Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/ 10.33395/sinkron.v6i1.11173
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Classification of the Human Development Index in Indonesia Using the Bootstrap Aggregating Method

Noor Ell Goldameir[1)*], Anne Mudya Yolanda[2)], Arisman Adnan[3)] Lusi Febrianti[4)]
[1)2)3)4)] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Riau University, Indonesia
[1)]noorellgoldameir@lecturer.unri.ac.id, [2)]annemudyayolanda@lecturer.unri.ac.id,
[3)]arisman.adnan@lecturer.unri.ac.id [4)]lusi.febrianti3593@student.unri.ac.id

Submitted : Sep 30, 2021 | **Accepted** : Oct 10, 2021 | **Published** : Oct 10, 2021

Abstract: Successful development of the quality of human life in a region is determined by the Human Development Index (HDI). Human development performance based on the HDI can be measured: long and healthy life, knowledge, and a decent standard of living. The HDI is usually grouped into several categories to facilitate the classification of the HDI level of each region. This study aimed to determine the ability of the bootstrap aggregating (bagging) method to classify the HDI by district/city. Bagging is a stochastic machine learning approach that can eliminate the variance of the classifier by producing a bootstrap ensemble to obtain better accuracy results. The dependent variable in this study was the HDI by district/city in 2020. In contrast, life expectancy at birth, expected years of schooling, mean years of schooling, and real expenditure per capita are adjusted as independent variables. Bagging was applied to the high and low categories of HDI data. The bagging method demonstrated good classification performance due to only eight classification errors, namely the HDI data which should be in the high category but classified into the low category by the bagging method. Based on the results of calculations with 25 replications, it can be concluded that the bagging method has a very good performance, with an accuracy value of 92.3%, the sensitivity of 100%, and specificity of 83.33%. The bagging method is considered very good for the classifying the HDI by district/city in Indonesia in 2020 because it has a balanced accuracy of 91.67%.

Keywords: Adjusted real expenditure per capita; bootstrap aggregating; classification; expected years of schooling; human development index; mean years of schooling; life expectancy at birth

INTRODUCTION

National income and economic growth are important indicators of development. The purpose of the development is for the benefit of society. In addition, development should not only focus on economic aspects but also human development. This concept, the Human Development Index (HDI), was introduced by the United Nations Development Program (UNDP) in 1990 through a report entitled Human Development Report (HRD).(Badan Pusat Statistik, 2021c).

The HDI is an annual human development performance measure published by the Central Bureau of Statistics (CBS). HDI describes how people can access the country's development in terms of health, education, income, and other aspects of life. High and low HDI values depend on development programs that are continuously implemented and monitored. For HDI to be classified as good, the program's implementation must be consistent with the target and regional priorities based on the HDI categories within the region. Therefore, classification is needed to facilitate the classification of the HDI level.

Human development means expanding people's choices and abilities to live freely in dignity and achieve their desires. Human development also refers to changes in the welfare of society and is the goal of all kinds of development. Human development performance can be measured from three dimensions: long and healthy life represented by the Life Expectancy at birth (LE) indicator, knowledge represented by the Expected Years of Schooling (EYS), and Mean Years of Schooling (MYS) indicators, and a decent standard of living represented by

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021 e-ISSN : 2541-2019
DOI : https://doi.org/ 10.33395/sinkron.v6i1.11173 p-ISSN : 2541-044X

an adjusted indicator of real expenditure per capita. These indicators are grouped into single values known as the HDI. CBS divides HDI into four categories, namely low category (HDI <60), medium category (60≤HDI<70), high category (70≤HDI<80, very high category (HDI ≥80).

The HDI data are classified as categorical data, thus they can be classified using some methods. Several studies are conducted related to HDI, and the methods used are also varied, such as support vector machine, random forest, k-nearest neighbor, and others. This study classified HDI in Indonesia in 2020 based on District/City using the bagging method. Classification refers to the process of identifying an observation in a certain predetermined class. The HDI classification aims to divide District/City in Indonesia into several levels, which later will be served as a reference for development targets. For example, as an indicator in measuring the key performance of the Regional Incentive Fund (RIF) and determining the General Allocation Fund (GAF).

