JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 6. NO. 2 FEBRUARY 2021
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i2.1935

IMPLEMENTATION OF ENSEMBLE TECHNIQUES FOR DIARRHEA CASES CLASSIFICATION OF UNDER-FIVE CHILDREN IN INDONESIA

Andriansyah Muqiit Wardoyo Saputra [1]; Arie Wahyu Wijayanto [2*]

Program Studi D4 Statistika[1]; Program Studi D4 Komputasi Statistik[2*]
Politeknik Statistika STIS
www.stis.ac.id
211709554@stis.ac.id [1]; ariewahyu@stis.ac.id [2*]
(*) Corresponding Author

Abstract—Diarrhea is an endemic disease in Indonesia with symptoms of three or more defecations with the consistency of liquid stool. According to WHO, diarrhea is the second largest contributor to the death of under-five children. Data and cases of children under five years who have diarrhea are very difficult to find, so the data analysis process becomes difficult due to the lack of information obtained. Difficulties in the data analysis process can be overcome by rebalancing, so the category ratios are balanced. The method that is popularly used is SMOTE. To solve imbalanced data and improve classification performance, this study implements the combination of SMOTE with several ensemble techniques in diarrhea cases of under-five children in Indonesia. Ensemble models that are used in this study are Random Forest, Adaptive Boosting, and XGBoost with Decision Tree as a baseline method. The results show that all SMOTE-based methods demonstrate a competitive performance whereas SMOTE-XGB gains a slightly higher accuracy (0.88), precision (0.96), and f1-score (0.86). The implementation of the SMOTE strategy improved the recall, precision, and f1-score metrics and give higher AUC of all methods (DT, RF, ADA, and XGB). This study is useful to solve the imbalanced problems in official statistics data provided by BPS Statistics Indonesia.

Keywords: Diarrhea, Ensemble techniques, Imbalanced class, XGBoost.

Abstrak—*Penyakit diare merupakan penyakit endemis Indonesia dengan gejala buang air besar sebanyak tiga kali atau lebih dengan konsistensi tinja cair. Menurut WHO, diare menjadi penyumbang terbesar kedua kematian anak di bawah lima tahun. Data dan kasus balita yang mengalami diare sangat sulit ditemukan sehingga proses analisis data menjadi sulit karena kurangnya informasi yang didapatkan. Kesulitan dalam proses analisis data dapat diatasi dengan melakukan rebalancing agar rasio kategori menjadi berimbang. Metode yang populer digunakan adalah SMOTE. Untuk menyelesaikan masalah data tidak berimbang dan meningkatkan kinerja klasifikasi, penelitian ini mengimplementasikan kombinasi SMOTE dengan beberapa teknik ansambel pada kasus diare balita di Indonesia. Model ansambel yang digunakan adalah Random Forest, Adaptive Boosting, dan XGBoost dengan Decision Tree sebagai metode baseline. Hasil penelitian menunjukkan bahwa semua metode berbasis SMOTE menunjukkan kinerja kompetitif dengan SMOTE-XGB mendapatkan akurasi yang sedikit lebih tinggi (0,88), presisi (0,96), dan f1-score (0,86). Implementasi metode SMOTE jelas meningkatkan ukuran metrik recall, presisi, dan f1-score dan memberikan AUC yang lebih tinggi dari semua metode (DT, RF, ADA, dan XGB). Penelitian ini berguna sebagai pembelajaran dalam mengatasi data tidak berimbang pada data statistik resmi yang disediakan oleh BPS Statistics Indonesia.*

Kata Kunci: *Diare, Ensemble techniques, Imbalanced class, XGBoost*

INTRODUCTION

Diarrheal disease is the second-largest cause of death in children under the age of five and contributes to the death of approximately 525,000 children each year. Diarrhea is defined as a condition in which an individual experiences defecation with a frequency of three times or more with the consistency of liquid stool, which can be accompanied by blood and or mucus [1].

