Data balancing process to strengthen a malaria control prediction system in Senegal

Kodzo Parkoo[1], Bamba Gueye[2], Cheikh Sarr[1], Ibrahima Dia[3]

[1] Université Iba Der Thiam,
[2]Université Cheikh Anta Diop,
[3]Institut Pasteur de Dakar

Abstract. Malaria is a public health problem in Senegal. Despite the implementation of prevention and treatment programs, the prevalence rate remains high, although there has been a noticeable decrease over the years.
To further strengthen our efforts in the fight against malaria, we have previously developed a prediction system aimed at assessing the presence or absence of Anopheles larvae in specific sites. This system, is a crucial component of our anti-larval control (ALC) strategy, which involves gathering physico-chemical parameters from the site and using them to predict the likelihood of larvae presence. Given that, our prediction system relies on these physico-chemical parameters, ensuring the reliability and quality of the data is paramount. In our previous study, although we had access to reliable and high-quality data, we encountered an issue with data imbalance. To validate the accuracy of our prediction system, it is essential to address this data imbalance.

Keywords: SMOTE algorithm, balanced data, data prediction, malaria.

1 Introduction

In 2021, the World Health Organization (WHO) reported an estimated 247 million cases of malaria worldwide, resulting in 619,000 deaths [1]. Africa accounted for about 95% of all malaria cases and 96% of the associated deaths [2]. Particularly in Senegal, there were 536,850 reported cases and 399 deaths in 2021. These numbers represented an increase compared to the year 2020 with the 445,313 confirmed cases and 373 deaths reported in 2020 [3]. Thus, despite the diligent efforts made through the national malaria control program, malaria is still present in Senegal.

One unique particularity of malaria control in Senegal is its focus on combating the adult stage of the Anopheles. An essential strategy involves area-specific interventions targeting the larval development phase, known as Larval Control (LAL). To enhance LAL efforts, previous work has aimed to develop a predictive model to determine the presence or absence of larvae. The prediction process relies on physico-chemical parameters from Anopheles breeding sites [4].

However, it should be noted, that the quality of predictions from the previous model is questionable due to the confusion matrix of the latter. Indeed, the results from the

confusion matrix of the prediction model yield four results: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), with respectively TP corresponding to the presence of larvae and TN to the absence of larvae. An analysis made on the results of the confusion matrix shows that the value of TN is equal to zero (0). This value revealed an imbalance in the data source used to set up the prediction model [4]. There are various potential solutions to address the data source imbalance. One option is to create a new database, which would involve gathering and processing new data. However, this approach needs substantial costs. In order to maintain the reliability of the results of the prediction model, it is therefore necessary to balance the existing data and re-test the model. The efficiency of this prediction model will be a benefit in the LAL.

The remaining sections of this document are organized as follows. Section 2 deals with the motivations for choosing the SMOTE algorithm. Afterwards, Section 3 discusses the different tools, methods used, and implementations made in this research. Section 4 presents the outcomes and findings from the implementation and comprehensive discussion of the results. Finally, Section 5 depicts the conclusion presents some perspectives and challenges.

2 Various SMOTE algorithms

Faced with the need to balance data in order to improve prediction, the use of SMOTE seems appropriate. SMOTE (Synthetic Minority Oversampling Technique) appears to be a well-suited approach for oversampling minority observations [5]. As data imbalance is a frequent problem in classification and affects the performance of machine learning models, a suitable solution should be proposed.

In the field of medicine, data analysis methods have proved extremely useful in healthcare for early diagnosis to provide better medical treatment and thus minimize the mortality rate in cases such as breast cancer, diabetes, coronary heart disease, kidney disorders, and more. However, a survey of existing models reveals shortcomings in data processing analysis and also in learning classification algorithms. This is due to the imbalance in the data, leading to unbalanced results. To address these challenges and ensure the reliability of predictions, the SMOTE algorithm, more precisely Distance-based SMOTE (D-SMOTE) and Bi-phasic SMOTE (BP-SMOTE) coupled with learning algorithms, have enabled Sowjanya A. M. et al. to propose hybrid sampling techniques that enhance prediction accuracy in unbalanced health data [6].

