JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 5, No. 1. December 2022 DOI: https://doi.org/10.34288/jri.v5i1.453
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

Implementation of Machine Learning Algorithms for Early Detection of Cervical Cancer Based on Behavioral Determinants

Duwi Cahya Putri Buani[-1*], Indah Suryani[-2]

Informatika
Universitas Nusa Mandiri
Jakarta, Indonesia
duwi.dcp@nusamandiri.ac.id, Indah.ihy@nusamandiri.ac.id
(*) Corresponding Author

Abstract

Cervical cancer is a disease that affects women and has the highest mortality rate after breast cancer. Early detection of cervical cancer is critical at this time, so cervical cancer patients are decreasing. Many women, especially in Indonesia, are less concerned about the dangers of cervical cancer, even though if detected earlier, this disease will be easier to treat. One alternative for early detection can use machine learning algorithms. The machine learning algorithms used in this study are Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), SVM, and Random Forest. In this study, a random under-sampling method was employed, which had no uses in any prior research. This technique makes the accuracy of the five algorithms even better. The research results show that NB has an accuracy rate of 91.67%, LR has an accuracy rate of 87.5%, DT has an accuracy rate of 81.81%, SVM has an accuracy rate of 75%, and RF has the highest accuracy rate of 94.45%. This research shows that the best model is RF or Random Forest.

Keywords: Cervical Cancer; Machine Learning; Random Forest

Abstract

Kanker servik merupakan penyakit yang diidap oleh wanita memiliki tingkat kematikan terbesar di dunia setelah kanker payudara. Deteksi dini kanker serviks sangat penting untuk saat ini, agar pasien kanker serviks semakin berkurang. Banyak wanita terutama di Indonesia kurang peduli dengan bahayanya kanker serviks, padahal jika dideteksi lebih dini penyakit ini akan lebih mudah untuk ditangani. Salah satu alternatif untuk melakukan deteksi dini dapat menggunakan algortima machine learning. Algortima machine learning yang digunakan dalam penelitian ini adalah Naïve Bayes (NB), Logistic Regerson (LR), Decision Tree (DT), SVM dan random Forest. Dalam penelitian ini juga menggunakan teknik Random Under Sampler yang pada penelitian sebelumnya tidak digunakan, teknik ini menjadikan akurasi dari ke-lima algortima menjadi semakin baik. Dari hasil penelitian yang dilakukan menunjukan bahwa NB memiliki tingkat akurasi 91.67%, LR memiliki tingkat akurasi 87.5%, DT memiliki tingkat akurasi 81.81%, SVM memiliki tingkat akurasi 75% dan RF memiliki tingkat akurasi yang paling tinggi yaitu 94.45%. Dari penelitian ini menunjukan bahwa model yang paling baik adalah RF atau Random Forest.

Kata kunci: Kanker servik; Machine Learning; Random Forest

INTRODUCTION

GLOBOCAN *(Global Cancer Observatory)* stated that Asian countries, including Indonesia, contribute most significantly to cancer cases worldwide. Data sourced from Darmais hospital in 2018 showed that the most cancer cases were breast cancer at 19.18%, cervical cancer at 10.69%, and lung cancer at 9.89% (Agustyawati, Fauzi, & Pratondo, 2021; Pangribowo, 2019; Wongkar, Angka, & Angeline, 2022). The WHO *(World Health Organization)* states that cervical cancer is a deadly disease that ranks second only to breast cancer. About 50,000 women have diagnosed with cervical cancer annually (Sobar, Machmud, & Wijaya, 2016)(Setyani, 2018). The high number of cervical cancer patients is influenced by the lack of knowledge among the public, especially women, to carry out early detection before cancer spreads (Aisah, Hafiyusholeh, & Ulinnuha, 2022; Winarni & Suratih, 2020). This data shows that cervical cancer is one of the most common cases of cancer in Indonesia, so it needs to be detected early (Arifin, Siregar, Ratna, & Mudzakir, 2021; Hidayah, Cholissodin, & Adikara, 2019). To perform early detection using machine learning. Machine learning

is used as a classifier to detect the probability of cervical cancer risk based on its behavior and determinants (Feblian & Daihani, 2017). Previous research using the naïve Bayes (NB) and Logistic Regression (LR) algorithms showed the following results in Table 1:

Table 1. Previous Research

Algorithm	Accuracy	Auc
NB	91.67%	0.96
LR	87.5%	0.97

Source: (Sobar et al., 2016)

Table 1 shows previous studies using the same data but only two popular algorithms in 2016: NB and LR. In this study, the authors will compare NB and LR algorithms with other algorithms to find the best model for the early detection of cervical cancer.

Previous research conducted by Sober used the same data but only made comparisons with two algorithms, NB and LR, which were prevalent then. This study compares three additional algorithms: Decision Tree (DT), SVM, and Random Forest (RF). These three additional algorithms have their respective advantages that can cover the weaknesses of the Naïve Bayes algorithm, which in previous studies had the highest level of accuracy, namely 91.67%. In addition to adding algorithms for comparison, this study also uses the Random Under Sampler technique to overcome class imbalance, although this method had not been employed in earlier investigations.

Previous research using the same data using the SVM algorithm has an accuracy rate of 87%, this research uses a sample data of 59 data and four attributes without using the random under sampler technique and data processing using python (Arifin et al., 2021). Previous research using cervical cancer risk classification data with feature selection based on expert interviews used the Extreme Learning Machine algorithm to classify and measure using the Confusion Matrix curve, resulting in an accuracy of 91.76% (Hidayah et al., 2019). Previous research used the Decision Trees algorithm. The results of the accuracy error in the study were 0% using 19 attributes, and the data was hospital patient data. Dr. Wahidin Sudirohusodo Makassa, in this study, used symptoms and signs to determine the stage of cervical cancer suffered by patients (Irmayani1, 2017).

The medical lens is typically used in cervical cancer studies, not a lifestyle perspective. In this study, the authors used data from habits carried out in everyday life. Of course, the information was collected from both people with and without cervical cancer.

RESEARCH METHODS

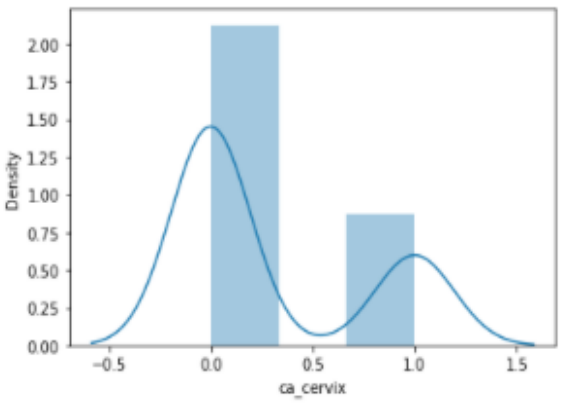
In this study, the CRISP-DM (Cross Industry Standard Process for Data Mining) model was used, which consisted of six stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, And Deployment (Firqiani, Kustyo, & Giri, 2008; Hasanah, Soim, & Handayani, 2021; Matovani & Hadiono, 2018).

A. Stages of business understanding

Based on data from the UCI machine learning repository with a total of 72 respondents, 22 were cancer patients, and 50 were cancer survivors. All respondents were residents of a city in Jakarta, Indonesia. Examining sufferers must be done so that the disease can be detected early to reduce the risk of cervical cancer. Using data mining with classification algorithms with a high level of prediction and accuracy can help overcome these problems so that the diagnostic results obtained are accurate. This study used algorithm comparisons to obtain high accuracies, such as logistic regression, naïve Bayes, SVM, decision trees, and random forests.

