water

MDPI

Article

Extreme Rainfall Event Classification Using Machine Learning for Kikuletwa River Floods

Lawrence Mdegela [1,2,*], Esteban Municio [3], Yorick De Bock [1], Edith Luhanga [4], Judith Leo [2] and Erik Mannens [1]

[1] Department of Computer Science, University of Antwerp-imec IDLab, Sint-Pietersvliet 7, 2000 Antwerp, Belgium
[2] The Nelson Mandela African Institution of Science and Technology, Arusha P.O. Box 447, Tanzania
[3] i2CAT Foundation, 08034 Barcelona, Spain
[4] Carnegie Mellon University Africa, Kigali P.O. Box 6150, Rwanda
* Correspondence: lawrencenehemiah.mdegela@uantwerpen.be; Tel.:+32-494594736

Abstract: Advancements in machine learning techniques, availability of more data sets, and increased computing power have enabled a significant growth in a number of research areas. Predicting, detecting, and classifying complex events in earth systems which by nature are difficult to model is one such area. In this work, we investigate the application of different machine learning techniques for detecting and classifying extreme rainfall events in a sub-catchment within the Pangani River Basin, found in Northern Tanzania. Identification and classification of extreme rainfall event is a preliminary crucial task towards success in predicting rainfall-induced river floods. To identify a rain condition in the selected sub-catchment, we use data from five weather stations that have been labeled for the whole sub-catchment. In order to assess which machine learning technique is better suited for rainfall classification, we apply five different algorithms in a historical dataset for the period of 1979 to 2014. We evaluate the performance of the models in terms of precision and recall, reporting random forest and XGBoost as having the best overall performances. However, because the class distribution is imbalanced, a generic multi-layer perceptron performs best when identifying heavy rainfall events, which are eventually the main cause of rainfall-induced river floods in the Pangani River Basin.

Keywords: heavy rainfall; river floods; machine learning

Citation: Mdegela, L.; Municio, E.; De Bock, Y.; Luhanga, E.; Leo, J.; Mannens, E. Extreme Rainfall Event Classification Using Machine Learning for Kikuletwa River Floods. *Water* **2023**, *15*, 1021. https://doi.org/10.3390/w15061021

Academic Editors: Fazlul Karim, Zaved Khan and Tawatchai Tingsanchali

Received: 16 January 2023
Revised: 25 February 2023
Accepted: 28 February 2023
Published: 8 March 2023

1. Introduction

Rainfall-induced river floods are among Earth's most common and most catastrophic natural hazards [1]. Worldwide, flash floods account for more than 5000 deaths annually, with a mortality rate more than 4 times greater than other types of flooding [2], and subsequently, their social, economic, and environmental impacts are significant. According to the Tanzania Meteorological Agency, in the last decade, the northern part of the country has experienced its heaviest rainfall accompanied by strong winds, causing the most severe floods of the past 50 years [3]. It is without a doubt that with the changing climate, such events are likely to become more frequent, not only in Tanzania, as evidenced in several reports (Burundi and Tanzania—Floods Leave Homes Destroyed, Hundreds Displaced. https://floodlist.com/africa/burundi-tanzania-floods-late-february-2021 (accessed on 16 December 2022), Tanzania—Severe Flooding in Mtwara Region After Torrential Rainfall. https://floodlist.com/africa/tanzania-flood-mtwara-january-2021 (accessed on 16 December 2022), Tanzania—12 Killed in Dar Es Salaam Flash Floods. https://floodlist.com/africa/tanzania-daressalaam-floods-october-2020 (accessed on 16 December 2022)), but across the globe. The effects of floods are notably severe in developing or low-income countries such as Tanzania because of their vulnerability to the occurrence of these phenomena. The vulnerability is partly due to limited human capacity and limited

resources invested in managing the problem [4]. Understanding the trends and key patterns in the occurrence of rainfall events, is an important step towards better flood risk management plans that will help in designing more accurate early warning systems [5].

Machine learning (ML) presents the ability to identify the hidden patterns and trends in historical climate data [6] and may be used to classify and predict key rainfall events that are associated with the occurrence of floods. The potential of machine learning techniques to improve the classification and prediction of extreme rainfall events has been demonstrated by several studies in the past. The techniques provide valuable insights into the spatial and temporal patterns of extreme rainfall events and their impacts on flood generation, water resources management, and climate change impact assessment. In [7], the authors proposed an event-based flood classification method to study the global river flood generation processes. The approach is based on a combination of unsupervised and supervised machine learning methods that can provide event-based information for better understanding of flood generation processes. Another machine learning-based downscaling approach is demonstrated in Pham et al. [8], where a combination of random forest and least square support vector regression was found to improve the accuracy of extreme rainfall predictions at a local scale. The inspiration of the method used here is that it provides valuable insights into the extreme rainfall events and their spatial and temporal characteristics, which are useful for water resource management and flood risk assessment.

Similarly, Ref. [9] developed a machine learning-based classification method to categorize extreme precipitation events over Northern Italy. The study employed a k-means clustering technique to identify distinct clusters of extreme rainfall events and used decision trees to develop a classification scheme. Despite the fact that most of these studies were conducted in developed counties, where there is advancement in both technology and human resources, the results show significant potential for use as models for similar studies in other developing regions, such as Tanzania.

Furthermore, Ref. [10] presented a study that used three different machine learning algorithms (XGBoost, LightGBM, and CatBoost) to forecast daily stream flow in a mountainous catchment. The study compared the performance of the three algorithms and showed that machine learning can provide accurate stream flow forecasts, which are valuable for water management and flood prediction. An analysis of physical causes of extreme precipitation [11] can also be used to identify key climatic variables that drive extreme precipitation events, and machine learning based approaches can be applied to predict the occurrence of extreme precipitation.

We apply five machine learning techniques, namely random forest, extreme gradient boost, support vector machine, k-nearest neighbors, and multilayer perceptron to identify and classify rainfall events in the Karanga–Weruweru–Kikafu sub-catchment, located within the Pangani River Basin, Tanzania. Random forest is preferred for its robustness to large and noisy data sets [12] and its ability to handle imbalanced data sets [13,14]. XGBoost is computationally efficient [15], and it can also perform better on imbalanced data sets [16]. SVM is suitable for high-dimensional input space and modeling complex, non-linear relationships between inputs and outputs [17,18]. k-nearest neighbors, although considered to be one of the simplest machine learning algorithms, has been successful in a number of applications, from recognition of handwritten texts [19] to satellite image scenes [20] and mostly success in classification problems with irregular decision boundaries. MLP is a feed-forward neural network that has also shown success in classification problems, including extreme natural events such as droughts [21].