The existence of several variants of SMOTE allows for its adaptation to various types of data or problems. These variants include: classical SMOTE, proposed by Chawla et al. [5] as an alternative to random cloning of minority data, which can lead to overlearning.

SMOTE-NC (Nominal Continuous), which is an extension of SMOTE for mixed data, i.e., data containing both numerical and categorical variables. It was introduced by Chawla et al [7] as an improvement on classical SMOTE, which cannot handle categorical variables without encoding them in numerical form, giving rise to errors.

Therefore, Borderline-SMOTE [8], focuses on minority observations located at the border between classes since they are more difficult to classify.

Furthermore, ADASYN (Adaptive Synthetic Sampling) .[9], adapts the number of synthetic data to be generated according to the degree of nesting between classes whereas SVMSMOTE considers an SVM (Support Vector Machine) to figure out the most relevant minority observations to oversample.

On the other hand Safe-Level-SMOTE [10] defines a safety level for each minority observation, based on the number of neighbors in the same class, and favors observations with a high safety level. In contrast, Cluster-SMOTE, groups minority observations into clusters, and generates synthetic data within each cluster. In addition to adapting to different types of data, these different variants also pave the way for new SMOTE-based methods.

The ASN-SMOTE, as proposed by Yi et al. [11], represents a novel approach to oversampling unbalanced data. It is based on the classical SMOTE but improving the neighbor's selection and synthetic data generation. Its aim is to reduce noise and improve data quality for the minority class. The main idea is to use the majority class to perceive the decision frontier, and adoptively select qualified neighbors for each minority observation. This reduces noise and improves data quality for the minority class [11].

Regarding its advantages, the SMOTE algorithm provides several key benefits. It improves the performance of classification models despite the presence of imbalances. In addition, it prevents overtraining and preserves information. SMOTE generates synthetic samples by combining characteristics of neighbors and reduces classification bias that can occur when the model is strongly biased towards the majority class. These advantages bring SMOTE a highly suitable algorithm for addressing imbalances in preparation for predictive modeling.

3 Materials and methods

3.1 Experimental settings

The implementation of the SMOTE algorithm was carried out using Python. To do this we deployed a virtual machine, Windows 7, with 4 GB of RAM, 60 GB of storage and 4 Intel processors at 2.10 GHz. On this machine, we installed Anaconda 3, with different applications: Jupyter Lab and Jupyter Notebook.

The main steps of the data processing were: import of libraries, import of the initial database (unbalanced), removal of non-determinant columns, exploratory analysis for descriptive statistics of the data, implementation of the SMOTE algorithm, and verification of the balancing of the data.

3.2 Data examination for balancing

The data under consideration stems from measurements conducted in October 2020 within the Toubacouta district and its surroundings due to the sympatric presence of An. *arabiensis*, An. *gambiae* and An. *coluzzii* species and the observation of contrasting

hybridization rates between the latter two. In order to collect these physicochemical parameters, an inspection was made at each site, the larvae were collected using either the «dipping» or «pipetting» method, depending on the size of the sites, and were subsequently placed in labeled jars denoting the site number.

For each site, a comprehensive record was made of the presence or absence of Culicidae other than Anopheles, as well as observations regarding vegetation, turbidity, and sunlight. Following this, precise measurements of nests dimensions (length, width, depth) were taken with a decameter, alongside the assessment of additional factors such as their status, turbidity, sunlight exposure, and the presence of vegetation or other mosquito larvae.

Subsequently, the following parameters were then measured with a portable field tester (SD Card Real time Datalogger): temperature, amount of dissolved oxygen, salt content and pH. The larvae were then sorted in the laboratory and stored in tubes containing 70° ethanol. The dataset encompasses a representation of the physicochemical characteristics and the presence or absence of larvae.