B. Data understanding stage

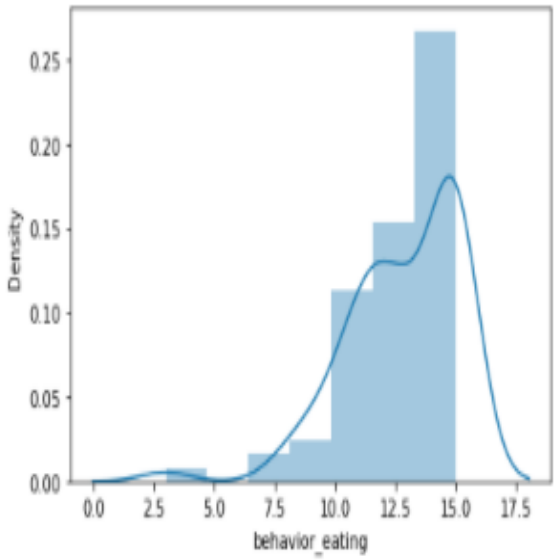
The data used are secondary data obtained from the survey results of cervical cancer patients, and the data comes from a questionnaire distributed to 72 respondents, of which 22 are cancer patients and 50 are Not cancer survivors. All respondents are residents of cities in Jakarta, Indonesia, which can be accessed publicly through UCI machine learning repositories. Data consists of 19 attributes and one attribute as a class.



Source: (Buani & Suryani, 2022)

Figure 1. Visualization of Ca Cervix Variables

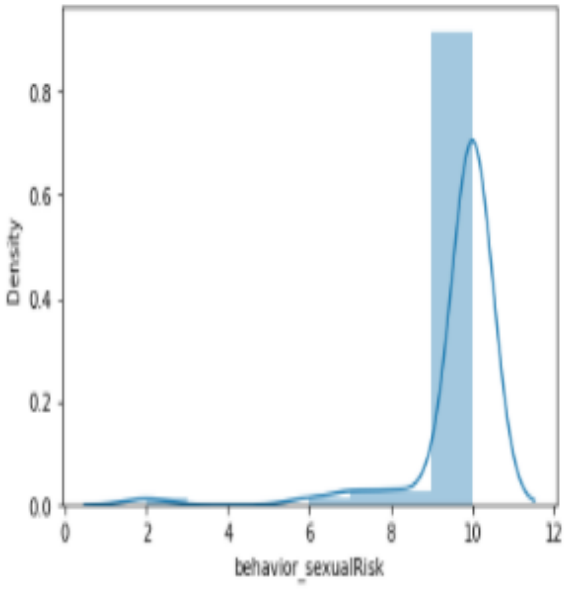
The Ca Cervix variable is a label variable used to classify having cervical or no cervical cancer.



Source: (Buani & Suryani, 2022)

Figure 2. Visualization of Eating Behavior Data

The eating Behavior Variable is a variable that describes the consumption of food. The food consumed is very influential on the health of the body.



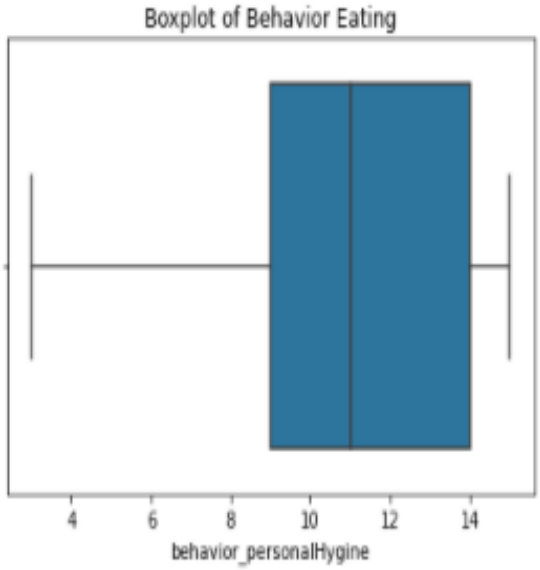
Source: (Buani & Suryani, 2022)

Figure 3. Visualization of Behavioral Sexual Risk Variable

Behavioral Sexual is a variable that most likely determines whether a person has cervical cancer.