We compare these techniques and discuss the suitability of each in successfully classifying rainfall events. To train these models, a historical labeled data set from the Pangani Water Board (PWB) and the Tanzania Meteorological Agency (TMA) collected from five stations located across the Karanga–Weruweru–Kikafu sub-catchment was used. The nature of the data set gives us an imbalanced multi-class classification problem. There are three categories in the target class (heavy, light and no rain). Of these, heavy rainfall is the smallest, making up just 0.32% of the whole data set. The distribution is highly skewed

towards the majority class, in this case, light rain, which makes up 83.22% of the whole data set, leaving 16.46% to the no rain class. This simply means for every single example of a heavy rainfall event, there are 51 examples of no rain and 260 examples of light rain.

2. Materials and Methods

The study area under consideration is situated in the northern part of Tanzania in the south of the Kilimanjaro region. The Karanga–Weruweru–Kikafu (KWK) sub-catchment (Figure 1) and the villages along the Kikuletwa river are intensely affected by flash river floods from heavy rainfall. The aim of this work was to classify rainfall intensity among three classes (heavy, light, none). The categories are based on the rate of precipitation per period, with precipitation of more than 64.5 mm in a day classified as heavy, anything below that up to precipitation greater than zero (0) millimeters per day classified as light rain, and 0.00 classified as no rain. Data records for the study covering the period from 1979 to 2014 was provided by Tanzania Meteorological Agency (TMA), and the Pangani Basin Water Board(PBWB).

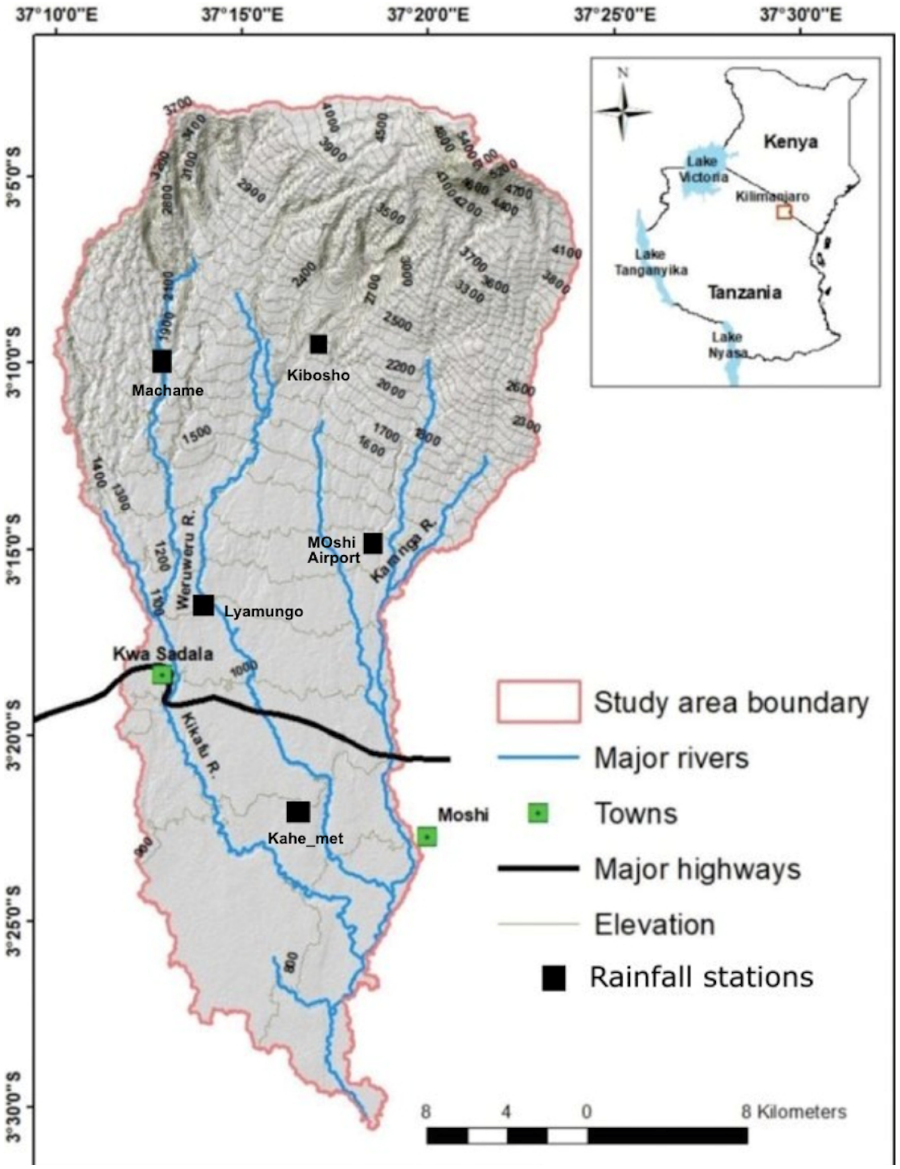

Figure 1. Karanga–Weruweru–Kikafu(KWK) sub-catchment.

The data contain daily weather data from the five stations (Machame, Kibosho, Lyamungo, Moshi Airport, and Kahe) for 35 years, with seven parameters, namely maximum

and minimum temperatures in Celsius (°C), precipitation in millimeters (mm), wind speed in meters per second (m/s), relative humidity expressed as percentage, solar irradiance in mega-joules per square metre (MJ/m^2), and rain category. The parameters are measured uniformly across all the locations and we applied data for the same period for all the stations.

