Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11792

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Stratified K-fold cross validation optimation on machine learning for prediction

Slamet Widodo[1)*], Herlambang Brawijaya[2)], Samudi[3)]
[1)2)]Universitas Bina Sarana Informatika, Indonesia, [3)]Universitas Nusa Mandiri, Indonesia
[1)]slamet.smd@bsi.ac.id, [2)]herlambang.hba@bsi.ac.id, [3)]samudi.smx@nusamandiri.ac.id

Submitted : Sep 12, 2022 | **Accepted** : Sep 28, 2022 | **Published** : Oct 3, 2022

Abstract: Cervical is the second most common malignant tumor in women, with 341,000 deaths worldwide in 2020, almost 80% of which occur in developing countries. One of the causes is infection with Human papillomavirus (HPV) types 16 and 18. The increasing incidence of cervical cancer in Indonesia makes this disease must be treated seriously because it is one of the main causes of death. In addition to the virus, external factors can be one of the causes. The high mortality rate in patients is caused by the patient's awareness of the emergence of cervical cancer which is only seen when it enters the final stage. One of the efforts to reduce the number of sufferers is to implement cervical cancer detection. Early detection of cervical cancer can also be identified by looking at external factors, such as behavioral factors, intentions, attitudes, norms, perceptions, motivations, social support, and empowerment. However, the data used has an imbalance in the distribution of the target class, namely more negative samples than positive ones. To overcome this, a technique called Stratified K-Fold Cross-Validation (SKCV) is used. Evaluation of the accuracy value using the Confusion matrix to determine the performance of each model. The best performance of the five classification algorithms used is 96 percent (RF), 94 percent (LR), 92 percent (XGBoost), 90 percent (KNN), and 88 percent (NB). The results show that the model formed by RF-based SKCV has the highest accuracy of other models.

Keywords: KNN, LR, NB, RF,Serviks, SKCV, XGBoost

INTRODUCTION

Cervical cancer is the second most common malignant tumor in women, with 341,000 deaths worldwide in 2020, almost 80% of which occur in developing countries (Sung et al., 2021). 90% of cervical cancer cases are associated with oncogenic HR-HPV virus infection. Before the occurrence of cancer, will be preceded by a condition called precancerous lesions or cervical intraepithelial neoplasia (NIS). One of the causes is infection with persistent Human papillomavirus (HPV) types 16 and 18 (Chen et al., 2021; Wang et al., 2021). The increasing incidence of cervical cancer in Indonesia makes this disease must be taken seriously because it is one of the main causes of death. In 2020, the World Health Organization (WHO) noted, that the incidence of cervical cancer in Indonesia was 36,633 new cases, or around 9.2% of all cancer cases, this is in second place after breast cancer (The Global Cancer Observatory 2020). In addition to viruses, factors that influence the incidence of cervical cancer include age, parity, education, use of hormonal family planning, smoking, hygiene, physical activity, place of residence, and heredity, most of these factors are modifiable risk factors so that prevention efforts can carry out (Setianingsih et al., 2022). The high mortality rate in patients with cervical cancer is due to the patient's awareness of the appearance of cervical cancer which is only seen when entering an advanced or late stage, however, prevention remains the best option to reduce mortality due to this disease (Guimarães et al., 2022). One of the efforts to reduce the number of sufferers is to implement cervical cancer detection.

The use of machine learning tools in medical diagnosis is increasing. By applying data mining and computational techniques to extract rules and patterns from various data sets. Some of these techniques have shown good results in data mining problems, this is very helpful for the medical profession to detect a disease. Early and accurate detection and identification of risk factors are the main goals of data mining applications in diagnostics. The results of this procedure form the basis for choosing the right treatment (Witten, 2017). Currently, effective machine learning techniques help develop several techniques that can be used to diagnose diseases that are widely used to identify a disease (Abdualgalil et al., 2022). Early detection of cervical cancer can also be identified by looking at external factors, such as behavioral factors, intentions, attitudes, norms, perceptions, motivations, social support, and empowerment as has been done (Sobar, Machmud, and Wijaya