Diarrhea is one of the endemic diseases in Indonesia. The disease is also classified into potential diseases of Extraordinary Events (KLB) which are often followed by death [2]. So, cases of diarrhea need to be of particular concern to be able to be prevented.

Data about toddlers with diarrhea is very difficult to obtain and cases are rare. The process of data analysis becomes difficult due to the lack of information obtained. When the statistical analysis

is done, information about a little data will cause the data to have an imbalanced class.

The application of a simple classification algorithm on data that has an imbalanced class will cause bias in larger classes/categories and treat fewer classes as noise [3]. Therefore, it is necessary to do special handling of the data. The most common way to handle imbalanced data is balancing the data using resampling techniques. This technique is also known as the data level approach to class imbalance learning. A popular method used in this resampling technique is the Synthetic Minority Oversampling Technique or SMOTE [4]. However, this method can raise some problems such as bias that negatively affect the model performance [5], overfitting, loss of important information from data, and so on [6]. In research conducted by [7], it is proven that applying oversampling to training set samples before learning can improve the recall rate of minority categories significantly. Thus, this approach can solve problems in the minority categories as well.

Another approach used to overcome the imbalanced class is to apply the ensemble method. In this method, an algorithmic approach is carried out by increasing the weight on samples undergoing misclassification to improve classification performance [8]. The ensemble methods proposed in this study are Random Forest [9, 10], AdaBoost [11], and XGBoost [12].

According to the previous studies [12, 13], applying a hybrid method (data level approach and ensemble) for imbalance learning was superior to other models. In [6], concluded that a hybrid method can increase AUC gradually with an increase in the percentage of oversampling. Related research was also conducted [15] by comparing SMOTE Random Forest and SMOTE XGBoost models on the HCV dataset. The results obtained that models with SMOTE can significantly increase the recall value from under 2% to more than 70%.

In this study, researchers aimed to overcome misclassification in cases of diarrhea of toddlers in Indonesia by combining SMOTE and ensemble-based classification methods. SMOTE is applied so that the data ratio becomes balanced and then the classification results are improved by several ensemble methods.

MATERIALS AND METHODS

This study is a further research from a scientific article entitled "Penerapan Metode Resampling dalam Mengatasi *Imbalanced* Data pada Determinan Kasus Diare pada Balita di Indonesia (Analisis Data SDKI 2017)" [16].

Data and Variables

The source of data used in this study is derived from Indonesia Demographic and Health Surveys (IDHS) 2017 by BPS Statistics Indonesia in collaboration with the National Population and Family Planning Board (BKKBN), Ministry of Health, and ICF [17].

The units of analysis in this study are households that had under-five children, formerly 11,340 households. Within 11,340 households there were 1,560 households had under-five children with diarrhea cases in two weeks recently during the survey. The research used eight predictor variables that can be seen in Table 1 below.

Table 1. Research variables [16]

Name	Symbol	Type	Category
Diarrhea case (Target)	Y	Categoric	0, No 1, Yes, 2 weeks recently
Predictor:			
Children's age	X1	Numeric	-
Mother's age	X2	Numeric	-
Children's sex	X3	Categoric	0, Female 1, Male
Residence's type	X4	Categoric	0, Rural 1, Urban
Educational attainment	X5	Categoric	0, No education 1, Complete elementary 2, Complete junior high school 3, Complete senior high school
Main floor material	X6	Categoric	0, Natural floor 1, Material floor 2, Finished floor
Type of toilet facility	X7	Categoric	0, No toilet 1, Good toilet 2, Shared toilet 3, Bad toilet
Source of drinking water	X8	Categoric	0, Proper 1, Not proper

Software Specification

The analysis and modeling process in this study was done using the Python programming language with Google Colaboratory backend. The specification of the Google Colaboratory backend shown in Table 2 below.

