JurnalMantik
Volume 4, Number 4, Feb 2021, pp. 2599-2603 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Prediction of Drinking Water Facility Conditions Using the Naive Bayes Algorithm

Nur Yulias[1], Septian Rheno Widianto[2]

[1,2]Jurusan Sistem Informasi Bisnis, Sekolah Tinggi Manajemen Ilmu Komputer LIKMI / Jl. Ir. H. Juanda No.96, Lebakgede, Kecamatan Coblong, Kota Bandung, Indonesia,

E-mail: djadjoe@gmail.com[1], septian.rheno@yahoo.de[2]

ARTICLEINFO

Article history:
Received: 10/01/2021
Revised: 20/01/2021
Accepted: 30/01/2021

Keywords:
Data Mining, Clasification, Naive Bayes, Drinking Water Facility, Pamsimas

ABSTRACT

This study aims to optimize the selection of raw water source options at the location of Pamsimas III program. This analysis will affect the condition of the facilities that will function properly. The data mining method used in this study is a classification model with the Naive Bayes algorithm using the Rapidminer application tool. The data processed is SIM Pamsimas III data with the object of research on the new village drinking water facilities for the 2017-2019 Pamsimas program. This research analyzes the prediction of the condition of drinking water facilities based on the option of raw water sources. So this research can helps to determine the level of potential facilities that will function properly based on the specified raw water source options. Based on the research conducted, predictive analysis using Naive Bayes has an accuracy rate of 85.05%. So that the selection of the raw water source option can predict the condition of the drinking water facilities being built will function properly.

1. Introduction

The Indonesian government has an SDGs agenda for the 2016-2030 period which mandates to realize universal access to safe and sustainable drinking water [1]. The Sustainable Development Goals (SDGs) agenda includes access to air and sanitation which previously replaced the MDGs, this affects national policies to achieve universal access to drinking water as measured by the World Health Organization (WHO) and United Nations Children's Fund (UNICEF) through the Joint Monitoring Program for Water and Sanitation (JMP) [2]. However, the guarantee of communal water supply in rural areas is very limited on the ability of the community so that it requires financial support as an investment cost in order to provide clean water services widely [3].

The demand for clean water needs to increase every time the population increases, while the implementation of the Government in providing clean water in Indonesia is still lacking [4, p. 5]. Based on the SDGs agenda and the need to provide clean water services to the community, this is a concurrent and mandatory government affair. One of them is the construction of drinking water facilities through the Community Based Water and Sanitation Program (PAMSIMAS) III which has succeeded in building drinking water facilities and is well known in several villages throughout Indonesia. From the results of the construction of drinking water facilities, it is found that the conditions of drinking water facilities are in good functioning status.

This study aims to perform data mining analysis to determine the options of raw water sources in well-functioning drinking water facilities. So that it can be seen the source of raw water which is a source of drinking water that is functioning properly. Using the data mining classification method with the Naive Bayes algorithm model.

2. Literature Review

This study uses the Prediction of Drinking Water Facility Conditions Using the Naive Bayes Algorithm

2.1 Literature Method

The method of collecting library data is done by collecting data from sources or books that are relevantto the research.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

2.2 Metode Wawancara

The interview method is carried out by asking directly to the data manager and Drinking Water Facility Conditions Using the Naive Bayes Algorithm .

3. Research Methodology

Data mining is the process of analyzing data from multiple perspectives and summarizing it into useful information, patterns, relationships, or relationships among all these data can provide information [5]. To carry out the data mining process at the research stages carried out include: (1) data collection, (2) data cleaning, (3) model implementation, (4) evaluation of classification performance

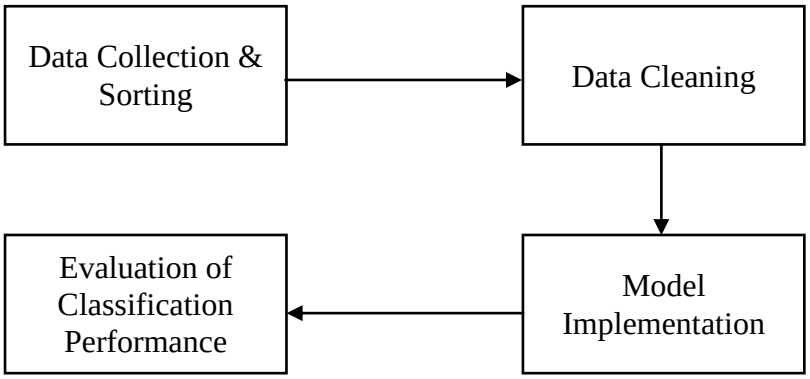


Fig 1. Research Stages

3.1 Data Collection & Sorting

The data used in this study are SIM PAMSIMAS III data specifically for new villages for 2017-2019. Data is taken from several tables and stored in one table.Then determine the attributes that will be used and get rid of the unnecessary attributes such as the name of the province, district, sub-district, and village. The attributes used in this study are the attributes of raw water sources and the functioning of drinking water facilities which then become data training and data testing.