*name of corresponding author

2016) using Naive Bayes (NB) and Logistic Regression (LR) results show that NB predicts very well with an accuracy of more than 92%. Classification techniques such as Naive Bayes (NB), Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimation (SMO), Random Forest (RF), Multilayer Perceptron (MLP), and Simple Logistic Regression (SLR) used to classify cervical cancer datasets. The classification performance was evaluated using 10 fold cross validation where accuracy, precision, and recall as evaluation metrics were measured using a confusion matrix to determine the strength of the performance of all classification techniques. As a result, RF has achieved the highest level of accuracy, precision, and recall compared to the other six classification algorithms while NB has the lowest (Razali et al., 2020). Based on the above, this study proposes to evaluate the early detection of cervical cancer based on behavioral risk data using data mining techniques. However, the data used has an imbalance in the distribution of the target class, namely more negative samples than positive ones. To overcome this, a technique called Stratified K-Fold Cross-Validation (SKCV) has been proposed (Prusty et al., 2022). SKCV is used to ensure that relative class frequencies are effectively maintained across each series and fold validation when using stratified sampling rather than random sampling. This technique is widely used for classification problems. This method uses stratified sampling, which divides the data set into k groups or folds, of nearly equal size. The use of a random subset of data in cross-validation, also known as k-fold cross-validation, is a powerful way to test the success rate of models used for medical data classification.

In this study, the analysis was carried out using SKCV. The confusion matrix is used to evaluate the accuracy value by using several machine learning techniques to determine the performance of the machine. Machine learning used is Naïve Bayesian (NB), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (xgboost) and K-nearest neighbors (KNN). From the results of testing and evaluation, it is hoped that it can be a benchmark in the development of machine learning that can provide recommendations about the occurrence of cervical cancer in someone based on their behavior or habits.

.

LITERATURE REVIEW

Machine Learning

Machine learning provides the technical basis for data mining. It is used to extract information from raw data into a database. Generally in a form that can be understood and can be used for various purposes. In many applications, machine learning derives an image of the shape from an example. The type of description found can be used for prediction, explaining what and why, and understanding. Some data mining applications focus on prediction: predicting what will happen in a new situation from data that describes what happened in the past, by guessing from examples (Witten, 2017). Different types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning are in the area. In addition, deep learning is part of a broader machine learning method, able to intelligently analyze data at a large scale (Sarker, 2021).. In practice, machine learning means applying computer algorithms whose weights are adjusted as new data comes in. Machine learning algorithms learn from data sets to make predictions about species classification, stock markets, corporate profits, decision-making, subatomic particles, optimal traffic routes, and more. Machine learning is the best tool for turning big data into accurate prediction tools (Wade, 2020)

Random Forest (RF)

Random forest is a popular machine learning procedure that can be used to develop predictive models. First introduced by Breiman in 2001, RF is a collection of classification and regression trees, which are simple models using binary separation of predictor variables to determine outcome predictions. The advantage of using RF for predictive modeling is the ability to handle data sets with a large number of predictor variables (Speiser et al., 2019). RF is a substantial modification of bagging that builds a large collection of uncorrelated trees and then averages them. In many cases, RF performance is very similar to optimation and is easier to train and adjust. As a result, RF has become popular and widely implemented (Genuer & Poggi, 2020).