It is worth noting that the data were consolidated from two main sources: the ground gauges and the satellite estimates. The basis of the consolidation and the decision to compliment the ground gauge data with the satellite estimates were based on several previous studies that were performed to test the validity of the satellite estimates over the region. Three studies, Refs. [22–24], investigated the spatio-temporal characteristics and accuracy of satellite-derived rainfall estimates in Tanzania in comparison to ground-based measurements. The studies revealed a positive correlation between the two data sources, indicating the potential of satellite-based rainfall estimates to be a useful complement to ground-based measurements, especially in areas with complex topography and limited ground-based measurement stations. Nevertheless, the satellite estimates exhibited a tendency to overestimate the ground-based measurements, and their accuracy varied in different locations. The findings suggest that the use of satellite-based rainfall estimates can enhance rainfall monitoring and prediction in regions where traditional measurement methods are sparse or challenging to implement, albeit with the need for continual improvements in their accuracy and uncertainty estimation.

2.1. Data Preparation

The data from each station were checked for missing values before being merged into a single data set. Simple line plots were used to verify whether all stations had similar patterns in the features and to identify any outliers. Simple statistical analysis was performed on the numerical features and are summarized in Table 1.

Table 1. Descriptive statistics of weather variables used in training.

Variable	Count	Mean	Std. Dev.	Min	25th Percentile	75th Percentile
Max temperature (°C)	10,389	22.71	3.08	13.26	20.53	24.93
Min temperature (°C)	10,389	12.90	1.96	5.71	11.61	14.41
Precipitation (mm)	10,389	3.17	6.34	0.00	0.15	3.66
Wind (m/s)	10,389	2.49	0.55	0.65	2.14	2.87
Relative humidity (%)	10,389	0.76	0.10	0.32	0.69	0.84
Solar (MJ/m^2)	10,389	16.92	7.23	0.00	11.18	22.19

Data were split into training and testing set in a ratio of 80% to 20%. In order to even out the distribution and to ensure that the distribution of the target variable is maintained in both the training and testing data sets, as there is an imbalance in the target class distribution, we applied stratified sampling. In total there were six features and one target class from each location. The features are maximum temperature, minimum temperature, precipitation, wind, relative humidity, and solar irradiance, and the target is rainfall category. Further feature engineering was performed, where all the object type columns were encoded to numeric type. The encoding was mainly for the target class which was the only object type. The target class had three classes, heavy rain, labeled "H", encoded as "0" (class 0); light rain, labeled "L", encoded as "1" (class 1); and no rain labeled "N", encoded as "2" (class 2). Pivoting was also performed to put the data in a format that is convenient for model training. Training data were normalized using *MinMaxScaler* from sklearn library.

2.2. Model Building

Two main things were considered during this stage before starting the models: first, the target class distribution and second, the multi-class classification. Our target class was

somehow severely imbalanced, with the distribution being highly skewed towards the majority class, light rain (83.22%), followed by no rain (16.46%), and heavy rainfall, which is our class of interest, is the minority (0.32%). This simply means for every single example of heavy rainfall, we had 51 examples of no rain and 260 examples of light rain.

Consideration on the part of multi-class classification was to use a multi-class strategy from scikit-learn (https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html, accessed on 25 November 2022) library known as the one-vs-the-rest (OvR) classifier. OvR is a heuristic technique of dealing with multi-class problems by fitting one classifier per class. For each classifier, the class is fitted against all the other classes. One of the implementations of OvR is from the sklearn library, which provides a separate *OneVsRestClassifier* class that allows the one-vs-rest strategy to be used with any classifier. A classifier that is inherently for binary classification is just provided to the *OneVsRestClassifier* as an argument.

Each model was then trained, tested, and evaluated. Because our problem falls under multi-class imbalanced classification, selecting a metric for evaluation was the most important step in the project. An incorrect metric would mean choosing the wrong algorithm and consequently solving a different problem from the one that you intend to solve.

2.3. Model Evaluation

Because we are dealing with a highly skewed data set, we chose precision and recall as our performance evaluation metrics. Precision (Equation (1)) is a ratio of the number of true positives divided by the sum of the true positives and false negatives. In other words, it provides information on how good a model is at predicting the positive class.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{1}$$

Recall (Equation (2)) on the other hand is the ratio of the number of true positives divided by the sum of the true positives and the false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False negatives}} \tag{2}$$

One important aspect of precision and recall to take note is that the calculations do not consider the use of the *truenegatives*. The focus is on the correct prediction of the minority class. A precision–recall curve is a plot of the precision on the *y*-axis and the recall on the *x*-axis for different thresholds. They give a more informative picture of an algorithm in skewed data sets, as has also been evident in a number of studies [25,26]. In that sense, we identified our positive class to be H for heavy rainfall and other collectively as negative classes (no rain and light rain). Nevertheless, precision and recall are in a trade-off relationship: at some point you may need to optimize one at the expense of the other [27]. Contextually, at some point you would want classifier that is good at minimizing both the false positives and false negatives, meaning that it would make more sense to have a model that is equally good at identifying cases were a false alarm of a heavy rainfall event occurs and when an alarm is not on when there is an event coming. In the view of that, we applied another metric called F1_score. F1_score is the harmonic mean (Equation (3)) of precision and recall and ranges from 0 to 1.

$$\text{F1\_score} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision + recall}} \tag{3}$$

3. Results

In our experiments, five different machine learning algorithms were used to identify and classify extreme rainfall events between three rainfall categories. As stated in the Introduction, the ability to identify extreme rainfall events is crucial for predicting rainfall-

induced river floods. The results from our evaluation show that overall random forest and XGBoost performed better than the rest, as we can see from the F1_scores summarized in Table 2.

Table 2. Summary of F1_score measures for the models.

Random Forest	XGBoost	Support Vector Machine	KNN	Multi-Layer Perceptron
0.998	0.998	0.878	0.898	0.950

Ideally, the scores from F1_score mean that both XGBoost and random forest have perfect precision and recall when you give equal importance to both false negatives and false positives. However, that is not the case in our problem. In classifying the heavy rains, which in our case is the minority class, false negatives were the most important. Intuitively, in our context, it is not helpful if we are successful in predicting all data points as negative, that is, not a heavy rainfall event. Instead, we focused on identifying the positive cases: the occurrence of a heavy rainfall event. Referring back to the definitions of the metrics, this simply means that we maximized the *recall*, the ability of our model to find all the relevant cases within a given data set. This notion is supported by a number of past studies [28–31]. In the view of that, F1_score is not the determinant for the appropriate model to use in this scenario; as was previously mentioned, it is the harmonic mean, and thus it takes into account both the precision and recall. Our main goal is to favor the minimization of the false negatives and not to cast equal importance to both the false negatives and false positives. We focused on having a model with high recall which is able to identify most of the heavy rainfall events (true positives), that way saving lives and properties from the consequences that accompany such events.