Table 1. Presentation of the data collected

PH	Temp	Conductivity	Saltiness	Dissolved oxygen	Turbidity	Sunshine	Vegetation	Status
7,45	32,5	43,9	0	11	0	0	0	1
7,68	34,1	227	0	8,7	1	0	1	0
7,26	30,2	52,6	0	20,2	1	0	0	1

In Table 1, only the physico-chemical characteristics and the presence of larvae are of interest. Furthermore, our final data source will retain only the latter [4]. The main reason for the imbalance in the data comes from the fact that out of 4700 record lines are composed mostly by the presence of larvae

3.3 Imbalance concept and balancing algorithm

Unbalanced data is a common situation when dealing with real data. We can evoke the imbalance when we have observations distributed in two (02) classes and the frequencies of these two (02) classes are not in a ratio of 50% each. However, in real data, the notion of imbalance is evoked if the ratio is between 10% and 90%, i.e. if the imbalance exceeds 10% for the minority class .[12].

In case of data imbalance, the need to rebalance the data goes through two (02) methods: subsampling and oversampling.

Subsampling consists of removing part of the majority class in order to give more importance to minority individuals. Oversampling consists in increasing the number of minority individuals so that they have more importance in the modeling. These two (02) balancing methods are implemented thanks to the SMOTE algorithm on which we have previously exposed. In our particular case, an oversampling will be performed on the minority class characterized by the absence of larvae at the status level. This allows to densify the population of minority individuals in a more homogeneous way [5].

3.4 SMOTE Implementation

The data that require balancing were collected in October 2020 by a team from the Pasteur Institute of Dakar in Toubacouta, Senegal. These initial data include information on the deposit, the physico-chemical characteristics of the deposit and the presence or not of larvae. For our prediction, we focused on the physico-chemical characteristics and the presence or absence of larvae. It is worth noticing that fixed chemical parameters and a couple of physical parameters will be kept as well as the presence or absence of larvae.

Table 2. Retained physico-chemical parameters.

pH	Temperature	Conductivity	Salinity	Dissolved. Oxygen	Turbidity	Status
7.45	32.5	43.9	0	11	0	1
7.68	34.1	227	0	8.7	1	0
7.26	30.2	52.6	0	20.2	1	1
6.87	32.7	25.5	0	5.2	1	1
7.70	33.6	68.9	0	23.6	1	1

The Table 2 informs about the physico-chemical parameters retained in our prediction.

Table 3. Database description

	pH	Temperature	Conductivity	Salinity	Dissolved. Oxygen	Turbidity	Status
count	4794.0	4794.0	4794.0	4794.0	4794.0	4794.0	4794.0
mean	7.387	31.3489	208.6042	0.0080	21.4931	0.531915	0.9148
std	0.9886	2.732198	271.673	0.0192	8.421062	0.49932	0.2790
min	3.630	25.00	15.80	0.0	5.20	0.0	0.0
25%	6.980	29.80	58.10	0.0	15.7	0.0	1.0
50%	7.360	32.10	112.70	0.0	21.10	1.0	1.0
75%	7.790	33.60	255.0	0.010	26.80	1.0	1.0
max	9.90	34.30	1520.0	0.10	35.70	1.0	1.0

In Table 3, we can notice that our database consists of 4794 occurrences, with an average value of approximately 0.914894 for the «Status» column. Notably, this column represents the presence and absence, denoted by 0 and 1, respectively. The average value of 0.9148 for this column is indicative of the extent of the data imbalance.

Table 4. Status options count

Status value	Count
1	4386
0	408

Table 4 illustrates valuable insights into the extent of the data imbalance. Specifically, we observe that there are 4386 occurrences indicating the presence of larvae compared to only 408 for the absence of larvae. To correct this imbalance, we resort to oversampling, thanks to the SMOTE algorithm. The objective of the execution of the SMOTE algorithm is to achieve parity between the occurrences of the minority class and the majority class.

3.5 Verifying data balance

The execution of the oversampling with the SMOTE algorithm on our data provided us with a result. In order to know whether the balancing was successful, we performed a descriptive analysis of our data source.

