water

Article

Data-Driven Community Flood Resilience Prediction

Moustafa Naiem Abdel-Mooty [1,*], Wael El-Dakhakhni [2] and Paulin Coulibaly [3]

[1] Department of Civil Engineering, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4L7, Canada
[2] INTERFACE Institute for Multi-Hazard Systemic Risk Studies, Department of Civil Engineering and School of Computational Science and Engineering, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4L7, Canada; eldak@mcmaster.ca
[3] NSERC FloodNet, Department of Civil Engineering, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4L7, Canada; couliba@mcmaster.ca
* Correspondence: abdelmom@mcmaster.ca

Abstract: Climate change and the development of urban centers within flood-prone areas have significantly increased flood-related disasters worldwide. However, most flood risk categorization and prediction efforts have been focused on the hydrologic features of flood hazards, often not considering subsequent long-term losses and recovery trajectories (i.e., community's flood resilience). In this study, a two-stage Machine Learning (ML)-based framework is developed to accurately categorize and predict communities' flood resilience and their response to future flood hazards. This framework is a step towards developing comprehensive, proactive flood disaster management planning to further ensure functioning urban centers and mitigate the risk of future catastrophic flood events. In this framework, resilience indices are synthesized considering resilience goals (i.e., robustness and rapidity) using unsupervised ML, coupled with climate information, to develop a supervised ML prediction algorithm. To showcase the utility of the framework, it was applied on historical flood disaster records collected by the US National Weather Services. These disaster records were subsequently used to develop the resilience indices, which were then coupled with the associated historical climate data, resulting in high-accuracy predictions and, thus, utility in flood resilience management studies. To further demonstrate the utilization of the framework, a spatial analysis was developed to quantify communities' flood resilience and vulnerability across the selected spatial domain. The framework presented in this study is employable in climate studies and patio-temporal vulnerability identification. Such a framework can also empower decision makers to develop effective data-driven climate resilience strategies.

Keywords: community resilience; data-driven methods; machine learning; resilience; flood hazard

check for updates

Citation: Abdel-Mooty, M.N.; El-Dakhakhni, W.; Coulibaly, P. Data-Driven Community Flood Resilience Prediction. *Water* **2022**, *14*, 2120. https://doi.org/10.3390/w14132120

Academic Editors: Stefano Morelli, Veronica Pazzi and Mirko Francioni

Received: 5 May 2022
Accepted: 28 June 2022
Published: 2 July 2022

1. Introduction

The severity of climatological and hydrological hazards has been increasing over the past decades, with an especially higher frequency of flood hazard over the past three decades, heavily impacting the livelihood of exposed communities [1–3]. The changing climate has been significantly affecting the weather conditions and climatological factors (i.e., mean temperature, humidity, and precipitation) [4,5]. Data records since 1996 show that in North America, and similarly around the world, the rate of extreme weather events and rainfall (i.e., more than 100 mm of rainfall in 24 h) is alarmingly increasing, accompanied by an increased frequency of floods [6]. This is attributed to the higher rate of urbanization into flood-prone areas, where the urban environment now hosts over 50% of the world's population, with an expected increase to 70% by the year 2050, boosting the probability of flood-related disasters through the vulnerable community's exposure [7,8].

As a direct consequence of such increase in flood exposure and related losses, flood disaster management stakeholders have been moving to adopt a proactive risk-mitigation response, rather than a reactive post-disaster response approach [9,10]. However, flood risk needs first to be quantified in order to efficiently develop better mitigation strategies and eventually enhance resilience. In this respect, flood risk is identified as the expected damage (i.e., consequence), resulting from a hazard's probability of occurrence, coupled with the at-risk-community's exposure and vulnerabilities, considering different uncertainties [11–13].