On the other hand, of course that is at the expense of issuing false alarms of heavy rainfall events as though they would happen (false positive) when they will not. Potentially, the associated costs of false positives will be unnecessary anxiety for people and at the worst, costs associated with taking unnecessary precautions. In most cases, the false positives will not be fatal. Therefore, because false negatives will results in fatalities and destruction, we want to have our classification threshold to favor the optimization of recall over precision.

This is the point where we turn our attention to the precision–recall curves for more insight. In a precision–recall curve, the goal is to maximize the area under the curve (AUC), which represents the overall performance of the classifier. A higher AUC indicates better performance in terms of balancing precision and recall and therefore a better ability to identify instances of the minority class.

In addition to the overall AUC, the shape and position of the curve can also provide insights into the performance of the classifier. A curve that is close to the top-right corner of the plot indicates a classifier with high precision and high recall, whereas a curve that is close to the bottom-left corner indicates a classifier with low precision and low recall. The shape of the curve can also reveal whether the classifier is biased towards precision or recall, which can inform the selection of a decision threshold that balances the two [32]. In Figures 2–6, we show the precision–recall curves for each of the considered classification algorithms.

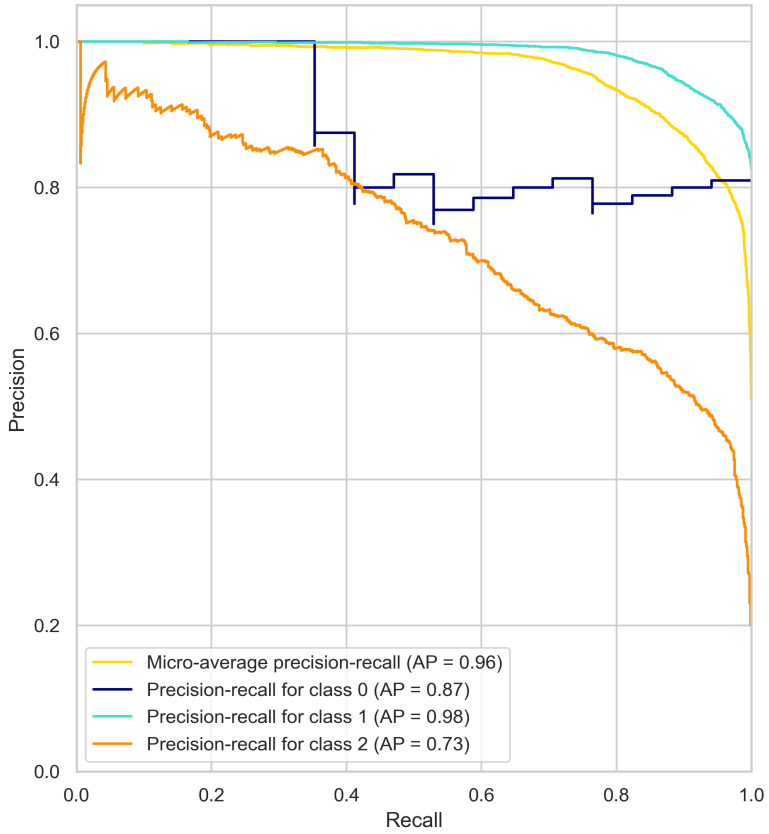

Figure 2. Precision–recall curve for SVM.

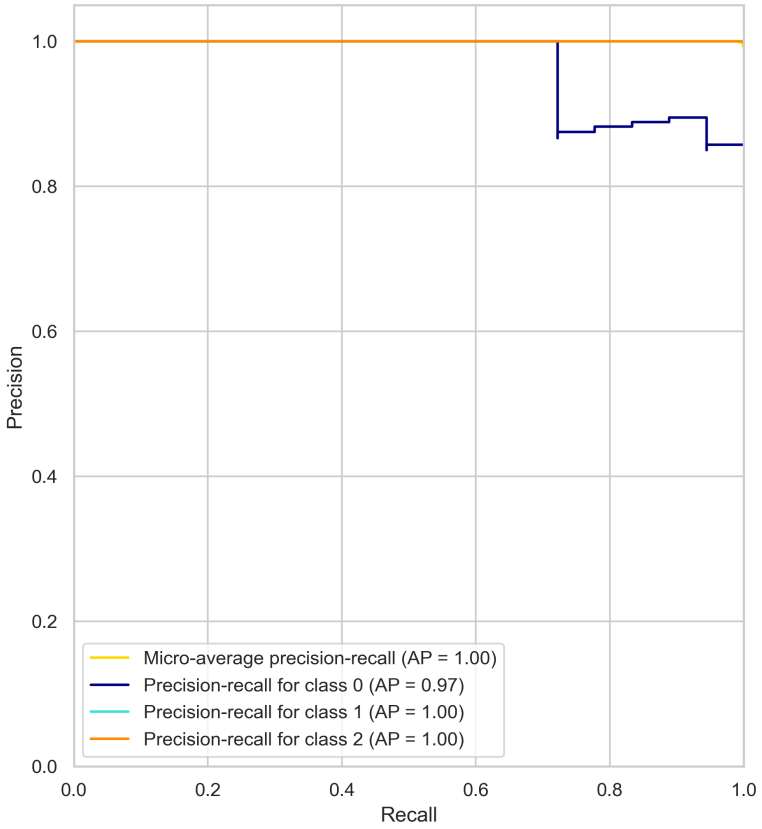

Figure 3. Precision–recall curve for random forest.