Table 5. Database description after SMOTE

	pH	Temperature	Conductivity	Salinity	Dissolved. Oxygen	Turbidity	Status
count	8772	8772.0	8772.0	8772.00	8772.0	8772.0	8772.0
mean	7.442376	31.484302	199.656920	0.006651	20.517004	0.516986	0.500
std	0.749027	2.661693	219.109651	0.015416	9.630774	0.499740	0.500029
min	3.63	25.000	15.800	0.000	5.200	0.000	0.000
25%	7.27	29.800	58.100	0.000	10.500	0.000	0.000
50%	7.36	32.100	120.500	0.000	21.100	1.000	0.500
75%	7.79	34.000	255.000	0.010	26.800	1.000	1.000
max	9.90	34.300	1520.000	0.100	35.700	1.000	1.000

As shown in Table 5, the average related to the distribution of the status (regrouping the presence or absence of larvae, represented by 0 and 1 is 0.5.

Our main objective is to confirm the reliability of the prediction system implemented in [4]. Indeed, it is mandatory to recalibrate the confusion matrix using the balanced data source.

3.6 Confusion matrix

The confusion matrix, often referred to as a contingency table, serves as a vital tool for assessing the performance of a classification model. In its basic form, it compares the actual data for a target variable with the predictions made by the model [13].

After splitting our balanced database by SMOTE into two for training and prediction, we implement the logistic regression algorithm to the data.

Table 6. Confusion matrix result

Result	Count
True Positive (TP)	1181
True Negative (TN)	1100
False Positive (FP)	1094
False Negative (FN)	1011

The confusion matrix result in Table 6 indicates the reliability level of our possible prediction tests.

3.7 ROC curve

The ROC (Receiver Operating Characteristic) curve is a graphical tool used in the context of classification problems. It allows us to evaluate the performance of different classification models. Its use also includes the AUC (Area Under the Curve), which helps to compare different models [14].

In our specific case, we have access to the results of the confusion matrices for our data before and after applying the SMOTE algorithm for balancing. Table 6 shows the results obtained after applying SMOTE, while Table 7 presents the results before SMOTE was applied.

Table 7. Confusion matrix result without SMOTE [4]

Result	Count
True Positive (TP)	2193
True Negative (TN)	0
False Positive (FP)	204
False Negative (FN)	0

To construct the ROC curve, we need to determine the True Positive Rate (TPR) and the False Positive Rate (FPR). The respective mathematical representations are as

follows: TPR = TP/ (TP + FN) and FPR = FP/ (TN + FP) [14]. After evaluating our model, we obtain the classification thresholds as detailed in Table 8.

Table 8. Threshold's classification

Class. Thresh.	0.1	0.2	0.2	0.3	0.4	0.5	0.6	0.7	0.8
t_v_posi	0.01	0.11	0.13	0.02	0.03	0.04	0.05	0.06	0.8
t_f_posi	0.01	0.03	0.08	0.015	0.25	0.42	0.65	0.08	0.95
t_v_posi_smote	0.05	0.18	0.42	0.63	0.76	0.87	0.92	0.96	0.99
t_f_posi_smote	0.01	0.03	0.08	0.15	0.25	0.42	0.65	0.8	0.95

4 Results and discussion

4.1 Balanced data

The results presented in table 5 indicate an average value between the presence and the non-presence at 0.5. This leads to an equality of distribution on the presence and non-presence of larvae. This enables to correct the imbalance that was posed on all our collected data, Table 9.

Table 9. Status options count after balancing

Status value	Count
1	4386
0	4386

4.2 Reliability of the prediction system

The reliability of the prediction system implemented in [4] is based on the data balance on one hand and on the confusion matrix on the other. The interpretation of this matrix is based on True Positive (TP) and Positive False (PF). TP presents the accuracy and PF the recall on the accuracy of the predictions. Thus, with a high accuracy and a high recall we can assume that our data are well managed by the model.

4.3 Representation of the ROC Curve

The result of the verification of our classification thresholds grouped in Table 8 gives the curve illustrated in Figure 1.