The analysis used bagging method. Bagging is a relatively simple but effective ensemble method and has been applied in many applications. The use of the bagging method is very helpful to overcome the instability of the parameters in the classification. The bagging method's basic idea is to repeatedly use random resampling with replacement to the original data to obtain new data to be used in classification analysis. This study resulted in the classification of HDI data and the accuracy of the prediction results.

LITERATURE REVIEW

Previous research on the HDI classification of districts/cities in Indonesia was carried out by (Mauludiyah, 2020) using the random forest method. This study used HDI data in 2018 and 4 attributes related to HDI in 514 districts/cities in Indonesia. HDI data is then divided into two parts: training data by 60% and testing data by 40%. The best accuracy value at m=2 and k=100 is 93.69%. The HDI in Indonesia in 2018 was 71.39 where the HDI level was very high at 6%, high at 32%, medium at 57%, and low at 5%.

Subsequent research was conducted by (Darsyah, 2014) discusses the classification of Central Java Province HDI in 2016 using the K-Nearest Neighbor (KNN) method. The results obtained are the average HDI of 70.61. The best performance results were obtained at values of k=5 and k=10 at an accuracy rate of 91.43%, sensitivity 100%, and specificity of 83.33%.

The next research was conducted by (F Fauzi, 2017) which discusses the HDI classification of Central Java Province in 2013-2014 by comparing the K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) methods. The average HDI of Central Java Province in 2013-2014 was 68.68. The results of the HDI classification using the K-Nearest Neighbor method obtained an accuracy of 91.64%, a sensitivity of 88.57%, and a specificity of 97.14% at values of k = 10, 15, and 20. Meanwhile, the SVM method obtained accuracy with the Radial Basis Function kernel (RBF) of 95.36%, the sensitivity of 97.14%, and the specificity of 91.43% with parameter values of gamma 1 and C = 1, 10, and 100. Based on the results, the accuracy of the SVM method is better than the K-NN method in this HDI case.

Study (Fatkhurokhman Fauzi et al., 2017) discussed the HDI classification of districts/cities across Indonesia in 2015. The method used is the Smooth Support Vector Machine (SSVM) kernel RBF. The accuracy rate is 100%.

A study (Muttaqin & Zulkarnain, 2020) discussed on classification of HDI in 514 districts/cities in Indonesia in 2018. This study used cluster analysis, k-means as the cluster analysis method used. There are three groups of the results of this study, namely high areas, medium areas, and low areas. The first group or low area consists of 19 cities. The second group or the middle area consists of 381 districts/cities. The third group or high area consists of 114 districts/cities.

A study (Verawaty, Muji Gunarto, Rolia Wahasusmiah, 2021) discussed the HDI of districts/cities in the province of South Sumatra Indonesia in 2014-2016. This study used multiple regression analysis. The study results showed that capital expenditure, general allocation funds, local own-source revenue, Revenue sharing funds affect HDI.

METHOD

Bagging Method

Bootstrap aggregating invented by Breimanin 1996 (Xindong Wu, 2009). Bagging is a method applied to classification and regression. The purpose of this method is to reduce the variance of the independent variables to improve the quality of the predictions. Bootstrap is a random sampling method with resampling. The results obtained through the bootstrap process are followed by an aggregation process to make combined predictions (Otok et al., 2020).