Table 2. Software specification

Disk	RAM	CPU
107,77 GB (31.94 used)	13 GB (0.75 used)	Intel Xeon Processor 2 cores @ 2.30 GHz

The machine learning tool that is used in this study for classification algorithms is scikit-learn API in python [18].

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

VOL. 6. NO. 2 FEBRUARY 2021
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i2.1935

Research Framework

This section will explain the framework of the research to get the results. First, the imbalanced dataset is carried out to the data preprocessing step. In this step, we prepare the data before performing the analysis process. Then, create a balanced dataset that has gone through by SMOTE. This dataset is used to compare to the original data. The two datasets perform in the training process using several ensemble methods. Last, this study applies the evaluation and validation process to obtain how SMOTE and ensemble methods affect the classification result. The process is shown in Figure 1 as follows.

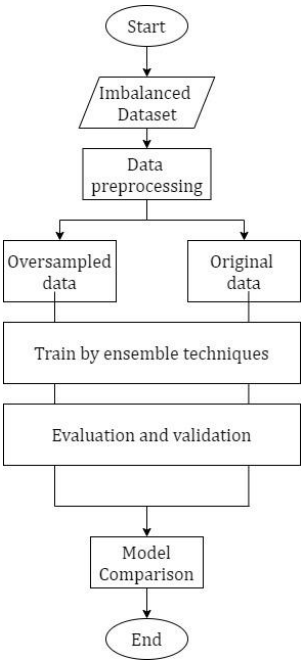

Figure 1. Research framework

Methods

This section explains the brief method that is used in this study. Based on the research framework, the explanation of the analysis steps are as follows:

1. Data preprocessing

The initial stage before an in-depth analysis is the initial processing step or preprocessing. This step aims to prepare the data ready to be processed and provide results with maximum performance. The process carried out is as follows:

a. Standardization

Numeric variables are standardized using the Z-score method with the formula:

$$Z = \frac{X-\mu}{\sigma} \quad \text{.....(1)}$$

b. Creating dummy variables

Categorical variables that have more than two categories are converted into dummy variables with the first category as the reference.

2. Rebalancing with SMOTE

The SMOTE method is applied in this step to overcome imbalanced class in diarrhea cases data in toddlers. The resampling process is done only to the minority class of category 1 (Yes, two weeks recently). The result of this process is new data with a balanced ratio of 50:50.

3. Data modeling

Binary classification is applied in this research as a learning method. Because the learning process focuses to solve the problem in minority classes (cases of toddler's diarrhea), then the algorithm used is ensemble-based methods. The methods used are Random Forest [9, 10], AdaBoost [11], and XGBoost [12]. All parameter settings in Random Forest and AdaBoost are set to default. While in XGBoost are set n_estimator to 100 and max_depth to 8. Also, the Decision Tree is applied in training as a baseline algorithm with default parameter setting [19].

Model training with Decision Tree, Random Forest, AdaBoost, and XGBoost algorithms are also applied to the dataset without SMOTE. This is an assessment to prove that the hybrid method (data level approach and ensemble) can overcome imbalanced class well.

4. Data evaluation and validation

The final step in this research is the evaluation and validation of the proposed model at the modeling step. The data is divided into two sets, namely the training set and the test set. Learning will be done using a training set while evaluation and validation using the test set.

The evaluation of the model in this study used a confusion matrix [20] with details in Table 3 below.

Table 3. Confusion matrix

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

Based on the confusion matrix can be derived into several metrics as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(2)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(3)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{.....(4)}$$

$$F1-score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad \text{.....(5)}$$

In addition to the confusion matrix and some of its derivative metrics, the study also used ROC to look at the comparison of models based on

the Area Under Curve (AUC). This curve is proven to be more sensitive to compare different learning models [20, 21]. Then, validation is also carried out by looking at the learning curve of cross-validation results as mitigation in the case of model overfitting.

RESULTS AND DISCUSSION

Oversampling Results

Table 4 is shown regarding the number of samples used in the study. Diarrhea cases data in toddlers have an imbalanced ratio of 6.27:1. This imbalance is necessary to be rebalanced by SMOTE.