C. Stages of data preparation

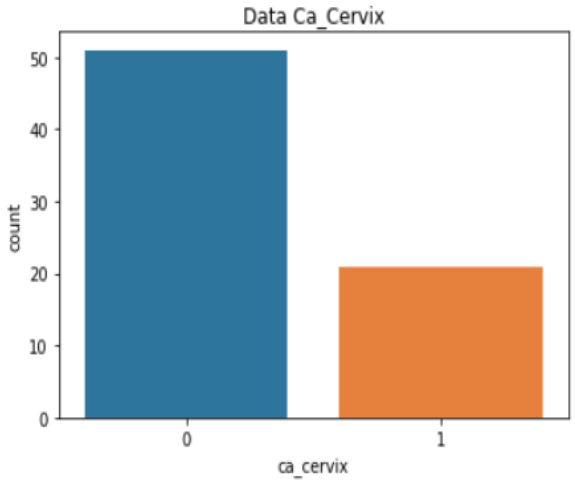
The total data in this study was 72, which already has a label where respondents have a risk of cancer and respondents who do not. However, this data still contains duplicate data, outliers, and anomalous or inconsistent data. Therefore, this stage is necessary to obtain quality data to produce a more effective and efficient model. An example of outlier data can be seen in Figure 4 below:



Source: (Buani & Suryani, 2022)

Figure 4. Eating Behavior Variable

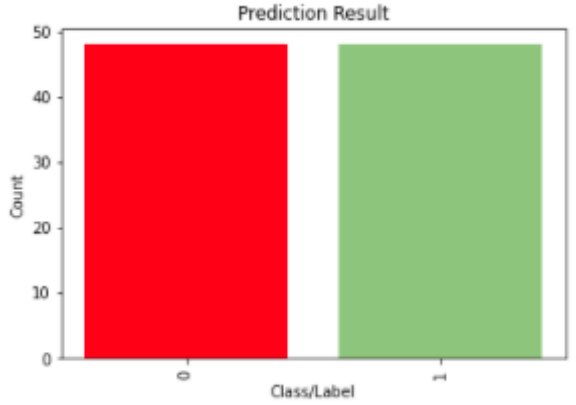
Figure 4 shows that the Behavior Eating Variable has no outlier data, so no data is far from observation.



Source: (Buani & Suryani, 2022)

Figure 5. Distribution of Ca-Cervix Data

In figure 5, If blue is data labeled no cervical cancer and the orange bar is cervical cancer. From the figure, the data must equate first. Here the author uses the Random Under Sampler technique. Then the result is as in Figure 6.

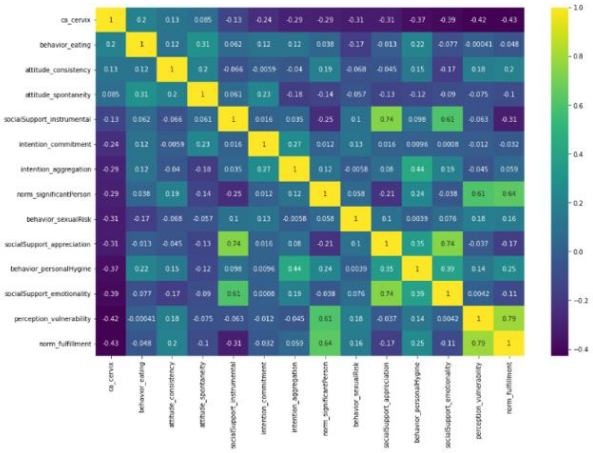


Source: (Buani & Suryani, 2022)

Figure 6. Data Distribution after Random Under Sampler

Figure 6 shows the use of the Random Under Sampler technique data distribution. The Random Under Sampler deals with classes/labels that are not identical in number. In the Random Under Sampler process, data is the division into testing data and training data, and the data division is 70% for training data and 30% for testing data.