Bagging is an ensemble method using bootstrap sampling. The bagging method is used effectively to resolve the problem of unbalanced data classes (Issam H. Laradji, Mohammad Alshayeb, 2015). This method is used in the classification to separate training data into multiple new random sampling training data and build a new model based on training data (Wahono & Suryana, 2013).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/ 10.33395/sinkron.v6i1.11173
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Bagging is using training data manipulation. Then, the training data is duplicated d times sampling with replacement to produce d new data. Furthermore, classifiers are built which can be regarded as bagged classifiers (Naomi Altman, 2017). Bagging uses sampling with replacement, the size of the data is the same as the original data, but the distribution of data from each bagging data is different, some data from the training data may appear several times or may not appear at all (Galar et al., 2012). Bagging is a stochastic machine learning approach that can eliminate variance from the classifier by producing a bootstrap ensemble to obtain better accuracy results (Windridge & Nagarajan, 2017). Bagging can reduce the variance of the dataset by sampling with replacement (Mordelet & Vert, 2013).

Model Evaluation and Validation

Classification analysis is expected to all data can be classified correctly. However, sometimes there are some errors in classification. Validation and evaluation are needed to assess whether the model is correct or not. To avoid errors following the application of the model to new unknown data, namely overfitting, it is necessary to assess the model. Dividing the dataset into training data and test data is an approach taken to avoid overfitting. Training data is used to build the model and test data to validate the built model (Mohammed et al., 2017).

The matrix in Table 1 serves as a basis for comparison in identifying the optimal method based on the data that has been analyzed. Model validation can be done by comparing new built model size by using confusion matrix. (Luque et al., 2019). The confusion matrix is used to present the performance of a classifier information. (Bramer, 2007).

The confusion matrix is used in machine learning for the classification or determination of the behavior of the classification model (Hasnain et al., 2020). The confusion matrix is represented by rows and columns, where the row is the actual class and the column is the prediction class (Caelen, 2017).

Tabel 1
Confusion Matrix for Two Classes

		Prediction Class	
		Positive (P)	Positive (P)
Actual Class	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

A good model has the highest number of true values and the fewest false values. Accuracy, sensitivity, and specificity were calculated using the numbers in Table 1. The definition of classification performance in Table 2 can be seen in formula (1) for sensitivity (SNS), formula (2) for specificity (SPC), and formula (3) for accuracy (ACC).

$$SNS = \frac{TP}{TP+FN} \quad (1)$$

$$SNS = \frac{TN}{TN+FP} \quad (2)$$

$$ACC = \frac{TP+TN}{TP+FN+TN+FP} \quad (3)$$

The positive prediction class level is the positive actual class classified correctly. The negative prediction class level is the incorrectly classified positive actual class. The positive prediction class level is the negative actual class classified incorrectly. The negative prediction class level is the negative actual class classified incorrectly (Galar et al., 2012).

Data Description

This study uses secondary data obtained from CBS, namely HDI, life expectancy at birth, expected years of schooling, mean years of schooling, and adjusted real expenditure per capita in 2020 total of 514 districts/cities data in Indonesia. The dependent variable in this study is the HDI, which consists of low (low HDI and moderate HDI) which is categorized as 1 (one) and high (high HDI and very high HDI) is categorized with 2 (two). Life expectancy at birth (LE), expected years of schooling (EYS), mean years of schooling (MYS), and adjusted real expenditure per capita (Expenditures) as independent variables. The units of LE, EYS and MYS are in years while expenditures are in thousands of rupiah. Table 2 shows a description of the data.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/ 10.33395/sinkron.v6i1.11173
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Tabel 2
Data Description

Variable	Definition	Scale
Y	Human development index (HDI)	Nominal
X_1	Life expectancy at birth (LEB)	Ratio
X_2	Expected Years of Schooling (EYS)	Ratio
X_3	Mean Years of schooling (MYS)	Ratio
X_4	Adjusted real per capita expenditure (Expenditure)	Ratio

Stage of Analysis

Figure 1 is the stage of analysis carried out in this study, namely collecting data. Determine variables, namely dependent variables, and independent variables. Perform descriptive analysis. Next, divide the data, namely training data by 80% (411 districts/cities in Indonesia) and testing data by 20% (103 districts/cities in Indonesia). Build a bagging model from training data. Determine the outcome of the prediction. Calculate the accuracy, sensitivity, and specificity of the test data. Drawing conclusions.