Table 4. Information of data sample

# Total	# Majority	# Minority	IR
11,340	9,780	1,560	6.27:1

The comparison of the number of cases before and after resampling is shown through the bar chart in Figure 2. It appears that the result of oversampling has obtained new data with a balanced ratio (50:50).

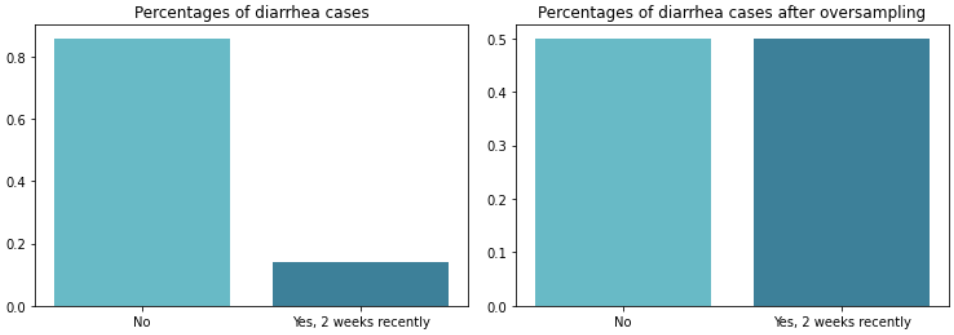

Figure 2. Percentages of diarrhea cases in each category before and after oversampling

Table 5. Model's performance comparison

Model	Accuracy	Recall	Precision	F1-score
DT	0.80	0.13	0.20	0.15
RF	0.83	0.07	0.20	0.33
ADA	0.81	0.11	0.20	0.14
XGB	0.86	0.01	0.19	0.02
SMOTE-DT	0.85	**0.83**	0.86	0.84
SMOTE-RF	0.82	0.78	0.86	0.82
SMOTE-ADA	0.85	0.82	0.87	0.85
SMOTE-XGB	**0.88**	0.79	**0.96**	**0.86**

Model Comparison

After modeling with the proposed classification method, the evaluation of each model is carried out. The evaluation is done by making predictions on the test data based on several measures that have been described in the research method.

Table 5 showing the comparison of models' performance using eight classifiers with four metrics evaluation derived from the confusion matrix. All SMOTE-based methods demonstrate a competitive performance whereas SMOTE-XGB gains a slightly higher accuracy (0.88), precision (0.96), and f1-score (0.86). Overall, the implementation of the SMOTE strategy improved the recall, precision, and F1-score metrics of all methods (DT, RF, ADA, and XGB).

Based on the ROC curve shown in Figure 3, all SMOTE-based methods give higher AUC than other methods (DT, RF, ADA, and XGB). This result is related to the previous study [6] that explained applying SMOTE to imbalanced data can improve the AUC value.

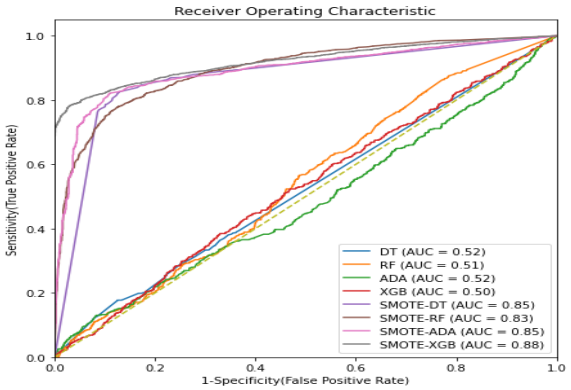

Figure 3. ROC Curve

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

VOL. 6. NO. 2 FEBRUARY 2021
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i2.1935

Model Validation

To avoid overfitting issues, the researcher applies a learning curve into the SMOTE-based model by compare training scores and cross-validation scores. The cross-validation option is set to 100 splits and 30% samples for the test set. Based on Figure 4, the learning curve will converge into a specific score as the training size increases. SMOTE-XGB converges faster than another model. It happens because XGBoost has scalability that runs faster than other models [12].