After the Random Under Sampler, perform techniques and data in the same class. The next thing to do is look at the correlation between attributes or variables using HeatMap, seen in Figure 7.



Source: (Buani & Suryani, 2022)

Figure 7. HeatMap Correlation Between Variables

Figure 7 shows that the darker the heatmap color, the more attributes or variables have a stronger association with data classes or labels if and only if. Figure 3 shows that the variables/attributes are Behavioral Sexual Risk, Commitment of Intentions, Aggregation of Intentions, Norm Significant People, appreciative social support, self-hygiene behavior, perceptual susceptibility, and fulfillment of norms. Eight variables/attributes are the attributes that most influence the class/label of cervical cancer.

A. Dataset

The dataset used in this study is public data from the UCI Machine Learning Repository with a total of 72 data, data consisting of 20 attributes, and one attribute is a label, which sees in table 2:

Table 2. Data Description

Variable
behavior_sexualRisk
behavior_eating
behavior_personalHygine
intention_aggregation
intention_commitment
attitude_consistency
attitude_spontaneity
norm_significantPerson
norm_fulfillment
perception_vulnerability
perception_severity
motivation_strength
motivation_willingness
socialSupport_emotionality
socialSupport_appreciation
socialSupport_instrumental
empowerment_knowledge
empowerment_abilities
empowerment_desires
ca_cervix (this is a class attribute, 1=have cervical cancer, 0=no cervical cancer)

Source: (Sobar et al., 2016)

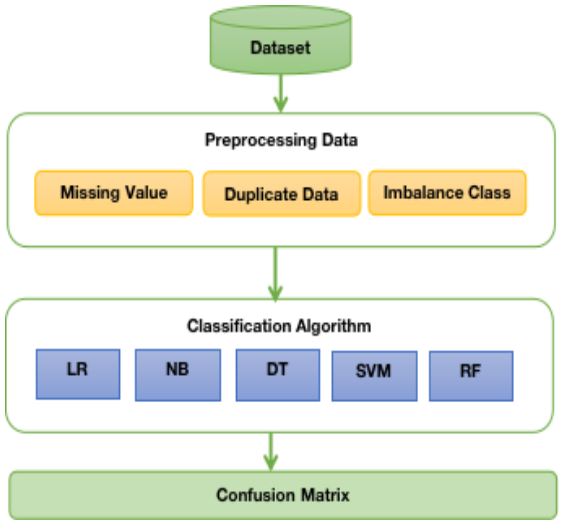
Table 2 shows the variables or attributes used in the study. These attributes include Behavioral Sexual Risk, Behavior Eating, Behavior Personal Hygiene, Aggregation of Intentions, Commitment of Intentions, Attitude Consistency, Attitude Spontaneity, Norm Significant People, Norm Fulfillment, Perceptual Vulnerability, Perceived Severity, Motivation strength, volitional motivation, emotional, social support, appreciative social support, instrumental social support, empowerment knowledge, empowerment ability, empowerment desire and ca_cervix (these are class attributes, 1=cervical cancer, 0=no cervical cancer).

B. Research Methods

The model used in this study is the Application of Machine Learning Algorithms for the early detection of cervical cancer. The algorithms used include Decision Tree (DT), SVM, Random Forest (RF), and two algorithms from previous studies, NB and LR. Then from the five algorithms selected the

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i1.453
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

best model, the results are accurate. Figure 1 shows the flow chart of this study.



Source: (Buani & Suryani, 2022)

Figure 8. Research Flowchart

Figure 8 shows research from data preparation, such as checking blank data, duplicate data, and imbalance classes.