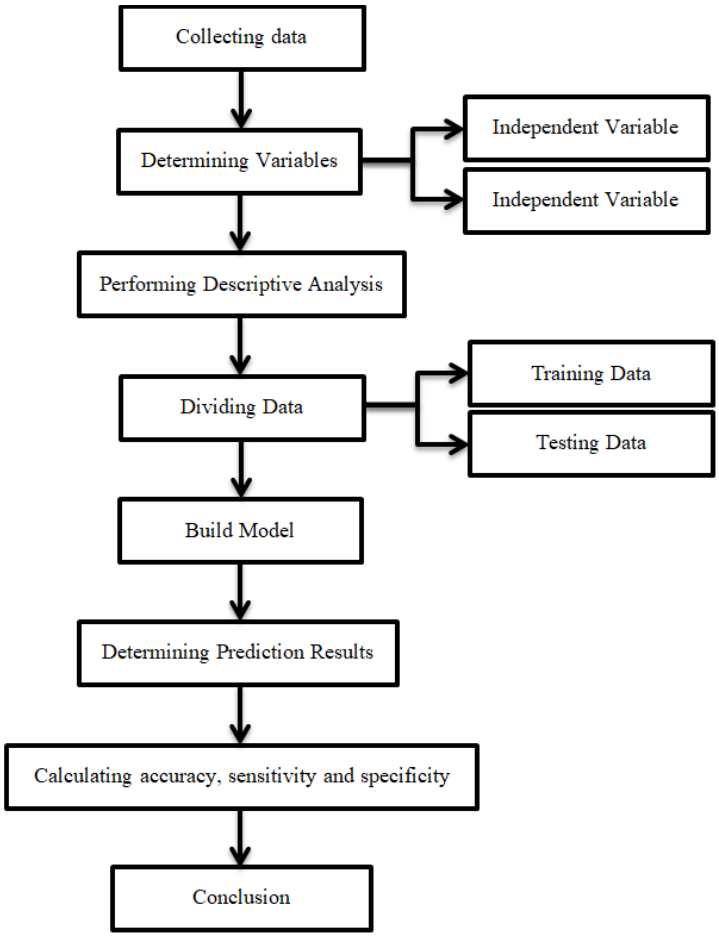

Figure 1. Stages of Analysis

RESULT

Descriptive Analysis

The results of the descriptive analysis in this study can be seen in Table 3.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/ 10.33395/sinkron.v6i1.11173
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3
Descriptive Analysis

Variable	N	Minimum	Maximum	Average
LEB	514	55.27	77.65	71.47
EYS	514	3.61	17.79	12.98
MYS	514	1.13	12.65	8.48
Expenditure	514	3,975	23,575	11.013
HDI	514	31.55	86.61	71.94

Table 3 showed that there are 514 districts/cities in Indonesia on the variables of LEB, EYS, MYS, Expenditure, and HDI. Life expectancy at birth was lowest at 55.27 years, the highest at 77.65 years, and an average of 71.47 years. The shortest expected years of schooling were 3.61 years, the longest was 17.79 years, and the average was 12.98 years. The shortest mean years of schooling were 1.13 years, the longest was 12.65 years, and the average was 8.48 years. The adjusted real expenditure per capita was at least Rp. 3,975,000,-, the maximum is Rp. 23,575,000,- and an average of Rp. 11,013,000,-.

In 2020, Indonesia's HDI nationally was 71.94, increasing from 2019's HDI of 71.92. At the provincial level, the highest HDI in 2020 was achieved by DKI Jakarta province at 80.77, while the lowest was Papua province at 60.44. (Badan Pusat Statistik, 2021b). At the District/City level, the average HDI in districts/cities across Indonesia was 71.94 with the highest HDI being 86.61 and the lowest being 31.55 (Badan Pusat Statistik, 2021a).