With the increasing climatological disasters and flood risk, community resilience research is steadily gaining more traction worldwide. While a community is defined as a "Place designated by geographical boundaries that function under the jurisdiction of a governance structure (e.g., town, city, or county)" [14], community resilience is the ability of a community to adapt to, predict, and rapidly recover from future disruptions, back to a predefined target state [14]. Flood risk is a result from the simultaneous realization of three aspects: (i) flood hazard: the potential, or probability, of a flood event of certain characteristics occurring at a given location, (ii) flood vulnerability: a measure of the susceptibility, and the adaptability, of the exposed community to the flood hazard, and finally (iii) flood exposure: the assets, humans, and otherwise (i.e., infrastructure systems) that are located in a flood-prone area [11,13,15]. This indicates that a severe flood hazard does not necessarily yield a high-risk flood, as it can occur in an area with a low number of exposed elements, but flood risk can be quantified only when the exposed and vulnerable community prone to said hazard is coupled with the hazard realization [12,15]. As an extension, resilience analysis evaluates the extended functionality loss and recovery trajectory of communities prone to flood hazards, taking into account the direct and indirect losses as well as restoration costs [5,12].

Previously, resilience has been defined differently across different fields; however, in the context of this study, resilience is defined as the ability to resist being affected by, and rapidly recover from, some external disturbance [16]. Resilience is quantified through the four attributes including: two objectives (i.e., goals) of resilience: robustness and rapidity, enabled by two means: resourcefulness and redundancies [17,18]. Robustness is the inherent ability of the system to retain its functionality level when exposed to stress or extreme demand; rapidity is the time needed for the system to bounce back to a certain predefined target functionality level; resourcefulness is the availability of adequate resources within the system to maintain its functionality under extreme demand levels, and finally, redundancy is the availability of alternate components to maintain functionality during the external hazard [17,19]. It is worth noting that rapidity measures the total time needed for the system to bounce back to its target functionality, including the downtime of the system (i.e., the duration of the hazard itself).

Over the years, numerous researchers have embarked on flood categorization and prediction studies [20–23]. However, most such studies focused on the hazard's features and, to a lesser extent, on the direct impact and losses due to the flood hazard or long-term recovery cost and time [24–29]. In this respect, this study aims at developing a prediction framework that classifies the long-term potential impacts, recovery, and resilience of the exposed community, a categorization that captures the resilience of the exposed communities rather than simply the hazard's characteristics. To achieve that, having reliable data is imperative to accurately incorporate said damage and characteristics within an objective data-driven resilience prediction framework [30]. The incorporation of the hazard, system vulnerability, and exposure employed in this framework would result in a comprehensive assessment of the short-term potential impacts, direct and otherwise, of the flood event through robustness assessment (i.e., flood risk), as well as the long-term impact on the exposed community through rapidity evaluation (i.e., resilience assessment). The study presented herein is employable in vulnerability identification and flood prediction studies, providing an imperative decision support tool for stakeholders and policymakers to allocate adequate resources and potentially save billions of dollars.

2. Flood Resilience Prediction Framework

2.1. Framework Design and Layout

The aim of this research is to develop a flood resilience prediction framework that captures the probable and resulting impacts of floods on respective exposed communities. Such a framework would serve as a practical data-driven tool for quick and actionable early-warning system. Such a system will subsequently aid policy and decisionmakers in developing resilience-guided risk management strategies, accounting for the four attributes of resilience. Classification and data driven models require a sufficient number of observations in a dataset to allow for meaningful classification and clustering [23]. While this necessitates the accessibility to a large volume of high quality data, there are also alternative ways to account for missing data within an employable dataset.

As can be seen in Figure 1, the framework presented herein is comprised of two main parts: (a) resilience-based categorization and (b) resilience-based prediction, and each part of the framework is comprised of different stages.