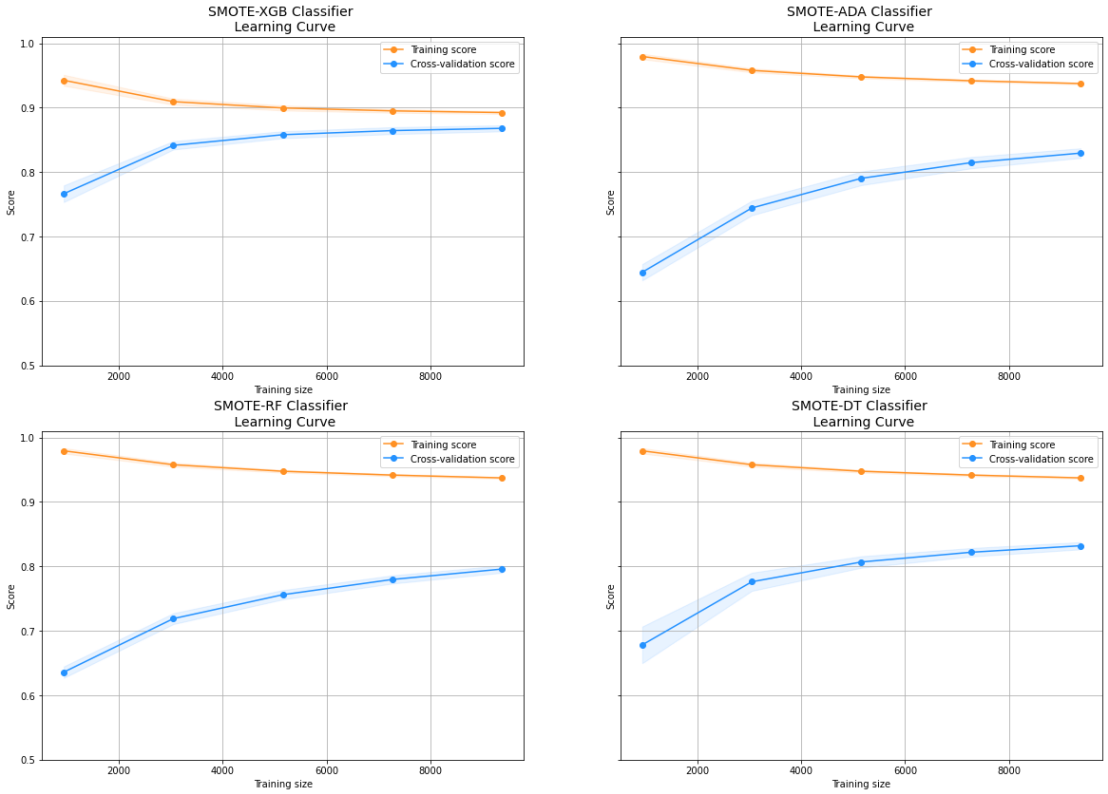

Figure 4. Learning curves

CONCLUSION

From the results and discussions, it can be concluded that the application of the hybrid method can handle an imbalanced class. This explains by the implementation of the SMOTE strategy improved the recall, precision, and F1-score metrics of all methods (DT, RF, ADA, and XGB) with the SMOTE-XGB gains slightly higher accuracy. SMOTE-based methods also get higher AUC values than others. Belonging to the learning curve, SMOTE-XGB runs competitively to other methods that converge fast. This study is useful for doing assessments and solve the imbalanced data in official statistics data provided by BPS Statistics Indonesia. For further studies, this study suggests applying a more complex algorithm to solve imbalanced data. Furthermore, statistical testing (parametric or nonparametric) with a specific level of significance is needed to choose whether the best model.