Prediction Results, Accuracy, Sensitivity, and Specificity

In this study, the bootstrap process was carried out 25 times (replication) to form a classification tree. Each replication was carried out on training data taken randomly with repetition. In the aggregating stage, predictions are made using test data on each classification tree that has been obtained by applying majority voting in classifying the data. The confusion matrix can be seen in Table 4.

Table 4
Confusion Matrix

		Prediction Result Class	
		Low	Tall
Actual Class	Low	55	0
	Tall	8	40

Table 4 showed that there was a misclassification in the high class. Eight observations should be in the high class but are classified in the low class. The values of sensitivity, specificity, and accuracy using the bagging method can be seen in Table 5.

Table 5
Value of Accuracy, Sensitivity, and Specificity

Sensitivity (%)	Specificity (%)	Accuracy (%)
100	83.33	92.23

DISCUSSIONS

The sensitivity value obtained is 100%. This value showed the sensitivity of the bagging model in detecting data in the positive class, namely the low class. Based on the sensitivity value, it is known that the proportion of data that is predicted to be labeled low from the actual data that is labeled low is 100%.

The specificity value indicates the reliability of the bagging model in detecting data with a negative labeled category (high HDI category) correctly. Based on the analysis obtained a specificity of 83.33%. This matrix showed that the model is quite good at detecting high-class data.

The accuracy value is the proportion of the number of data that is correctly predicted from all test data. The bagging model applied provides an accuracy of 92.23%, meaning that the model is very good at classifying data. Overall, a balanced accuracy of 91.67% was obtained so that it can be stated that the bagging model has a good performance in classifying HDI according to district/city in Indonesia.

*name of corresponding author

CONCLUSION

The results showed that the classification of HDI in Indonesia in 2020 based on district/city using the Bagging method obtained an accuracy value of 92.3%, sensitivity of 100%, and specificity of 83.33%. The bagging method is highly efficient for classifying the HDI with a balanced accuracy of 91.67%. The results of this classification can be used to predict HDI in the following year and determine development programs that must be implemented according to regional priorities based on the HDI category within the region. Suggestions for further research may be made by other methods that may produce a higher balanced accuracy.

ACKNOWLEDGMENT

This research was funded by DIPA FMIPA, Riau University with grant number 87/UN19.5.1.1.3/KPT/2021.

REFERENCES

Badan Pusat Statistik. (2021a). *[Metode Baru] Indeks Pembangunan Manusia 2019-2020*. Badan Pusat Statistik. https://www.bps.go.id/indicator/26/413/1/-metode-baru-indeks-pembangunan-manusia.html

Badan Pusat Statistik. (2021b). *[Metode Baru] Indeks Pembangunan Manusia menurut Provinsi 2018-2020*. Badan Pusat Statistik. https://www.bps.go.id/indicator/26/494/1/-metode-baru-indeks-pembangunan-manusia-menurut-provinsi.html

Badan Pusat Statistik. (2021c). *Apa Itu Indeks Pembangunan Manusia?* Badan Pusat Statistik. https://www.bps.go.id/subject/26/indeks-pembangunan-manusia.html

Bramer, M. (2007). Principles of Data Mining. In M. Bramer (Ed.), *Springer Science+Business Media springer.com* (Issue January 2007). Springer Science+Business Media springer.com. https://doi.org/10.1007/978-1-84628-766-4

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3–4), 429–450. https://doi.org/10.1007/s10472-017-9564-8

Darsyah, M. Y. (2014). Klasifikasi Indeks Pembangunan Manusia (IPM) dengan Pendekatan K-Nearset Neighbor (K-NN). *Seminar Nasional Pendidikan, Sains Dan Teknologi Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang*, 29–35. https://www.researchgate.net/publication/339470520_KLASIFIKASI_INDEKS_PEMBANGUNAN_MANUSIA_IPM_DENGAN_PENDEKATAN_K-NEARSET_NEIGHBOR_K-NN