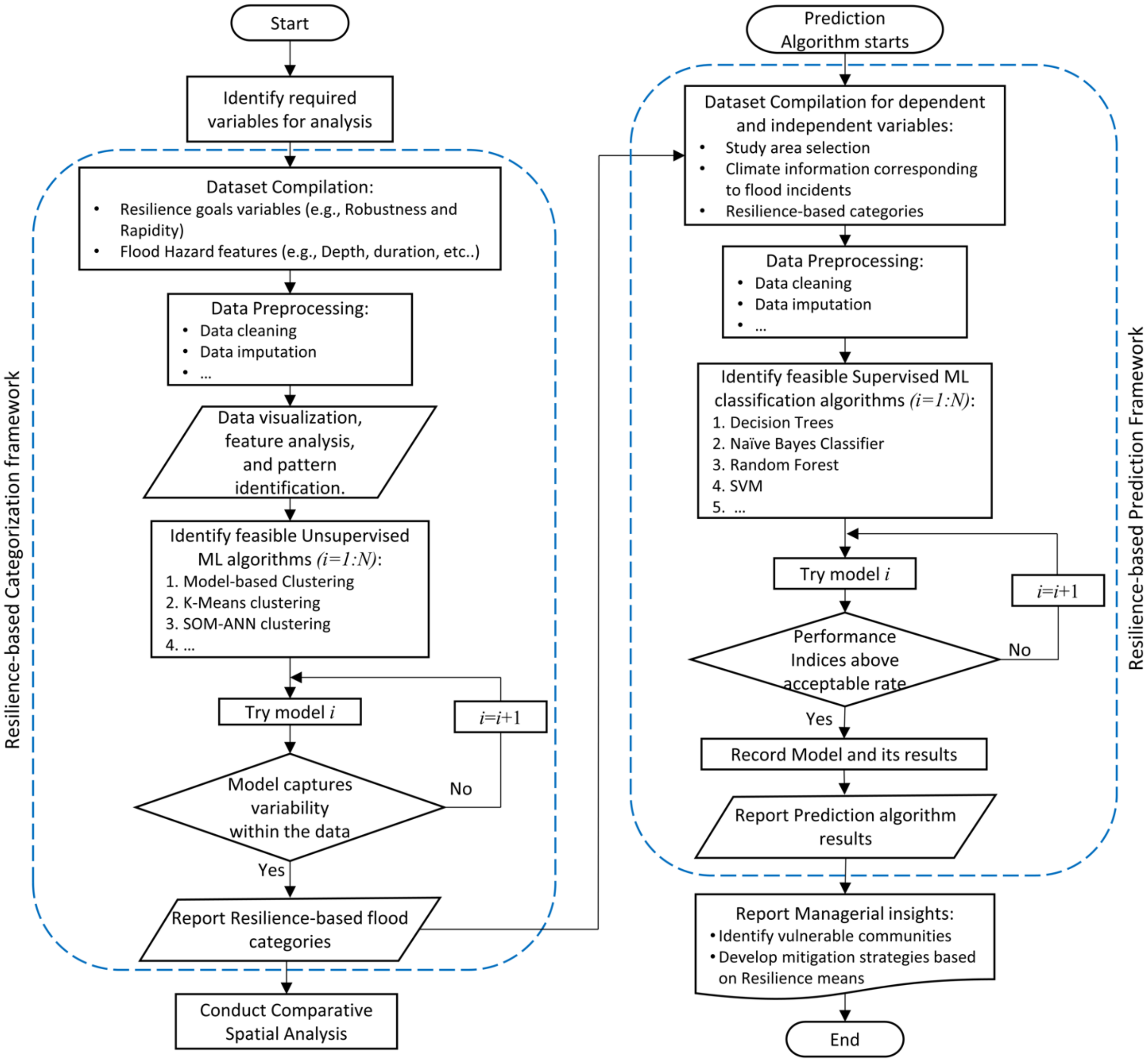
Start
Identify required variables for analysis
Dataset Compilation:
• Resilience goals variables (e.g., Robustness and Rapidity)
• Flood Hazard features (e.g., Depth, duration, etc..)
Data Preprocessing:
• Data cleaning
• Data imputation
• …
Data visualization, feature analysis, and pattern identification.
Identify feasible Unsupervised ML algorithms (i=1:N):
1. Model-based Clustering
2. K-Means clustering
3. SOM-ANN clustering
4. …
Try model i
i=i+1
Model captures variability within the data
No
Yes
Report Resilience-based flood categories
Conduct Comparative Spatial Analysis
Resilience-based Categorization framework
Prediction Algorithm starts
Dataset Compilation for dependent and independent variables:
• Study area selection
• Climate information corresponding to flood incidents
• Resilience-based categories
Data Preprocessing:
• Data cleaning
• Data imputation
• …
Identify feasible Supervised ML classification algorithms (i=1:N):
1. Decision Trees
2. Naïve Bayes Classifier
3. Random Forest
4. SVM
5. …
Try model i
i=i+1
Performance Indices above acceptable rate
No
Yes
Record Model and its results
Report Prediction algorithm results
Report Managerial insights:
• Identify vulnerable communities
• Develop mitigation strategies based on Resilience means
End
Resilience-based Prediction Framework