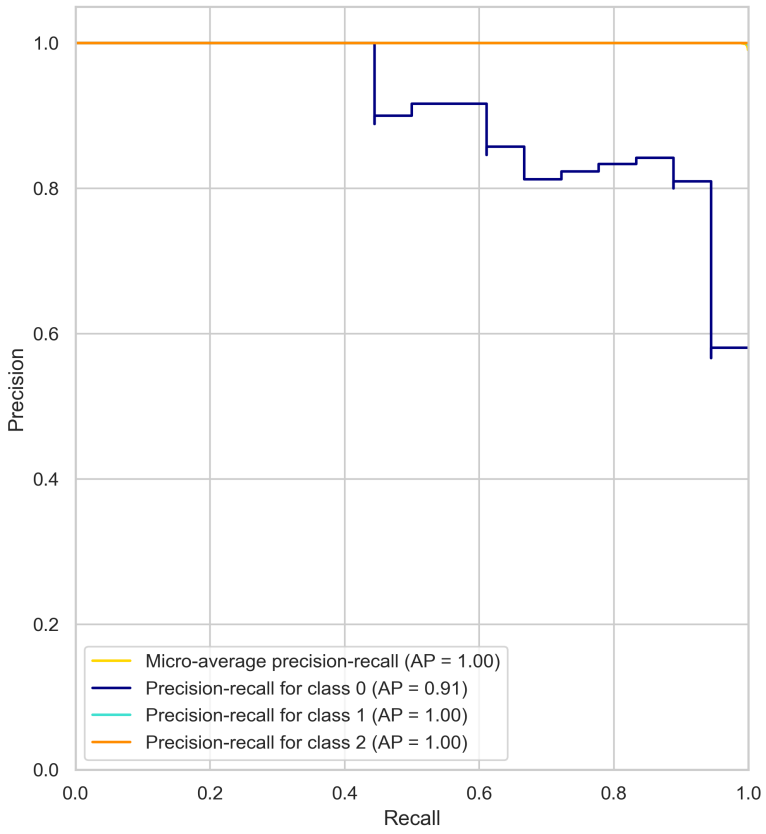

Figure 4. Precision–recall curve for XGBoost.

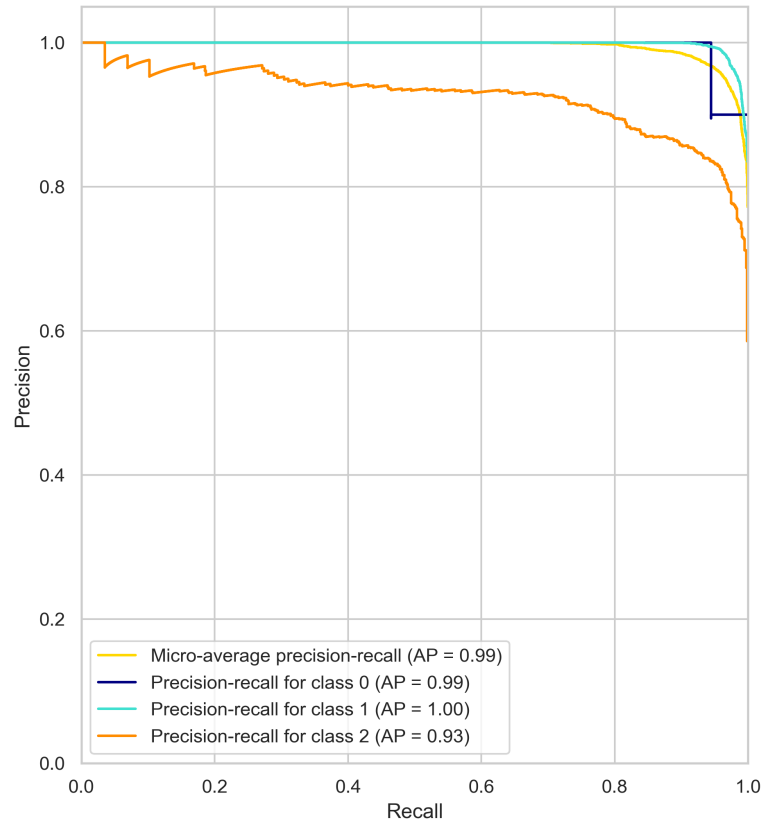

Figure 5. Precision–recall curve for multi-layer perceptron.

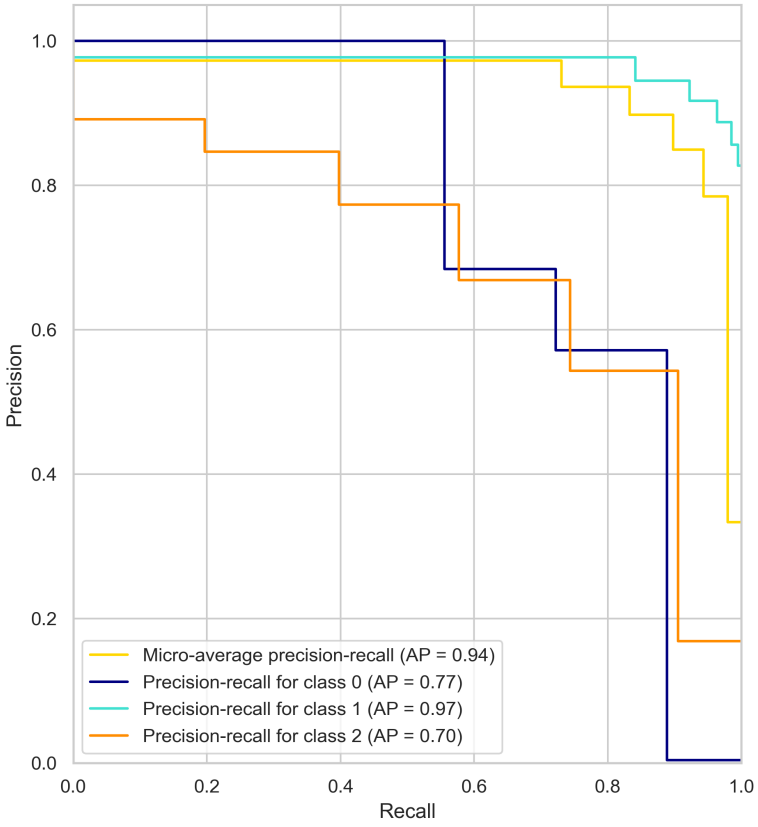

Figure 6. Precision–recall curve for KNN.