Table 1.
Attributes of Raw Water Sources and Function of Drinking Water Facilities

Attribute	Attribute Value
Water Springs	{Y,N}
Dug Well	{Y,N}
Shallow Drilling Well	{Y,N}
Deep Drilling Well	{Y,N}
Water Level	{Y,N}
TappingPDAM	{Y,N}
The functioning of Drinking Water Facilities	{Functions Good, Functions Partly, Doesn't Work}

3.2 Data Cleaning

Data quality is a major problem in datasets, especially in pattern discovery. Because data quality problems can produce wrong output if data analysis is done incorrectly [6].Before carrying out the data mining process, it is done first before processing the data to make improvements to the data to be analyzed. This selection process is important so that raw data tends not to be mined. Like the case of missing values in the dataset of raw water sources and the functioning of drinking water facilities, the attributes that are not needed are carried out by the process of taking data objects (undersampling).

3.3 Model Implementation

In the process of applying the classification model algorithm with the Naïve Bayes algorithm. The classification model is a data mining method for assigning sample data to one of several categories. There are many classifications of classifications that were developed to outperform one another. But all of them are based on mathematical techniques, for example like Naive Bayes [7].The algorithm process uses the Rapid Miner application. Rapid Miner is a system that supports the design and documentation of the entire data

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

mining process. It offers not only the operator a nearly incomplete set, but also a structure expressing the flow of process control [8, p. xx].

The Naive Bayes algorithm is a simple probability classification that calculates a set of probabilities by calculating the frequency and combination of values in a particular data set. This algorithm uses the Bayes theorem and assumes that all variables are independent by considering the value of the class variable. These conditional independent assumptions are rarely valid in real-world applications, so they are characterized as Naive, but algorithms tend to learn rapidly in a variety of controlled classification problems [9].

3.4 Evaluation of Classification Performance

This section provides a measure to assess whether or how accurate the classification is in predicting tuple class labels so that the level of accuracy of the predictions can be determined. Evaluation is carried out using a confusion matrix.A confusion matrix is a useful tool for analyzing how well a classification of tuples can be made of several classes. TP and TN will notify when the classification is doing something right, while FP and FN will notify when the classification has made a mistake [10, p. 365].

Table 2.
The Confusion Matrix Model [10, p. 366]

		Predicted Class		
		yes	*no*	Total
Actual Class	*yes*	*TP*	*FN*	P
	no	*FP*	*TN*	N
	Total	P'	N'	P + N

Next, take evaluation measurements, starting with calculating accuracy. The classification accuracy of a given test set is the proportion of the data set correctly classified. The accuracy calculation on the confusion matrix table is as follows:

$$accuracy = \frac{TP+TN}{P+N} \qquad [10]$$

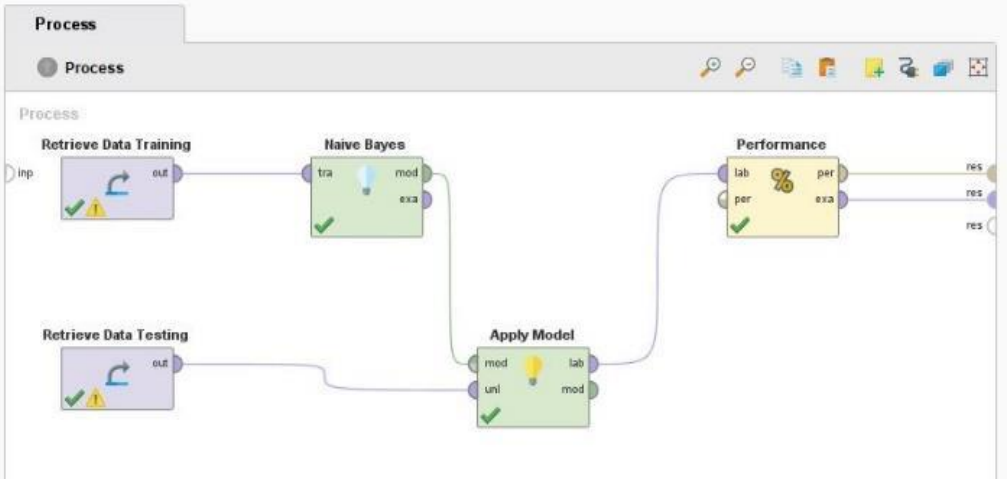
Precision measures and recall are also widely used in classification. Precision can be given as a measure of precision (that is, what proportion of tuples labeled as positive is actually), whereas recall is a measure of completeness (what proportion of positive tuples are labeled as such). If recall looks familiar, it is because it equals sensitivity (or true positive ratio) [10]. This measure can be calculated as follows:

$$Precision = \frac{TP}{TP+FP} \qquad [10]$$

$$recall = \frac{TP}{TP+FN} = \frac{TP}{P} \qquad [10]$$

4. Results and Discussion

Conduct a classification assessment. In this study, it can be seen that the Naïve Bayes algorithm can be used and good results can be obtained from the classification algorithm using the RapidMiner application



Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Fig 2. The classification process with Rapid Miner

In this study it can be observed that the functioning of water facilities based on the selection of raw water source options is classified with an accuracy of 85.05% with the Naive Bayes algorithm using the parameters obtained, thus indicating that the algorithm can be used to detect the function of air facilities predicted from the raw water source option.

Table 3.
ConfusionTable

accuracy: 85.08%

	true Berfungsi Baik	true Berfungsi Sebagian	true Tidak Berfungsi	class precision
pred. Berfungsi Baik	5615	606	379	85.08%
pred. Berfungsi Sebagian	0	0	0	0.00%
pred. Tidak Berfungsi	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

5. Conclusions

Based on the analysis using the Naive Bayes algorithm, its function can be predicted to determine the option of raw water sources for the construction of new drinking water facilities. The evaluation results using a confusion matrix show a significant level of accuracy. Based on these results it can also be ignored that there is a strong relationship and influence on the selection of raw water sources for functioning drinking water services.

6. References

[1] E.W.Purwanto,"PembangunanAksesAirBersihPascaKrisisCovid-19,"*J.Perenc.Pembang.Indones.J.Dev.Plan.*,vol.4,no.2,pp.207–214,2020,doi:10.36574/jpp.v4i2.111.
[2] J.Weststrate,G.Dijkstra,J.Eshuis,A.Gianoli,andM.Rusca,"TheSustainableDevelopmentGoalonWaterandSanitation:LearningfromtheMillenniumDevelopmentGoals,"*Soc.Indic.Res.*,vol.143,no.2,pp.795–810,2019,doi:10.1007/s11205-018-1965-5.
[3] S.Saskya,"PenentuanModelSistemPenyediaanAirMinumPerdesaan,"*J.Wil.danPerenc.Kota*,vol.21,no.2,pp.81–94,2010.
[4] P.TriJuwonoandA.Subagiyo,*SumberDayaAirdanPengembanganWilayah:InfrastrukturKeairanMendukung...-PitojoTriJuwono,ArisSubagiyo-GoogleBooks*,1sted.Malang:UBPress,2018.
[5] M.Sharma,"DataMining :ALiteratureSurvey,"vol.9359,no.2,pp.1–4,2014.
[6] S.KumarandR.Kumar,"DataMining :DirtyDataandDataCleaning,"2020.
[7] L.DhanabalandS.P.Shantharajah,"AStudyonNSL-KDDDatasetforIntrusionDetectionSystemBasedonClassificationAlgorithms,"*Int.J.Adv.Res.Comput.Commun.Eng.*, vol.4,no.6,pp.446–452,2015,doi:10.17148/IJARCCE.2015.4696.
[8] Chapman,*RapidMiner:DataMiningUseCasesandBusinessAnalyticsApplications-GoogleBooks*.CRCPress,2014.
[9] M.M.SandA.Yasar,"PerformanceAnalysisofANNandNaiveBayesClassificationAlgorithmforDataClassification,"pp. 0–1,2019,doi:10.1039/b000000x.
[10] J.Han,M.Kamber,andJ.Pei,*DataMining.ConceptsandTechniques,3rdEdition(TheMorganKaufmannSeriesinDataManagementSystems)*,Third.Waltham:MorganKaufmann,2011.
[11] Widianto, Septian Rheno & Waluyo, Sabar Yoyok (2015). Analisis Serangan SQL Injection pada Server Universitas Nasional. Seminar Nasional Teknik Informatika dan Komputer, JTIK PNJ. Hal. 226-229. ISSN: 2460-9951.
[12] Sinambela, Y., Herman, S., Takwim, A., & Widianto, S. (2020). A STUDY OF COMPARING CONCEPTUAL AND PERFORMANCE OF K-MEANS AND FUZZY C MEANS ALGORITHMS (CLUSTERING METHOD OF DATA MINING) OF CONSUMER SEGMENTATION. Jurnal Riset Informatika, 2(2), 49-54. https://doi.org/10.34288/jri.v2i2.116.
[13] Abdullah, Thoip & Qidri, Sulhan & Nuryadi, Wadi &Widianto, Septian Rheno. (2020) Failover Cluster Nodes and ISCSI Storage Area Network on virtualization Windows Server 2016. JOIN (Jurnal Online Informatika) Volume 5 No.1. Juni 2020: 89-96. DOI: 10.15575/join. v5IL.564. p-ISSN: 2528-1682. E-issn: 2527-9165.
[14] Utami, Amalia & Pratama, Bayu & Widianto, Septian. (2020). DATA MART DESIGN IN BKPP BANDUNG USING FROM ENTERPRISE MODELS TO DIMENSIONAL MODELS METHOD. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). 5. 279-284. 10.33480/jitk.v5i2.1219.