Figure 1. Multi-stage framework layout for resilience-based flood categorization and prediction.

Part (a): resilience-based categorization framework: this part is divided into three main stages: Stage (i) Data compilation, cleaning, and visualization: the first step is to compile a comprehensive dataset, with enough variables to capture the resilience attributes, as well as the features of the flood events (e.g., flood depth and duration). Following data gathering, data preprocessing starts to ensure data suitability for a reliable analysis and data imputation for missing values. Datasets are investigated for the identification of any biases or skewness within the dataset, as well as the accommodation for missing data. Missing data can induce disruptions to the ML algorithm, rendering replacing or removing observations with missing variables. Accounting for missing variables can be performed through multiple approaches, 1) by removing observations with missing variables altogether, 2) by averaging the readings from other nearby observations with similar conditions to the observation with missing variables, or 3) by using unsupervised learning to cluster the dataset and take the average of the cluster variables as the reading for the missing variables. In this study, a combination of approaches 1 and 2 was employed [31–33]. Finally, data visualization was conducted to identify inherent characteristics and interdependencies within the dataset, which is pivotal in choosing an appropriate model for the following stage.

Stage (ii) Selection of Machine Learning (ML) model: ML models are designed to analyze high-dimensional data. They have been utilized across different fields such as engineering, biology, and medicine and in different applications such as banking, targeted advertisement, social networks, and image and pattern recognition [34–37]. ML models are used to identify pattens and discover behaviors in large datasets, while continuously adapting to new data features to enhance model performance. ML models are expected to handle large datasets with complex interdependent features and identify hidden patterns [38]. ML models are divided into supervised and unsupervised algorithms (also named classification and clustering algorithms, respectively) and will be discussed in more detail in the following section. In the developed framework, the categorization in part (a) employs unsupervised (clustering) techniques, while part (b) employs supervised (classification) algorithms [38,39].

Stage (iii) Features and clusters analysis: the results of Stage (ii) in Part (a) are used in developing the features of each category (cluster). By conducting a feature analysis, the developed clusters can be used in developing a spatial analysis to identify vulnerable communities based on the considered resilience metrics. The deployment of the clustering algorithm results ensures the development of unbiased managerial insights, facilitating the decision-making process for utilizing the resilience means (i.e., redundancies and resourcefulness) to better enhance the resilience of the more vulnerable communities. The developed clusters in Part (a) are vital in the development of the predictive analysis in Part (b), where this categorization framework can aid decision makers in translating predicted flood hazards and risks into actionable plans, increasing the robustness by reducing the loss of functionality, and ensuring a quick recovery to the target state.

Part (b): Resilience-based prediction framework: similar to Part (a), Part (b) is also comprised of different stages; while these stages are similar in concept with their counterparts in Part (a), the details and the nature of the algorithms differ greatly.