*name of corresponding author

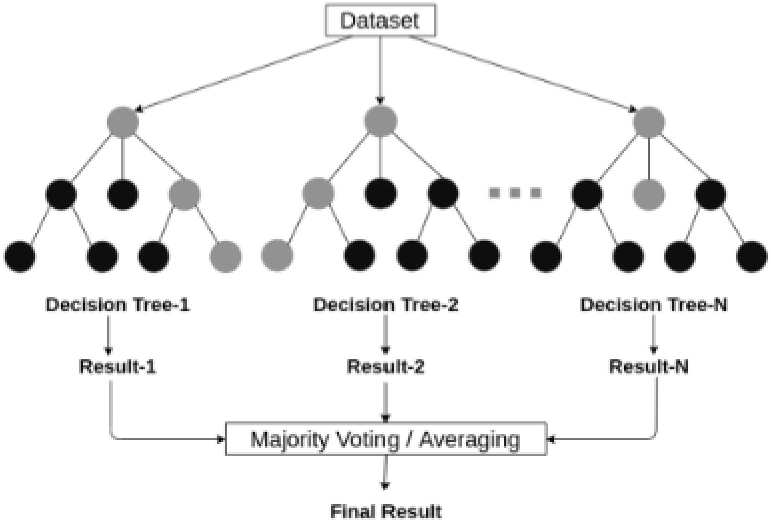


Fig. 1 RF with multiple decision trees (Sarker, 2021)

Logistic Regretion(LR)

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio independent variables (Cleophas & Zwinderman, 2018). LR is the most basic classification algorithm. Mathematically, LR works in a similar way to linear regression. For each column, LR finds the appropriate weights or coefficients, which can maximize model accuracy (Wade, 2020). The sigmoid function used in the LR is as follows:

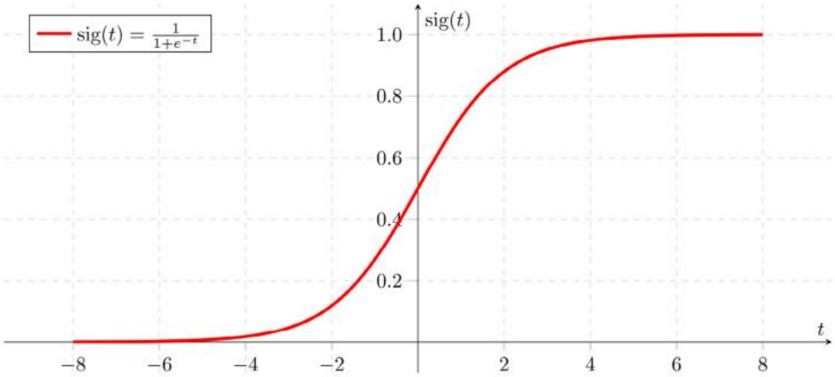


Fig. 2 Sigmoid function (Wade, 2020)

The sigmoid function is usually used for classification. All values greater than 0.5 are matched to 1, and all values less than 0.5 are matched to 0.

Extreme Gradient Boosting(xgboost)

XGBoost adalah kependekan dari Extreme Gradient Boosting.Ekstrim dimaksudkan untuk mendorong batas komputasi untuk mencapai kelebihan dalam akurasi dan kecepatan. XGBoost mengintegrasikan prediksi pengklasifikasi "lemah" (tree model) untuk mencapai pengklasifikasi "kuat" (tree model) melalui proses pelatihan serial. Hal ini dapat menghindari over-fitting dengan menambahkan istilah regularisasi. Komputasi paralel dan terdistribusi membuat proses pembelajaran lebih cepat untuk memberikan proses pemodelan yang lebih cepat (Mo et al., 2019). Gambar 3 skema dari proses komputasi XGBoost:

XGBoost stands for Extreme Gradient Boosting. Extreme is meant to push the limits of computing to achieve excellence in accuracy and speed. XGBoost integrates the predictions of the "weak" classifier (tree model) to

*name of corresponding author

achieve the "strong" classifier (tree model) through a serial training process. It can avoid over-fitting by adding regularization terms. Parallel and distributed computing make the learning process faster to provide a faster modeling process (Mo et al., 2019). Figure 3 schematic of the XGBoost computing process:

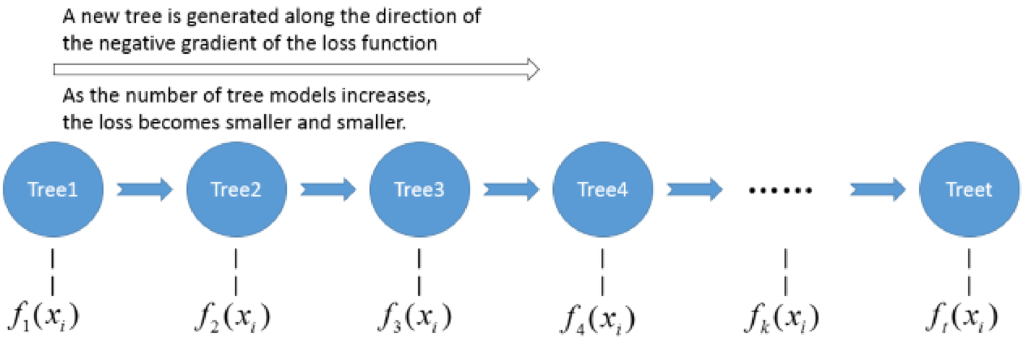

A new tree is generated along the direction of the negative gradient of the loss function
As the number of tree models increases, the loss becomes smaller and smaller.
Tree1
Tree2
Tree3
Tree4
......
Treet
$f_1(x_i)$
$f_2(x_i)$
$f_3(x_i)$
$f_4(x_i)$
$f_k(x_i)$
$f_t(x_i)$

Fig. 3 xgboost algorithm schematic (Mo et al., 2019)

where fk(xi) represents the model tree.

K-nearest neighbors(KNN)

K-Nearest neighbor is an example of instance-based learning, where the training data set is stored so that unclassified classification records can be found by comparing them with the most similar records in the training set (Larose & Larose, 2015). In instance-based classification, each new instance is compared to an existing one using the distance metric, and the closest existing instance is used to assign a class to the new one. This is called the nearest neighbor classification method. Sometimes more than one nearest neighbor is used, and the majority class of the k nearest neighbors (or distance-weighted average, if the class is numeric) is assigned to a new instance (Witten, 2017). Classification is calculated from the majority of the k nearest neighbors of each point. This algorithm is quite good for training noise data, and the accuracy depends on the quality of the data. The biggest problem with KNN is the consideration of choosing the optimal number of neighbors. KNN can be used for both classification and regression (Sarker, 2021).

Naïve Bayes (NB)

A naive Bayes algorithm is a probability method that could happen in the future, based on the Bayes theorem with the assumption of independence between each pair of features. NB has been used as an effective classifier for many years. Unlike many other classifiers, NB is very easy to construct, because as a priori, it does not require structured learning procedures (Gorunescu, 2011).

K-Fold Stratified Cross Validation

Stratified K-Fold Cross-Validation (SKCV) is an extension of Cross-Validation (CV). The SKCV guarantees that each class is evenly distributed throughout the k-fold (Allen et al., 2021). In other words, the datasets are not randomly distributed into k-folds, but in a way that does not interfere with the sample distribution ratio across the classes within SKC. SKCV was proposed by (Prusty et al., 2022) to ensure that relative class frequencies are effectively maintained across each train and validation fold when using stratified sampling rather than random sampling. This method uses stratified sampling, which divides the cervical cancer data set into k groups or folds of nearly equal size. As a result, SKCV is preferred over CV for dealing with classification problems with unequal class distributions. this CV object returns nested folds and is a variant of K-Fold. Folding is achieved by keeping the sample fraction with each class constant.

Confusion matrix

The confusion matrix provides decisions obtained in training and testing, the confusion matrix provides an assessment of classification performance based on objects correctly or incorrectly (Gorunescu, 2011). The confusion matrix contains actual (actual) and predicted (predicted) information on the classification system, the test sequence is tabulated in a confusion matrix where the predicted class is displayed at the top of the matrix and the observed class is on the left. Each cell contains a number indicating how many actual cases of the observed class to predict.