Fauzi, F. (2017). K-Nearset Neighbor (K-NN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah. *Jurnal Mipa*, *40*(2), 118–124. https://journal.unnes.ac.id/nju/index.php/JM/article/view/12884/7338

Fauzi, Fatkhurokhman, Yamin, M., & Wahyu, T. (2017). Klasifikasi Indeks Pembangunan Manusia Kabupaten / Kota Se-Indonesia dengan Pendekatan Smooth Support Vector Machine (SSVM) Kernel Radial Basis Function (RBF). *Seminar Nasional Pendidikan, Sains Dan Teknologi Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang*, 88–97. https://jurnal.unimus.ac.id/index.php/psn12012010/article/view/2986

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. *IEEE Access*, *8*, 90847–90861. https://doi.org/10.1109/ACCESS.2020.2994222

Issam H. Laradji, Mohammad Alshayeb, L. G. (2015). Software defect prediction using ensemble learning on selected features. *Information and Software Technology ScienceDirect*, *58*(September 2019), 388–402. https://doi.org/10.1016/j.infsof.2014.07.005

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition Elsevier Ltd*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Mauludiyah, K. (2020). Klasifikasi Indeks Pembangunan Manusia Kabupaten/Kota di Indonesia Menggunakan Metode Random Forest. *Jurnal ILmiah*.

Mohammed, M., Khan, M. B., & Bashie, E. B. M. (2017). Machine learning: Algorithms and applications. In E. B. M. B. Mohssen Mohammed, Muhammad Badruddin Khan (Ed.), *Machine Learning: Algorithms and Applications*. CRC Press Taylor & Francis Group. https://doi.org/10.1201/9781315371658

Mordelet, F., & Vert, J. P. (2013). A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters Elsevier B.V.*, *37*(1), 201–209. https://doi.org/10.1016/j.patrec.2013.06.010

*name of corresponding author

Muttaqin, M. F. J., & Zulkarnain. (2020). Cluster Analysis Using K-Means Method to Classify Indonesia Regency/City based on Human Development Index Indicator. *ACM International Conference Proceeding Series*, 81–85. https://doi.org/10.1145/3400934.3400951

Naomi Altman, M. K. (2017). Points of Significance: Ensemble methods: Bagging and random forests. *Nature Methods*, *14*(10), 933–934. https://doi.org/10.1038/nmeth.4438

Otok, B. W., Musa, M., Purhadi, & Yasmirullah, S. D. P. (2020). Propensity score stratification using bootstrap aggregating classification trees analysis. *Heliyon*, *6*(7), 0–7. https://doi.org/10.1016/j.heliyon.2020.e04288

Verawaty, Muji Gunarto, Rolia Wahasusmiah, C. I. M. (2021). Determinants of Human Development Index in Indonesia. *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore*, *March*, 4199–4210. https://www.researchgate.net/publication/353750726_Determinants_of_Human_Development_Index_in_Indonesia

Wahono, R. S., & Suryana, N. (2013). Combining particle swarm optimization based feature selection and bagging technique for software defect prediction. *International Journal of Software Engineering and Its Applications SERSC*, *7*(5), 153–166. https://doi.org/10.14257/ijseia.2013.7.5.16

Windridge, D., & Nagarajan, R. (2017). Quantum Bootstrap Aggregation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10106 LNCS*(February), 115–121. https://doi.org/10.1007/978-3-319-52289-0_9

Xindong Wu, V. K. (2009). The Top Ten Algorithms in Data Mining. In V. K. Xindong Wu (Ed.), *Taylor & Francis Group, LLC* (Vol. 53, Issue 9). Taylor & Francis Group, LLC. https://doc.lagout.org/Others/Data Mining/The Top Ten Algorithms in Data Mining %5BWu %26 Kumar 2009-04-09%5D.pdf

*name of corresponding author