Stage (i) Data compilation: the first step is compiling the dependent and independent variables of the dataset. In this stage, the study area is identified for the development of the predictive model where the features, characteristics, and exposure are fairly similar. The dependent variables selected for this framework are the climate information corresponding to recorded flood events (e.g., maximum temperature, minimum temperature, precipitation, wind speed, air pressure, humidity, etc. . . .), whereas the independent variable would be the resilience-based categories developed in Part (a) of the proposed framework. Similar to most ML algorithms, the dataset should be comprehensive and of good quality and diversity to produce actionable results. Data imputation and cleaning are conducted to ensure the reliability of the data and avoid skewness and imbalances in the dataset.

Stage (ii) Data preprocessing and analysis: for this stage, the gathered dataset is studied to identify the interrelationship between the different variables and thoroughly examine which variables to be included in the analysis to reduce the noise in the data while ensuring that all the resilience metrics and the hazard features are comprehensively represented. This feature selection step can be achieved through exploratory and sensitivity data analyses, feature selection, or correlation analysis between different variables of the available data. Following that step, data cleaning and preprocessing commences. The performance of data-driven models is strictly tied to the quality and quantity of the dataset involved in the development of the model, whereas finding a readily available dataset that matches all the required criteria for analysis is typically very challenging. Therefore, numerous methods have been developed to deal with missing data, unbalanced data, and skewed data (e.g., data imputation, removing datapoints with missing variables, take average readings from nearby sources, etc.) [32,33].

Stage (iii) Development and testing of the ML models: in this stage, a supervised ML model is developed to predict flood resilience categories based on climate data corresponding to the recorded flood events. Supervised ML models can be used in predicting discreet, continuous, or categorical data. The classification required for the analysis herein falls under the multi-class classification category, where the dependent variables are used to predict a categorical independent variable of more than two classes (Wu et al., 2004). For this classification, different algorithms were validated and tested to determine the most suitable algorithm for the current dataset (e.g., Naïve Bayes classifier, Support Vector Machine, Decision Trees, Artificial Neural Networks, Ensemble techniques, etc.), where they were assessed based on a common performance criteria, which is to be explored further in the Methodology section [33,40–42].

2.2. *Methodology*

Machine Learning is an artificial intelligence tool designed to learn autonomously from a training dataset, mimicking the behavior of the human brain through the learning process. By deploying ML models on appropriate datasets, the model extracts the dataset's inherent features and adjusts itself to better enhance its performance [43]. As mentioned, ML models are broadly divided into two types, supervised and unsupervised learning models, where they use labelled and unlabeled data, respectively, for training and validation. In the field of natural hazard and community resilience, ML and data-driven models have been recently been employed in achieving the overarching goal of increasing community resilience in the face of natural and anthropic hazards [25,42–46]. For the framework developed herein, both ML model types are utilized, where the unsupervised learning is utilized in the development of the community resilience categories, and supervised ML techniques are employed to predict the community resilience metrics under future flood hazards.

2.2.1. Unsupervised Learning: Clustering

Unsupervised ML models use partitioning algorithms to cluster observations based on a predefined similarity measure such that observations with common features are placed in the same cluster [47]. This is an unguided process that does not require a predefined objective, ensuring that the clustering is based on inherent features of the dataset. This similarity measure is assessed by measuring the distance between different observations, where two, or more, observations are considered similar when the distance between them is minimal. Henceforth, observations within a cluster should be closer to one another than that of other clusters.

Choosing the similarity measure depends heavily on the type of data and objective of the study; such measures include the Euclidean, Cosine similarity, Manhattan, and Gower distances [48]. For this study, multiple similarity measures were explored to determine their applicability with the available mixed-type dataset (i.e., dataset containing both categorical and numerical data). For the Gower distance within the Partitioning Around Medoids algorithm, the developed dissimilarity matrix from the dataset was skewed, which results in a biased algorithm favoring seasonal clustering instead of resilience-based clustering. Eventually, weighted Euclidean distance was adopted in this study as it measures the weighted proximity of the observations within a three-dimensional space [48,49]. It is important to note that other approaches may also be employed in the current study.