RESULTS AND DISCUSSION

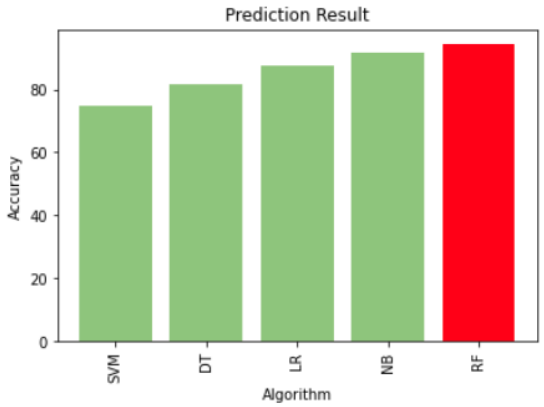
This study compared five algorithms: Logistic Regression, Naïve Bayes, SVM, Decision Tree, and Random Forest. For LR and NB, the accuracy results come from previous studies. SVM is a classification algorithm whose level of accuracy in a model depends on the kernel functions and parameters used, and the advantage of SVM is that it can classify and overcome regression with linear and non-linear (Dasmasela, Tomasouw, & Leleury, 2021)(Parapat, Furqon, & Sutrisno, 2018). Decision trees are prediction model techniques that use for task classification and prediction (Bahri & Lubis, 2020)(Wijaya, Bahtiar, Kaslani, & R, 2021)(Wuryani & Agustiani, 2021)(Schonlau & Zou, 2020). The results of the study can be seen in table 3.

Table 3. Predicted Results

Algorithm	Accuracy	AUC
NB	91.67%	0.96
LR	87.5%	0.97
DT	81.81%	0.81
RF	94.45%	0.75
SVM	75%	0.75

Source: (Buani & Suryani, 2022)

Table 3 is the result of the research conducted in this study. Table 3 describes the accuracy results after conducting experiments where Naïve Bayes (NB) has an accuracy of 91.67%, Logistic Regression (LR) of 87.5%, Decision Tree (DT) of 81.81%, SVM 75%, and Random Forest (RF) 94.45%, from the table. It shows that the highest accuracy in this study is RF.



Source: (Buani & Suryani, 2022)

Figure 9. Graph of Prediction Results

Figure 9 is a visualization of the prediction results made by NM, LR, DT, RF, and SVM from the image showing that Random Forest is the best algorithm in making predictions with a result of 94.45%

CONCLUSIONS AND SUGGESTIONS

The results of this research using the Random Under Sampler technique show that the model using the SVM algorithm is 75%, while the results of the model using the Decision Tree algorithm are 82%. For the model with the Random Forest algorithm, 94% of the results indicate that the random forest is a random forest model. The best method for early detection of cervical cancer in behavioral determinants.

REFERENCES

Agustyawati, D. N., Fauzi, H., & Pratondo, A. (2021). Perancangan Aplikasi Deteksi Kanker Serviks Menggunakan Metode Convolutional Neural Network. *EProceedings of Engineering*, *8*(4), 3908–3924.

Aisah, S. N., Hafiyusholeh, M., & Ulinnuha, N. (2022). Klasifikasi Kanker Serviks Menggunakan Metode Extreme Learning Machine (ELM). *Komputek*, *6*(3), 68–75. Retrieved from https://studentjournal.umpo.ac.id/index.php/komputek/article/view/68

Arifin, S. S., Siregar, A. M., Ratna, A., & Mudzakir, T.

Al. (2021). *Klasifikasi Penyakit Kanker Serviks Menggunakan Algoritma Support Vector Machine (SVM)*. (Ciastech), 521–528.

Bahri, S., & Lubis, A. (2020). Metode Klasifikasi Decision Tree Untuk Memprediksi Juara English Premier League. *Jurnal Sintaksis*, *2*(1), 63–70. Retrieved from http://www.jurnal.stkipalmaksum.ac.id/index.php/Sintaksis/article/view/47

Buani, D. C. P., & Suryani, I. (2022). *Independent Research Report*. Jakarta.

