan Malahayati*, vol. 6, no. 3, hlm. 376–382, Jul 2020, doi: 10.33024/jkm.v6i3.2667.

[4] W. Widarti, F. R. Rinawan, A. I. Susanti, dan H. N. Fitri, "Perbedaan Pengetahuan Kader Posyandu Sebelum dan Sesudah Dilakukan Pelatihan Penggunaan Aplikasi iPOSYANDU," *J. Pengabdi. Dan Pengemb. Masy.*, vol. 1, no. 12, hlm. 143, Feb 2019, doi: 10.22146/jp2m.43473.

[5] F. Martinez-Plumed *dkk.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, hlm. 1–1, 2020, doi: 10.1109/TKDE.2019.2962680.

[6] R. Wirth dan J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," hlm. 11.

[7] H. Wiemer, L. Drowatzky, dan S. Ihlenfeldt, "Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model," *Appl. Sci.*, vol. 9, no. 12, hlm. 2407, Jun 2019, doi: 10.3390/app9122407.

[8] S. Huber, H. Wiemer, D. Schneider, dan S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model," *Procedia CIRP*, vol. 79, hlm. 403–408, 2019, doi: 10.1016/j.procir.2019.02.106.

[9] S. Studer *dkk.*, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, hlm. 392–413, Apr 2021, doi: 10.3390/make3020020.

[10] M. Bohanec, M. Robnik-Šikonja, dan M. Kljajić Borštnar, "Decision-making framework with double-loop learning through interpretable black-box machine learning models," *Ind. Manag. Data Syst.*, vol. 117, no. 7, hlm. 1389–1406, Agu 2017, doi: 10.1108/IMDS-09-2016-0409.

[11] O. Niakšu, "CRISP Data Mining Methodology Extension for Medical Domain," hlm. 19.

[12] J. Demšar dan B. Zupan, "Orange: Data Mining Fruitful and Fun - A Historical Perspective," hlm. 6.

[13] M. G. Amala, "Orange Tool Approach For Comparative Analysis Of Supervised Learning Algorithm In Classification Mining," hlm. 10, 2019.

[14] D. Vaishnav dan B. R. Rao, "Comparison of Machine Learning Algorithms and Fruit Classification using Orange Data Mining Tool," dalam *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, Nov 2018, hlm. 603–607. doi: 10.1109/ICICT43934.2018.9034442.

[15] A. Naik dan L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Comput. Sci.*, vol. 85, hlm. 662–668, 2016, doi: 10.1016/j.procs.2016.05.251.

[16] J. Demšar *dkk.*, "Orange: Data Mining Toolbox in Python," hlm. 5.

CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers

ORIGINALITY REPORT

16%	10%	11%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Stefan Studer, Thanh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, Klaus-Robert Müller. "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology", Machine Learning and Knowledge Extraction, 2021 Publication	5%
2	www.coursehero.com Internet Source	2%
3	core.ac.uk Internet Source	1%
4	Submitted to Universitas Muhammadiyah Surakarta Student Paper	1%
5	Qili Chen, Guangyuan Pan, Wenbai Chen, Peiliang Wu. "A Novel Explainable Deep Belief Network Framework and Its Application for Feature Importance Analysis", IEEE Sensors Journal, 2021	1%

Publication

6	jurnal.stts.edu Internet Source	1 %
7	Amri Muliawan Nur, Muhammad Farid Wazdi, Bambang Harianto, Muhammad Farid Zaini. "Implementation of Naive Bayes Algorithm in Analyzing Acceptance of Poor Student Assistance", Journal of Physics: Conference Series, 2020 Publication	<1 %
8	Arief Zulianto, Ayi Purbasari, Neni Suryani, Ari Indra Susanti, Fedri R. Rinawan, Wanda G. Purnama. "Pemanfaatan Katalon Studio untuk Otomatisasi Pengujian Black-Box pada Aplikasi iPosyandu", Jurnal Edukasi dan Penelitian Informatika (JEPIN), 2021 Publication	<1 %
9	Submitted to Bridgepoint Education Student Paper	<1 %
10	e-jurnal.lppmunsera.org Internet Source	<1 %
11	jurnal.polibatam.ac.id Internet Source	<1 %
12	Grace Y. Smith, Christine M. Schubert Kabban, Kenneth M. Hopkinson, Mark E. Oxley, George E. Noel, Huaining Cheng. "Sensor Fusion for Context Analysis in Social Media COVID-19	<1 %

	Data", NAECON 2021 - IEEE National Aerospace and Electronics Conference, 2021 Publication	
13	Nutan Singh, Priyanka Tripathi. "Live Streaming of Face Mask and Body Temperature Detection System using Transfer Learning and IoT", Journal of Physics: Conference Series, 2022 Publication	<1 %
14	www.warse.org Internet Source	<1 %
15	www.researchsquare.com Internet Source	<1 %
16	ijetae.com Internet Source	<1 %
17	jurnal.untan.ac.id Internet Source	<1 %
18	Submitted to University of Salford Student Paper	<1 %
19	assets.researchsquare.com Internet Source	<1 %
20	Submitted to Botswana Accountancy College Student Paper	<1 %
21	Submitted to MCI Management Centre Innsbruck Student Paper	<1 %

22	Submitted to Pasundan University Student Paper	<1 %
23	Submitted to Westcliff University Student Paper	<1 %
24	Javad Rahnama, Eyke Hüllermeier. "Learning Tversky Similarity", Information Processing and Management of Uncertainty in Knowledge-Based Systems Internet Source	<1 %
25	Giordano d'Aloisio. "Quality-Driven Machine Learning-based Data Science Pipeline Realization: a software engineering approach", 2022 IEEE/ACM 44th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 2022 Publication	<1 %
26	www.kdd.org Internet Source	<1 %
27	Ayi Purbasari, Arief Zulianto, Achmad Nizar Hidayanto. "Revealed-Preference Activity Rule in Combinatorial Clock Spectrum Auction: A Review and New Research Opportunities", 2018 Third International Conference on Informatics and Computing (ICIC), 2018 Publication	<1 %

28	Veronika Plotnikova, Marlon Dumas, Fredrik Milani. "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements", Data & Knowledge Engineering, 2022 Publication	<1%
29	rinarxiv.lipi.go.id Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	Off		