METHOD

The data used in this study utilizes a secondary dataset entitled Cervical Cancer Behavioral Risk Cervical Dataset (Sobar et al. 2016). The dataset was obtained from the UCI machine learning repository accessed via the web https://archive.ics.uci.edu/ml/machine-learning-databases/00537/. This dataset has 72 data with 18 attributes (columns) and an attribute that is a label containing the number 0 (zero) as a patient without cervical cancer and 1 (one) representing a patient with cervical cancer. From the dataset, then it is divided into training and testing data. SKCV is used for the five model classifications. To evaluate the prediction results, the accuracy measurement uses a confusion matrix.

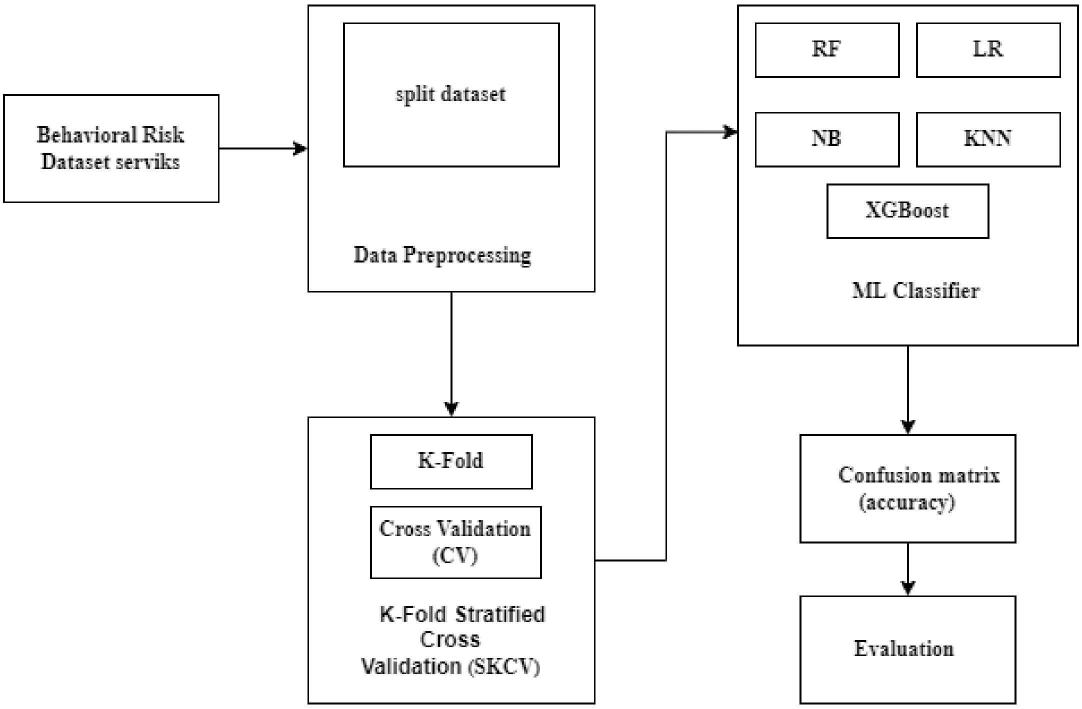

Behavioral Risk Dataset serviks
split dataset
Data Preprocessing
K-Fold
Cross Validation (CV)
K-Fold Stratified Cross Validation (SKCV)
RF
LR
NB
KNN
XGBoost
ML Classifier
Confusion matrix (accuracy)
Evaluation

Fig. 4 Process design for predicting cervical cancer.

RESULT

In this research, experimental research is used. Experimental research involves investigating causal relationships using tests controlled by the researcher himself. The data used in this study is data taken through a questionnaire from the Jakarta population, with as many as 72 respondents consisting of 21 patients who were detected with cervical cancer, while 51 other patients did not get cervical cancer. The data consists of 18 attributes with 7 behavioral determinants of cervical cancer. There are variables that are classified as predictive variables, namely variables that are used as determinants of cervical cancer, and objective variables, namely variables that are used as disease outcomes. The predicting variables are: 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, empowerment_details, empowerment_emotionality_appreciation. behavioral factors can be seen in table 1.