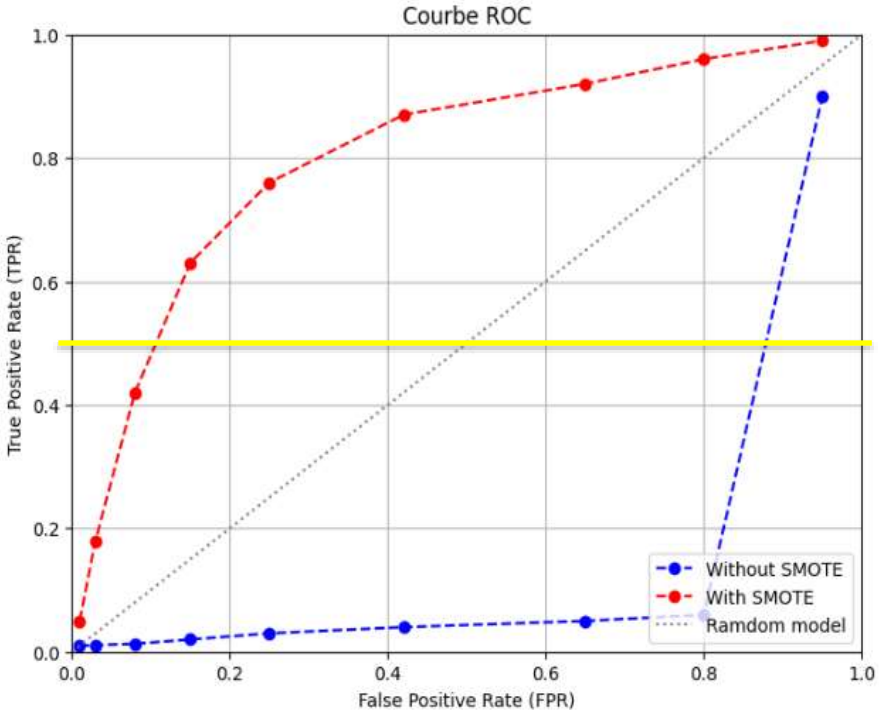

Fig. 1. ROC curve

By observation of Figure 1, we distinguish the curves representing data from our initial database and those for the database balanced using SMOTE. The point (0,0) represents the threshold where everything is classified as negative, i.e., no false positives (FPR = 0) and no true positives (TPR = 0). The point (1,1) represents the threshold where everything is classified as positive, i.e., no true negative (FPR = 1) and no false negative (TPR = 1). We can see that the ROC curve in red, representing data balanced by SMOTE, evolves gradually for TPR values tending towards 1. In our particular case our TPR is 0.95, expressing better logistic regression performance on the balanced database.

4.4 Discussions

The implementation of a reliable prediction system requires measures to ensure its sustainability and accuracy. The system set up in [4] faced reliability issues, particularly in its tendency to predict situations of larval presence. The root of the problem was the imbalance in the training data used for the model, it was therefore imperative to have a balanced data source in order to retest the prediction system. The first option to have a new source of data, was to make a new collection of on the field by taking this time, the care to collect on zones of which was sure of the absence of the larvae.

This processing enables to have a database with a near balance on the presence or non-presence of larvae. This first option could not be implemented, because of various constraints associated with scientific fieldwork. A second option was therefore proposed, to balance the existing data. To achieve this, we employed the SMOTE algorithm, utilizing oversampling techniques. This approach promotes building a new database, which in turn facilitates the retesting of our prediction system.

Based on the results obtained from the ROC curve, we can confidently conclude that the use of the SMOTE algorithm to balance our data enhanced the prediction system's quality and reliability. At this stage, we can affirm the system's dependability for predictive purposes.

5 Conclusion

The establishment of a reliable and balanced database for collecting physico-chemical parameters from Anopheles breeding sites and determining the presence or not of Anopheles larvae, proved to be a complex task.