For the framework presented herein, two clustering algorithms were employed to develop the resilience-based flood categories, namely *K*-means clustering and self-organizing Maps. The *K*-means clustering technique, and its variations, is the most heavily used partitioning (clustering) algorithm [50], where observations are divided into a predefined number of clusters (*K*). Prior to the partitioning algorithm, multiple values are assumed for *K*, and the optimal value is that with the minimum intra-cluster variation (i.e., the total within-cluster sum of squares (WSS)). For the current study, the WSS utilized the squared Euclidean distance between the observations and their respective cluster′s centroid [51–53].

SOM is a type of Artificial Neural Networks (ANN) algorithm trained to cluster data into groups in an unsupervised approach. The input space is organized according to a predefined topology of neurons, where each neuron is assigned a number of observations. ANN is an artificial intelligence technique by which complexinterrelationships within a dataset are uncovered automatically based on inherent patterns in the dataset [54,55] by mimicking the behavior of the human brain when transmitting signals through neurons, albeit through artificial neurons. There have been numerous ANN techniques developed to date, each of which may befit a specific application (e.g., self-organizing maps, recurrent neural networks, and feed-forward back-propagation neural networks). However, ANN is more commonly employed in predictive algorithms [54,56,57] and pattern recognition applications [23,36,55,58]. For the study presented herein, SOM was utilized using the Deep Learning Toolbox in MATLAB, where the Kohonen rule was adopted [55,59].

2.2.2. Supervised Machine Learning: Classification

Classification is a supervised ML technique that learns and utilizes features of a dataset to derive patterns and classify new input data. Supervised ML models learn from a training dataset, which is comprised of dependent (i.e., predictor variables) and independent variables (i.e., predictand variable) and applies the identified patterns on a testing dataset, while applying optimization techniques to increase the model's performance [41,60,61]. Numerous classification techniques have been developed to date (e.g., continuous, discreet, numerical, or categorical). In the present study, the independent variable is class-based; therefore, multiclass classification techniques will be employed in the current study (e.g., Naïve Bayes classifier, Classification Trees, Support Vector Machine, ANN, etc.). To improve the performance of said models, classification models employ ensemble techniques—bagging, random forest, or boosting [62–64].

Naïve Bayes Classification

The Naïve Bayes classifier algorithm employs Bayes' theorem with the assumption that the variables are conditionally independent given the value of the class variable (i.e., Naïve). The algorithm employs joint conditional probabilities of the dependent variable of the training dataset given their respective independent variable [65–67]. The output of said model is the conditional probabilities of the class labels assigned based on the highest class-label's joint probability for each observation in the dataset. The theorem employed in this algorithm calculates the conditional probability for class variable y using Equation (1), where $(x_1, \ldots, x_n)$ are the n dependent variables.

$$P(y|x_1,\ \ldots,\ x_n) = \frac{P(y)P(x_1,\ \ldots,\ x_n|y)}{P(x_1,\ \ldots,\ x_n)} \tag{1}$$

By applying the naïve assumption for all i, and substituting with $P(x_1,\ \ldots,\ x_n)$ as a constant, the resulting conditional probabilities can be expressed as Equation (2):

$$P(y|x_1,\ \ldots,\ x_n) \propto P(y)\Pi_{i=1}^{n}P(x_i\ |y) \tag{2}$$

This theorem can be interpreted such that a data record belongs to a certain class (M) when the conditional probability $P(\ M_i|x_1,\ \ldots,\ x_n)$ returns the highest value of all classes. The reader is referred to the studies by McCallum and Nigam (1998) [68] and Zhang (2004) [69] for further details on Naïve Bayes classification.

Decision Trees

Within the Classification and Regression Trees (CART) algorithm, classification trees are utilized to predict categorical (discriminate) data, unlike regression trees which deal with predicting continuous independent variables [41].