From the results of several experiments, the distribution of training data is 70 percent and testing is 30 percent with the parameter value set random_state = 4 (controls how the data is scrambled before being split) and StratifiedKFold cross validation n_split = 10, the most optimal results are obtained with an average accuracy of more than 90%, the accuracy results can be seen in Figure 5.

Table 1. List of Attributes of the Cervical Cancer Behavioral Risk Dataset

Determinant (variable)	**attribute**
Perception	Eating
	Personal hygine
intention	aggregation

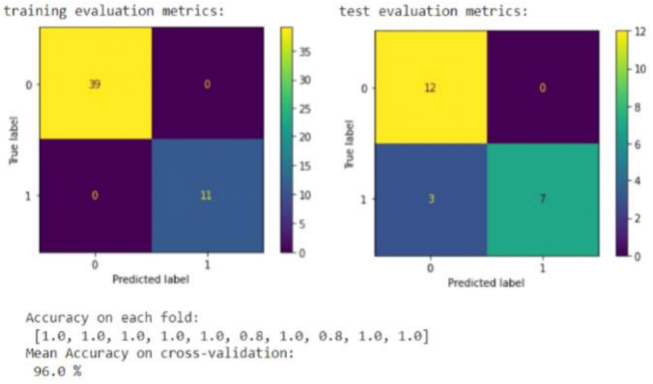
*name of corresponding author

	commitment
attitute	consistency
	spontaneity
norm	significant person
	fulfillment
perception	vulnerabilit
	severity
motivation	strength
	willingness
socialSupport	emotionality
	appreciation
	instrumental
empowerment	knowledge
	abilities
	desires

Table 2. Accuracy value with Stratified K-Fold Cross-Validation

Model	***CV mean Accuracy***	***Test Accuracy***	***Train Accuracy***
Random Forest	96.0%	86.36%	100.0%
Logistic Regression	94.0%	90.91%	100.0%
Extreme Gradient Boosting	92.0%	81.82%	100.0%
K-nearest neighbors	90.0%	81.82%	94.0%
Gaussian Naïve Bayes	88.0%	90.91%	94.0%

CV is widely used in ML to measure how well a model performs on untrained data. The application of the SKCV technique provides equal opportunity for each data point to be included in the test set by dividing it into k equal parts. Thus, it helps in reducing computation time, bias, and variance as the value of k increases. The table shows the best performance of the five supervised machine learning algorithms implemented with SKCV, the classification prediction results with an average accuracy of 96 percent (RF).



```
Accuracy on each fold:
 [1.0, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 0.8, 1.0, 1.0]
Mean Accuracy on cross-validation:
 96.0 %
```

Fig. 5 RF evaluation using the Confusion matrix

Based on Figure 5 of the confusion matrix training and testing above, it can be seen that 18 cases were correctly classified as cervical cancer (tp) out of a total of 21 cases, 3 cases were misclassified, which should have been misclassified as cervical cancer instead of cervical (fn), 51 cases were correctly classified as non-cervical. cervical cancer (tn) from a total of 51 cases while 0 cases in (fn). The accuracy value can be calculated in the following equation:

$$Accuracy = \frac{tp+tn}{tp+fn+tn+fp}$$

$$Accuracy = \frac{18+51}{18+3+51+0} = 0,958$$

*name of corresponding author

DISCUSSIONS

This study proposes to evaluate the early detection of cervical cancer based on risk behavior using data mining techniques. However, the data used has an imbalance in the distribution of the target class, namely more negative samples than positive ones. To overcome this, a technique called Stratified K-Fold Cross-Validation (SKCV) is used. Evaluation of the accuracy value using the Confusion matrix to determine the performance of each model. From the experimental results obtained the best performance using the classification algorithm (RF) with an accuracy of 96%.