First, Figure 2 shows a precision–recall curve for a support vector machine in which the PR curve for the heavy rain class (class 0) is slightly below the micro-average PR curve, indicating a lower precision value for this class. The PR curve for the light rain class (class 1) is higher than both the micro-average PR curve and the PR curve for the heavy rain class, indicating high precision and recall values for this class. The PR curve for the no rain class (class 2) is below the micro-average PR curve and the PR curve for the light rain class, indicating relatively low precision and recall values for this class. Overall, the micro-average PR curve is high, with an AP value of 0.96, indicating a high level of precision and recall across all three classes.

Figure 3 illustrates the PR curve for a random forest classifier, with a precision–recall curve for the heavy rain class slightly below the micro-average precision–recall curve, indicating slightly lower precision values for this class, but with relatively high recall values. The precision–recall curve for the no rain class and light rain are all aligned at the top-left corner with the micro-average precision–recall curve, indicating a perfect level of precision and recall values for these classes. The micro-average precision–recall curve, with an AP value of 1, indicates a perfect level of precision and recall across all three classes.

Figure 4 depicts the precision–recall curves for XGBoost, suggesting that the classifier is highly accurate and consistent in identifying both light rain and no rain classes, with high precision and recall values for both. The precision–recall curve for the heavy rain class is slightly below the micro-average precision–recall curve, indicating slightly lower precision values for this class but relatively high recall values. Overall, XGBoost shows high performance in identifying rainfall classes, particularly for the light rain and no rain classes.

Figure 5 shows the precision–recall curve for the multilayer-perceptron classifier, with light rain (class 1) closest to the top-right corner of the plot, followed by the precision–recall curve for heavy rain (class 0), and then the curve for no rain (class 2), which is the lowest and is to the left of the other two curves, indicating that MLP had a lower level of precision and recall in identifying the no rain class as compared to the other two classes. However,

overall the classification model has achieved high precision and recall rates for all three classes, as indicated by the high micro-average AP of 0.99.

Finally, Figure 6 illustrates the PR curve for KNN classification with an overall micro-average AP of 0.94, indicating a relatively good performance across all three classes. The curve for light rain (class 1) has the highest AP of 0.97, indicating the best performance for this class. The curve for heavy rain (class 0) has an AP of 0.77, which is lower than class 1 and suggests that the classifier is less accurate in identifying instances of heavy rain compared to light rain. The curve for no rain (class 2) has an AP of 0.70, which is the lowest of the three classes, indicating that the classifier has the most difficulty identifying instances of no rain.

Despite the fact that random forest and XGBoost were the best performers overall when we put equal weight on both the false negatives and the false positives, it was the generic multilayer perceptron that performed the best when we focused on the minority positive class. The multilayer perceptron (MLP) was 98% accurate in identifying our class of interest, as can be seen in the precision–recall curves of Figure 5.

4. Discussion

To obtain deeper insight into the results highlighted above, we see that we based the evaluation on two sets of metrics, the precision–recall (PR) curve and the F-scores, to assess the models' ability to classify the minority class in a data set. The evaluation of algorithms in classifying the minority class in imbalanced data sets is a topic of ongoing research in the field of machine learning. For example, in a study by Batista et al. [33], the authors found that precision–recall curves were more reliable for evaluating imbalanced data sets than other metrics such as ROC curves.

4.1. Precision–Recall Curve Results Analysis

The precision–recall curve is a valuable metric for evaluating algorithms in imbalanced data sets, particularly when the positive class is rare. The PR curve provides a graphical representation of the trade-off between precision and recall, where precision measures the proportion of correct positive predictions among all positive predictions, and recall measures the proportion of correct positive predictions among all actual positive samples.

The micro-average PR curves of each model (see again Figures 2–6) summarize the overall performance of the model in all classes. The micro-average PR curve is computed by treating all the classes as a single binary classification problem. The micro-average PR curve A.P (average precision) score for the support vector machine (SVM), random forest, XGBoost, multi-layer perceptron (MLP), and k-nearest neighbors (KNN) models were 0.96, 1, 1, 0.99, and 0.94, respectively. On the other hand, looking at the individual class PR curves, the models achieved high precision–recall performances for classes 1 and 2, indicating a high ability to classify the absence of rain (class 2) and light rain (class 1). However, all models showed lower performance in predicting the occurrence of heavy rain (class 0), which is the minority class in the imbalanced data set.

Among the models, SVM (Figure 2) achieved the lowest A.P score of 0.96, and its PR curve for class 0 had the lowest A.P score of 0.87, indicating that the SVM model has the lowest ability to predict the minority class. On the other hand, the random forest (Figure 3), XGBoost (Figure 4), and MLP (Figure 5) models showed high performance in predicting the minority class, with A.P scores of 1, 1, and 0.99, respectively. The PR curves for class 0 for these models also achieved high A.P scores of 0.97, 0.91, and 0.98, respectively. KNN (Figure 6) achieved a moderate A.P score of 0.94, and its PR curve for class 0 had an A.P score of 0.77.

4.2. F1_score Results Analysis

The F-score is a single number that summarizes the harmonic mean of precision and recall. It is another useful metric for evaluating model performance in imbalanced data

sets. The F-scores of the models (Table 2) were as follows: random forest (0.998), XGBoost (0.998), SVM (0.878), KNN (0.898), and MLP (0.95).

Comparing the results of the two sets of metrics, the random forest and XGBoost models achieved the highest F-scores of 0.998, indicating their superior overall performance in predicting the occurrence of heavy rain. These models also showed high PR curve performance, particularly for class 0. The MLP model achieved the second-highest F-score of 0.95, indicating high performance in predicting the minority class. The SVM model had the lowest F-score of 0.878, consistent with its lower PR curve performance, particularly for class 0.

Overall, the random forest and XGBoost models showed the highest performance in predicting heavy rain events, whereas the SVM model had the lowest performance. The MLP model also demonstrated a good performance in predicting the minority class. It is important to note that the imbalanced nature of the data set presented a challenge to all models in predicting the minority class, and thus, the models' performance in this aspect should be carefully considered.

The PR curve analysis revealed that the random forest, XGBoost, and MLP models had high precision and recall performance, particularly for the minority class, whereas the SVM and KNN models had lower precision and recall performance, especially for the minority class. These results suggest that the random forest, XGBoost, and MLP models are more suitable for the prediction of heavy rain events in an imbalanced data set.