Dasmasela, R., Tomasouw, B. P., & Leleury, Z. A. (2021). Penerapan Metode Support Vector Machine (SVM) untuk Mendeteksi Penyalahgunaan Narkoba. *Matematika, Statistik Dan Terapannya*, *1*(02), 93–101.

Feblian, D., & Daihani, D. U. (2017). Implementasi Model Crisp-Dm Untuk Menentukan Sales Pipeline Pada Pt X. *Jurnal Teknik Industri*, *6*(1). https://doi.org/10.25105/jti.v6i1.1526

Firqiani, H. N., Kustyo, A., & Giri, E. P. (2008). Seleksi Fitur Menggunakan Fast Correlation Based Filter pada Algoritma Voting Feature Intervals 5. *Jurnal Ilmiah Ilmu Komputer*, *6*(2), 245184.

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

Hidayah, U. R., Cholissodin, I., & Adikara, P. P. (2019). Klasifikasi Penyakit Kanker Serviks dengan Extreme Learning Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(7), 6575–6582. Retrieved from http://j-ptiik.ub.ac.id

Irmayani1, B. A. (2017). Klasifikasi Stadium Kanker Serviks Menggunakan Sistem Pengambilan Keputusan Decision Tree. *Prosiding Seminar Nasional*, *04*(1), 455–464. Retrieved from http://journal.uncp.ac.id/index.php/proceding/article/view/1281

Matovani, D., & Hadiono, K. (2018). Implementasi Algoritma Apriori Untuk Membantu Proses Persediaan Barang. *Jurnal Dinamika Informatika*, *10*(2), 53–59. https://doi.org/10.35315/informatika.v10i2.8133

Pangribowo, S. (2019). Beban Kanker di Indonesia. *Pusat Data Dan Informasi Kesehatan Kementerian Kesehatan RI*, 1–16.

Parapat, I. M., Furqon, M. T., & Sutrisno. (2018). Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(10), 3163–3169.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Setyani, R. A. (2018). Penerapan Program Deteksi Dini Kanker Serviks Sebagai Upaya Pemberdayaan Wanita Di Sleman Yogyakarta. *Kebidanan, Fakultas Ilmu Kesehatan Universitas Respati Yogyakarta*, *III*(2), 12.

Sobar, Machmud, R., & Wijaya, A. (2016). Behavior determinant based cervical cancer early detection with machine learning algorithm. *Advanced Science Letters*, *22*(10), 3120–3123. https://doi.org/10.1166/asl.2016.7980

Wijaya, Y. A., Bahtiar, A., Kaslani, & R, N. (2021). Analisa Klasifikasi menggunakan Algoritma Decision Tree pada Data Log Firewall. *Jurnal Sistem Informasi Dan Manajemen*, *9*(3), 256–264. https://doi.org/10.47024/JS.V9I3.303

Winarni, W., & Suratih, K. (2020). Mengenal Lebih Dini Kanker Leher Rahim Bersama Forum Kajian Dan Komunikasi Muslimah. *GEMASSIKA : Jurnal Pengabdian Kepada Masyarakat*, *4*(2), 186. https://doi.org/10.30787/gemassika.v4i2.569

Wongkar, R., Angka, R. N., & Angeline, R. (2022). Karakteristik Pasien Kanker Stadium 4 yang Mendapatkan Perawatan Paliatif di Rumah Sakit X. *Jurnal Kedokteran Meditek*, *28*(2), 126–132. https://doi.org/10.36452/jkdoktmeditek.v28i2.2235

Wuryani, N., & Agustiani, S. (2021). Random Forest Classifier untuk Deteksi Penderita COVID-19 berbasis Citra CT Scan. *Jurnal Teknik Komputer*, *7*(2), 187–193. https://doi.org/10.31294/jtk.v4i2