[15] Aditya, Adhisyanda M & Mulyana, Dicky R & Widianto, Septian Rheno (2020). Penggabungan Teknologi Untuk Analisa Data Berbasis Data Science. Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 51-56. ISBN: 978-602-52720-7-3.

[16] Tohirin &Widianto, Septian Rheno. (2020). Peran Trello dalam Adopsi Agile Scrum pada Pengembangan Sistem Informasi Kesehatan. Jurnal Multinetics. Vol 6. No.1. pg.32-39. https://doi.org/10.32722/multinetics.vol6i.2765.

[17] Gunadi, Faustina & Widianto, Septian Rheno. (2020). Efektifitas Pelaporan Pajak Online di Indonesia Berbasis Cobit 5.0 pada Domain MEA (Monitor, Evaluate, Assess). Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 82-85. ISBN: 978-602-52720.-7-3.

[18] Gunadi, Faustina &Widianto, Septian Rheno. (2020). Evaluasi Kualitas Pelaporan Manajemen pada Sistem Epicor Perusahaan Manufaktur Berbasis McCall.Jurnal Multinetics. Vol 6. No.1. pg.21-31. https://doi.org/10.32722/multinetics.vol6i.2765.

[19] Widianto, Septian Rheno. (2015). Perancangan Jaringan WLAN di PT. Gemopia Jewellery Indonesia. Jurnal Multinetics. Vol.1, No. 2. https://doi.org/10.32722/multinetics.Vol1.No.2.2015.pp.50-53.

[20] Mahardi, Sandi & Kuncoro, Adi M & Widianto, Septian Rheno. Integrasi Data Sektoral Pemerintah. (2020). Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 615-617. ISBN: 978-602-52720.-7-3.

[21] Widianto, Septian Rheno & Azzam, Abdullah Izzudin (2018). Analisis Upaya Peretasan Web Application Firewall dan Notifikasi Serangan Menggunakan Bot Telegram pada Layanan Web Server. Jurnal Elektra. Vol. 3, No.2, Juli 2018. Hal. 19-28. ISSN: 2503-0221.

[22] Utami, Sri Farida (2020). Penerapan Data Mining Algoritma Decision Tree BerbasisPSO. Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 677-681.ISBN: 978-602-52720.-7-3.

[23] Widianto, S., S.B.K, F. and Purwanto, A. (2020) "Analysis of Mobile Based Software Development Model: Systematic Review", *Jurnal Mantik*, 4(3, Nov), pp. 1703-1711. doi: 10.35335/mantik.Vol4.2020.973.pp1705-1713.

[24] Widianto, S. and Magdalena, M. (2020) "Online Disposition Data Based Management System", *Jurnal Mantik*, 4(3, Nov), pp. 1641-1648. doi: 10.35335/mantik.Vol4.2020.971.pp1641-1648.

[25] Widianto, S. and Warmayudha, I. P. (2020) "HSQL Database", *Jurnal Mantik*, 4(3, Nov), pp. 1717-1721. doi: 10.35335/mantik.Vol4.2020.982.pp1717-1721.

[26] Widianto, S., Sudiro, S., Suwandi, I. and Leiliawati, L. (2020) "Database Management System on Raw Material Transaction System Case Study : Sabana Fried Chicken", *Jurnal Mantik*, 4(3, Nov), pp. 1722-1727. doi: 10.35335/mantik.Vol4.2020.983.pp1722-1727.