CONCLUSION

The conclusion of this study is that SKCV has a good performance in prediction. This can be seen from the five classification algorithms which in the data have an unbalanced class distribution, with the RF algorithm model achieving an accuracy of 96%, LR 94%, XGboost 92%, KNN 90% and NB 88%. The results show that the model formed by RF-based SKCV has the highest accuracy compared to other models.

REFERENCES

Abdualgalil, B., Abraham, S., & M. Ismael, W. (2022). An Efficient Machine Learning Techniques as Soft Diagnostic for Tuberculosis Classification Based on Clinical Data. *Journal of Scientific Research*, *66*(02), 61–67. https://doi.org/10.37398/jsr.2022.660209

Allen, J., Liu, H., Iqbal, S., Zheng, D., & Stansby, G. (2021). Deep learning-based photoplethysmography classification for peripheral arterial disease detection: A proof-of-concept study. *Physiological Measurement*, *42*(5). https://doi.org/10.1088/1361-6579/abf9f3

Chen, C.-J., You, S.-L., Hsu, W.-L., Yang, H.-I., Lee, M.-H., Chen, H.-C., Chen, Y.-Y., Liu, J., Hu, H.-H., Lin, Y.-J., Chu, Y.-J., Huang, Y.-T., Chiang, C.-J., & Chien, Y.-C. (2021). *Epidemiology of Virus Infection and Human Cancer BT - Viruses and Human Cancer: From Basic Science to Clinical Prevention*. https://doi.org/10.1007/978-3-030-57362-1_2

Cleophas, T. J., & Zwinderman, A. H. (2018). Regression Analysis in Medical Research. In *Regression Analysis in Medical Research*. Springer International Publishing. https://doi.org/10.1007/978-3-319-71937-5

Genuer, R., & Poggi, J.-M. (2020). Random Forests with R. In *Use R*. Springer International Publishing. https://doi.org/10.1007/978-3-030-56485-8

Gorunescu, F. (2011). *Data Mining* (Vol. 12). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19721-5

Guimarães, Y. M., Godoy, L. R., Longatto-Filho, A., & Dos Reis, R. (2022). Management of Early-Stage Cervical Cancer: A Literature Review. *Cancers*, *14*(3). https://doi.org/10.3390/cancers14030575

Kusuma, E. J., Nurmandhani, R., & Handayani, S. (2021). JPKM Jurnal Profesi Kesehatan Masyarakat. *Jpkm*, *2*(1), 1–8.

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics* (second). Wiley.

Mo, H., Sun, H., Liu, J., & Wei, S. (2019). Developing window behavior models for residential buildings using XGBoost algorithm. *Energy and Buildings*, *205*, 1–23. https://doi.org/10.1016/j.enbuild.2019.109564

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*. https://doi.org/10.3389/fnano.2022.972421

Razali, N., Mostafa, S. A., Mustapha, A., Wahab, M. H. A., & Ibrahim, N. A. (2020). Risk Factors of Cervical Cancer using Classification in Data Mining. *Journal of Physics: Conference Series*, *1529*(2). https://doi.org/10.1088/1742-6596/1529/2/022102

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Setianingsih, E., Astuti, Y., & Aisyaroh, N. (2022). Literature Review : Faktor-Faktor Yang Mempengaruhi Terjadinya Kanker Serviks. *Jurnal Ilmiah PANNMED (Pharmacist, Analyst, Nurse, Nutrition, Midwivery, Environment, Dentist)*, *17*(1), 47–54. https://doi.org/10.36911/pannmed.v17i1.1231

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

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*(336), 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Wade, C. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn* (D. Sugarman (ed.)). Packt Publishing.

Wang, X., Wu, S., & Li, Y. (2021). Risks for cervical abnormalities in women with non-16/18 high-risk human papillomavirus infections in south Shanghai, China. *Journal of Medical Virology*, *93*(11), 6355–6361. https://doi.org/10.1002/jmv.27185

Witten, I. H. (2017). Data Mining (Fourth Edition). In *Practical Machine Learning Tools and Techniques*.