On the other hand, the F-score results revealed that the random forest and XGBoost models had the highest F-score, indicating their superior overall performance in predicting heavy rain events. The MLP model also showed a high F-score, but the SVM and KNN models had a lower F-score. These results suggest that the random forest, XGBoost, and MLP models are more suitable for predicting heavy rain events in an imbalanced data set based on the F1_score.

It is important to consider that the models' performance might be affected by the choice of evaluation metrics, and it is recommended to use multiple evaluation metrics to assess model performance. In this study, the PR curve and F-score were used to provide a comprehensive evaluation of the models' performance. Regarding the boundary and external conditions of our approach, we used a data set of meteorological features collected from a local weather station. These features include temperature, wind speed, humidity, pressure, and others, which are known to affect rainfall. We ensured the quality of the data set by removing missing values and outliers. Furthermore, we randomly split the data set into training and testing sets to evaluate the models' generalization performance.

We acknowledge that the external conditions, such as the terrain, geographical location, and topography, may affect the rainfall patterns, and our study did not specifically consider these factors. However, we believe that our approach provides a general framework that can be adapted to different settings by using relevant meteorological data.

On the other hand, it is important to note that these results were obtained using default hyperparameters, and there may be additional improvements that can be made by fine-tuning the models. However, these results still provide valuable insights into the relative performance of different algorithms in predicting rainfall classes in imbalanced data sets. Future studies could also investigate the impact of hyperparameter tuning on the models' performance in identifying extreme rainfall events.

Furthermore, the results consolidate the justification of applying ML models as compared to other models such as physical and numerical ones. This is consistent with studies such as that conducted by Chen et al. [34], where the authors found that the accuracy of machine learning models (MLP, RBF, SVM) was significantly better than that of the numerical model in both training and verification stages, as measured by root mean square error and R2. However, they noted that the numerical model's generalization ability is superior to the machine learning models' due to its inclusion of physical mechanisms. Physical-based models and machine learning are used for real-time irrigation management [35]. The authors compared the performance of these two modeling approaches in

predicting soil moisture content and optimizing irrigation scheduling. The study finds that machine learning models generally outperform physics-based models in predicting soil moisture content, particularly when the models are trained using large data sets with high temporal resolution. However, the authors note that physics-based models can still be useful in certain contexts, such as when the available data are limited or when detailed knowledge of the physical processes involved is required.

5. Conclusions

The objective of this study was to evaluate different machine learning techniques for detecting and distinguishing heavy rainfall events in a sub-region of the Pangani River Basin in Northern Tanzania. The study employed five different algorithms to identify heavy rainfall occurrences between 1979 and 2014. The models' performance was assessed using precision–recall metrics and F-score to determine the most suitable method for the task. Based on the evaluation results, random forest and XGBoost demonstrated superior overall performance. However, it was observed that the multi-layer perceptron (MLP) performed better in identifying heavy rainfall events, which are the leading cause of floods in the Pangani River Basin.

The study's results suggest that MLP, despite being outperformed by other algorithms in overall performance, was the most effective technique for identifying heavy rainfall events, highlighting the significance of precision and recall in detecting the minority class in imbalanced classification. The highly imbalanced class distribution in the data set makes it challenging to identify heavy rainfall events, making the use of MLP a vital approach in the process.

The findings of this study align with previous research that has emphasized the importance of selecting appropriate performance metrics to evaluate algorithms' effectiveness in detecting rare events. Moreover, the study contributes to the literature by demonstrating that the MLP approach is well-suited for recognizing heavy rainfall events in the Pangani River Basin. The research provides valuable insights into the potential of machine learning algorithms in identifying heavy rainfall events, enabling policymakers to take proactive measures in flood management and control. To the government of Tanzania, the study recommends that the ministry responsible for monitoring flood and water levels in rivers and other water bodies should collect water level data with respect to weather parameters. This will enable the replication of the developed model in other rivers and assist in future studies in similar areas.

Overall, these models have potential applications in various fields that require accurate predictions of rare events, such as climate prediction, disaster management, and risk assessment. However, further research is needed to explore the models' performance in different settings and under different conditions, such as changes in climate patterns and data sources.

Future research could also focus on developing more robust models that can handle highly imbalanced data sets and improving feature engineering techniques to enhance model performance. Additionally, ensemble techniques and meta-learning approaches could be explored to improve the models' generalization and transfer learning abilities. Overall, the study provides insights into the potential of machine learning algorithms in predicting rare events and highlights the need for further research to develop more accurate and robust models.

Author Contributions: Conceptualization, L.M. and E.M. (Esteban Municio); methodology, L.M.; software, L.M.; validation, L.M., E.M. (Esteban Municio) and Y.D.B.; formal analysis, L.M.; investigation, L.M.; resources, L.M. and E.M. (Esteban Municio); data curation, L.M.; writing—original draft preparation, L.M. and E.M. (Esteban Municio); writing—review and editing, E.M. (Esteban Municio), E.L., J.L. and Y.D.B.; visualization, L.M.; supervision, E.L., J.L. and E.M. (Erik Mannens). All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Flemish Interuniversity Council for University Development Cooperation (VLIR-UOS), Belgium (Grant number ZIUS2013AP029), through an institutional cooperation programme (IUC) with the Nelson Mandela African Institution of Science and Technology (NM-AIST), under the research project 'Institutional strengthening: ICT, Library and CIC maintenance for collecting, analyzing big data'.

Data Availability Statement: At the moment, data supporting reported results can be found on request; they will be publicly available at the end of the main project.