The database we have created, used for prediction, is the result of collecting physico-chemical parameters directly on the deposits. These data have a major particularity justifying their use: all parameters are correlated by the presence or absence of larvae. This point being crucial in the prediction, this database was suitable. However, based on its exploitation, the confusion matrix resulting from the execution of the logistic regression shows a significant imbalance in the database. This led to inaccuracy in the process of learning and predicting data, although the prediction model is good, the predictions are incorrect.

Faced with this challenge, two solutions were possible: the resumption of operations to collect physico-chemical parameters with this time the need to take equitably from roosts with the presence of larvae and breeding sites without the presence of larvae on the one hand and on the other hand to balance the data from our existing database. The first option involves huge mission costs coupled with the time required for processing in order to have a suitable database, we opted for the second option which is to balance the database.

Therefore, we used the SMOTE algorithm to balance our data. Thanks to the oversampling method applied to the data, we were able to obtain after execution of the logistic regression a correct confusion matrix to re-evaluate our data prediction system.

This reassessment was made possible thanks to the ROC curve. The results from the ROC curve indicate that the SMOTE algorithm significantly improved the accuracy of our prediction system. Furthermore, it confirmed that the learning and prediction system, based on the crucial parameters and logistic regression as previously reported in [4], is functioning correctly.

We plan conducting comprehensive field tests to further refine the system and effectively relaunch larval control efforts in order to reduce the malaria prevalence rate.

References

[1] "Despite continued impact of COVID-19, malaria cases and deaths remained stable in 2021." Accessed: Mar. 28, 2023. [Online]. Available: https://www.who.int/news/item/08-12-2022-despite-continued-impact-of-covid-19--malaria-cases-and-deaths-remained-stable-in-2021

[2] "Fact sheet about malaria." Accessed: Mar. 28, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/malaria

[3] "Le paludisme au Sénégal en 2021 (document)," Infomed. Accessed: Apr. 09, 2023. [Online]. Available: https://infomed.sn/le-paludisme-au-senegal-en-2021-document/

[4] K. M. Parkoo, B. Gueye, C. Sarr, and I. Dia, "Data prediction system in malaria control based on physio-chemical parameters of anopheles breeding sites," *EAI Endorsed Trans. Internet Things*, vol. 8, no. 4, pp. e3–e3, Dec. 2022, doi: 10.4108/eetiot.v8i4.2936.

[5] 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, Jun. 2002, doi: 10.1613/jair.953.

[6] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," *Appl. Nanosci.*, vol. 13, no. 3, pp. 1829–1840, Mar. 2023, doi: 10.1007/s13204-021-02063-4.

[7] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," in *Knowledge Discovery in Databases: PKDD 2003*, N. Lavrač, D. Gamberger, L. Todorovski, and H. Blockeel, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2003, pp. 107–119. doi: 10.1007/978-3-540-39804-2_12.

[8] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing*, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, pp. 878–887. doi: 10.1007/11538059_91.

[9] H. He, Y. Bai, E. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," presented at the Proceedings of the International Joint Conference on Neural Networks, Jul. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[10] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," in *Advances in Knowledge Discovery and Data Mining*, T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 475–482. doi: 10.1007/978-3-642-01307-2_43.

[11] X. Yi, Y. Xu, Q. Hu, S. Krishnamoorthy, W. Li, and Z. Tang, "ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection," *Complex Intell. Syst.*, vol. 8, no. 3, pp. 2247–2272, Jun. 2022, doi: 10.1007/s40747-021-00638-w.

[12] C. Tremblay, "Imbalanced data et Machine Learning," Kobia. Accessed: Apr. 10, 2023. [Online]. Available: https://kobia.fr/imbalanced-data-et-machine-learning/

[13] Alexandre, “Managing Unbalanced Classification Problems - Part 1,” Data Science Courses | DataScientest. Accessed: May 10, 2023. [Online]. Available: https://datascientest.com/en/management-of-unbalanced-classification-problems-i

[14] “ROC Curve — Machine Learning — DATA SCIENCE.” Accessed: Jul. 20, 2023. [Online]. Available: https://datascience.eu/machine-learning/understanding-auc-roc-curve/