REFERENCE

[1] World Health Organization, "Diarrhoeal disease," 2017. https://www.who.int/en/news-room/fact-sheets/detail/diarrhoeal-disease (accessed Dec. 16, 2020).

[2] Kementerian Kesehatan RI, "Profil Kesehatan Indonesia Tahun 2017," 2017. [Online]. Available: https://www.kemkes.go.id/.

[3] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, 2012, DOI: 10.1109/TSMCC.2011.2161285.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif.*

Intell. Res., vol. 16, pp. 321–357, 2002.

[5] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019, DOI: 10.1016/j.ins.2019.07.070.

[6] U. R. Salunkhe and S. N. Mali, "Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 725–732, 2016, DOI: 10.1016/j.procs.2016.05.259.

[7] H. Dong, D. He, and F. Wang, "SMOTE-XGBoost using Tree Parzen Estimator optimization for copper flotation method classification," *Powder Technol.*, vol. 375, pp. 174–181, 2020, DOI: 10.1016/j.powtec.2020.07.065.

[8] K. Li, G. Zhou, J. Zhai, F. Li, and M. Shao, "Improved PSO_AdaBoost ensemble algorithm for imbalanced data," *Sensors (Switzerland)*, vol. 19, no. 6, 2019, DOI: 10.3390/s19061476.

[9] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, DOI: 10.1023/A:1010933404324.

[10] R. Polikar, "Ensemble Learning," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA: Springer US, 2012, pp. 1–34.

[11] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997, DOI: https://doi.org/10.1006/jcss.1997.1504.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, DOI: 10.1145/2939672.2939785.

[13] Z. Chen, J. Duan, L. Kang, and G. Qiu, "A Hybrid Data-Level Ensemble to Enable Learning from Highly Imbalanced Dataset," *Inf. Sci. (Ny).*, 2020, DOI: 10.1016/j.ins.2020.12.023.

[14] S. Wang *et al.*, "A new method of diesel fuel brands identification: SMOTE oversampling combined with XGBoost ensemble learning," *Fuel*, vol. 282, no. May, p. 118848, 2020, DOI: 10.1016/j.fuel.2020.118848.

[15] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan Metode SMOTE Random Forest dan SMOTE XGBoost untuk Klasifikasi Tingkat Penyakit Hepatitis C pada Imbalance Class Data," *J. Gaussian*, vol. 9, no. 3, pp. 227–236, 2020.

[16] A. M. W. Saputra, I. P. Ananda, M. A. Rizki, Z. D. Hapsari, and R. Nooraeni, "Penerapan Metode Resampling dalam Mengatasi Imbalanced Data pada Determinan Kasus Diare pada Balita di Indonesia (Analisis Data SDKI 2017)," *J. Mat. dan Stat. serta Apl.*, vol. 8, pp. 19–27, 2020, DOI: 10.24252/msa.v8i1.13452.

[17] National Population and Family Planning Board - BKKBN, Statistical Indonesia - BPS, Ministry of Health - Kemenkes, and ICF, "Indonesia Demographic and Health Survey 2017 [Dataset]." 2018. Distributed by ICF. Available: http://dhsprogram.com/pubs/pdf/FR342/FR342.pdf.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[19] C. Gao and H. Elzarka, "The use of decision tree based predictive models for improving the culvert inspection process," *Adv. Eng. Informatics*, vol. 47, no. October 2020, p. 101203, 2020, DOI: 10.1016/j.aei.2020.101203.

[20] J. Han, M. Kamber, and J. Pei, "8 - Classification: Basic Concepts," in *Data Mining (Third Edition)*, Third Edit., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 327–391.

[21] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases.," *Radiology*, vol. 148, no. 3, pp. 839–843, Sep. 1983, DOI: 10.1148/radiology.148.3.6878708.

[22] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.," *Casp. J. Intern. Med.*, vol. 4, no. 2, pp. 627–635, 2013.