Acknowledgments: The authors would like to thank the Tanzania Meteorological Agency and Pangani River Board for providing some of the data for the study and technical assistance during field activities in the Pangani basin.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

XGBoost	eXtreme Gradient Boost
KNN	k-Nearest Neighbors
ML	Machine Learning
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
PBWB	Pangani Basin Water Board
TMA	Tanzania Meteorological Agency
KWK	Karanga–Weruweru–Kikavu
OvR	One-vs-the-Rest
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve

References

1. World Health Organization. Floods. 2021. Available online: https://www.who.int/health-topics/floods (accessed on 17 February 2023).
2. Jonkman, S.N. Global perspectives on loss of human life caused by floods. *Nat. Hazards* **2005**, *34*, 151–175. [CrossRef]
3. Tanzania Meteorological Agency. Annual Technical Report on Meteorology, Hydrology and Climate Services 2020–2021 Update. 2021. Available online: https://www.meteo.go.tz/uploads/publications/sw1628770614-TMA%20BOOK%202020%20-2021%20UPDATE.pdf (accessed on 17 February 2023).
4. Kimambo, O.N.; Chikoore, H.; Gumbo, J.R. Understanding the Effects of Changing Weather: A Case of Flash Flood in Morogoro on January 11, 2018. *Adv. Meteorol.* **2019**, *2019*, 8505903. [CrossRef]
5. Nayak, M.A.; Ghosh, S. Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. *Theor. Appl. Climatol.* **2013**, *114*, 583–603. [CrossRef]
6. Parmar, A.; Mistree, K.; Sompura, M. Machine learning techniques for rainfall prediction: A review. In Proceedings of the International Conference on Innovations in information Embedded and Communication Systems, Coimbatore, India, 17–18 March 2017; Volume 3.
7. Stein, L.; Pianosi, F.; Woods, R. Event-based classification for global study of river flood generating processes. *Hydrol. Process.* **2020**, *34*, 1514–1529. [CrossRef]
8. Pham, Q.B.; Yang, T.C.; Kuo, C.M.; Tseng, H.W.; Yu, P.S. Combing random forest and least square support vector regression for improving extreme rainfall downscaling. *Water* **2019**, *11*, 451. [CrossRef]
9. Grazzini, F.; Craig, G.C.; Keil, C.; Antolini, G.; Pavan, V. Extreme precipitation events over northern Italy. Part I: A systematic classification with machine-learning techniques. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 69–85. [CrossRef]
10. Szczepanek, R. Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost. *Hydrology* **2022**, *9*, 226. [CrossRef]
11. Davenport, F.V.; Diffenbaugh, N.S. Using machine learning to analyze physical causes of climate change: A case study of US Midwest extreme precipitation. *Geophys. Res. Lett.* **2021**, *48*, e2021GL093787. [CrossRef]
12. Fernández-Delgado, M.; Cernadas, E.; Barro, S.; Amorim, D. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.* **2014**, *15*, 3133–3181.

13. Khoshgoftaar, T.M.; Golawala, M.; Hulse, J.V. An Empirical Study of Learning from Imbalanced Data Using Random Forest. In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), Patras, Greece, 29–31 October 2007; Volume 2, pp. 310–317. [CrossRef]
14. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2009**, *39*, 539–550. [CrossRef]
15. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
16. Wang, C.; Deng, C.; Wang, S. Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognit. Lett.* **2020**, *136*, 190–197. [CrossRef]
17. Pilario, K.E.; Shafiee, M.; Cao, Y.; Lao, L.; Yang, S.H. A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring. *Processes* **2020**, *8*, 24. [CrossRef]
18. He, Y.; Ma, J.; Ye, X. A support vector machine classifier for the prediction of osteosarcoma metastasis with high accuracy. *Int. J. Mol. Med.* **2017**, *40*, 1357–1364. [CrossRef]
19. Chychkarov, Y.; Serhiienko, A.; Syrmamiikh, I.; Kargin, A. Handwritten Digits Recognition Using SVM, KNN, RF and Deep Learning Neural Networks. In Proceedings of the Fourth International Workshop on Computer Modeling and Intelligent Systems (CMIS), Zaporizhzhia, Ukraine, 27 April 2021.
20. Mcroberts, R. A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. *Remote Sens. Environ.* **2009**, *113*, 532–545. [CrossRef]
21. Azorin-Molina, C.; Ali, Z.; Hussain, I.; Faisal, M.; Nazir, H.M.; Hussain, T.; Shad, M.Y.; Mohamd Shoukry, A.; Hussain Gani, S. Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model. *Adv. Meteorol.* **2017**, *2017*, 5681308. [CrossRef]
22. Dinku, T.; Ceccato, P.; Grover-Kopec, E.; Lemma, M.; Connor, S.J.; Ropelewski, C.F. Validation of satellite rainfall products over East Africa's complex topography. *Int. J. Remote Sens.* **2007**, *28*, 1503–1526. [CrossRef]
23. Hamis, M.M. Validation of Satellite Rainfall Estimates Using Gauge Rainfall Over Tanzania. Master's Thesis, University of Nairobi, Nairobi, Kenya, 2013.
24. Lu, S.; ten Veldhuis, M.C.; van de Giesen, N. *Evaluation of Four Satellite Precipitation Products over Tanzania*; EGU General Assembly Conference Abstracts: Vienna, Austria, 2018; p. 1403.
25. Cook, J.; Ramadas, V. When to consult precision-recall curves. *Stata J.* **2020**, *20*, 131–148. [CrossRef]
26. Li, W.; Guo, Q. Plotting receiver operating characteristic and precision–recall curves from presence and background data. *Ecol. Evol.* **2021**, *11*, 10192–10206. [CrossRef] [PubMed]
27. Erenel, Z.; Altincay, H. Improving the precision-recall trade-off in undersampling-based binary text categorization using unanimity rule. *Neural Comput. Appl.* **2012**, *22*, 83–100. [CrossRef]
28. Brabec, J.; Komárek, T.; Franc, V.; Machlica, L. On Model Evaluation Under Non-constant Class Imbalance. In Proceedings of the Computational Science—ICCS 2020, Amsterdam, The Netherlands, 3–5 June 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 74–87.
29. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
30. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *30*, 25–36.
31. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
32. Davis, J.; Goadrich, M. The Relationship between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 233–240. [CrossRef]
33. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
34. Chen, C.; He, W.; Zhou, H.; Huang, Y.; Sun, H. A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. *Sci. Rep.* **2020**, *10*, 3904. [CrossRef]
35. Gumiere, S.J.; Camporese, M.; Botto, A.; Lafond, J.A.; Paniconi, C.; Gallichand, J.; Rousseau, A.N. Machine Learning vs. Physics-Based Modeling for Real-Time Irrigation Management. *Front. Water* **2020**, *56*. [CrossRef]