Decision Trees utilize a binary recursive partitioning algorithm, since each split (i.e., rule or partitioning step) depends on the prior splitting step. The data is partitioned into homogenous subgroups (i.e., nodes) using binary Yes-or-No questions about each feature of the sub-group, where this process is repeated until a suitable stoppage criterion is reached (e.g., maximum number of splits). For each split, the objective is to identify the optimum feature upon which the data can be split, where the overall error between the actual response and the predicted response is minimal. The analysis presented herein is concerned with classification trees, where the partitioning is set to maximize the cross-entropy or the Gini index [38,70]. The Gini index is a measure of purity (or lack thereof) in the classification model, where a small value indicates that a subgroup (i.e., node) contains predominantly observations from a similar class. High values of mean decrease in the Gini index correspond to a more important variable (i.e., feature) within the classification model [38]. The Gini index is relied upon herein given the type of data utilized in the demonstration application presented later in this study.

For model accuracy and performance enhancement, there exist numerous employable ensemble techniques (e.g., bagging, boosting, and random forest) [63,64]. Bagging is a bootstrap aggregating technique used for fitting multiple versions of the model drawn from the training dataset. Bootstrapping is a random sampling technique of the data, taken by replacement, such that a datapoint can still be available for selection in subsequent models while using all the predictors for the sampling technique [71]. Each model is then used to generate training for the DT model, and the averaging of all the predictions is subsequently used, resulting in a more robust model than a single tree [63,70,72].

Random forest further improves bagging techniques to enhance model performance, where the selection of the predictors is also randomized at each split at the node within the tree rather than using all the predictors. The size of the tree is maximized by repeating the aforementioned process iteratively, and the prediction is based on the aggregation of the prediction from the total number of trees [63,73–76].

Prediction Model Performance

For classification models, the overall model accuracy and misclassification errors are widely used. However, this criterion is not always suitable for asymmetrical or skewed datasets where the majority of the data falls within a single category. To introduce a more accurate measure of the predictive performance, the precision, recall, and F1-score for each category in the testing and training datasets were calculated. In this respect, precision is the number of correct predictions per class within multiclass classification, which is a measure of how accurate each class prediction is. Recall (i.e., sensitivity) on the other hand is the number of correct class predictions out of all correct examples in the dataset, and it captures the ratio between the correct classifications and the actual classification for the dataset. Finally, the F1-score is considered an integration between the precision and recall of the model, where it balances the concerns of both performance measures [77]. Precision, recall, and the F1-score are evaluated according to Equations (3)–(5), respectively, where the information can be extracted from the confusion matrix of each model.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{–}score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

In the equations above, *TP* refers to True Positive, which is the number of correctly predicted observations, and *FP* refers to False Positive, which is the number of predictions incorrectly assigned to a class, whereas *FN* refers to False Negative, which is the number of observations incorrectly assigned to a wrong class [60].

3. Framework Application Demonstration

To showcase the employability of the developed framework, the data from the National Weather Service (NWS) were adopted for the derivation of the resilience-based categories. Subsequently, these categories were then coupled with climate data extracted from the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information. The framework was thus applied to: (*i*) identify the features of the exposed communities along with their vulnerability using descriptive data analysis, (*ii*) identify interdependence between different features of the adopted dataset to appropriately choose a suitable ML model, (*iii*) categorize the communities' flood resilience by combining flood features with resilience metrics within the dataset (i.e., robustness and rapidity), and (*iv*) test the model performance in terms of accurately predicting the communities' resilience when exposed to flood hazard, using climate data as predictand.

The earlier work presented in the study by Abdel-Mooty et a. (2021) [59] serves as a foundation for the categorization stage of the prediction framework developed herein. In their study, Abdel-Mooty et al. (2021) developed a flood resilience categorization, resulting in five community flood resilience categories. These categories are thus employed through the second stage of the framework developed in the current study. In the following section, a brief summary of their findings is presented, followed by a description of the flood prediction demonstration.

3.1. Part (a): Resilience-Based Categorization

In the first stage, the dataset compiled by the NWS was employed. This dataset is one of the longest-run annual flood damage recorded in the United States [78]. The data were gathered through third party organizations and directly reported to the NWS database according to the predefined guidelines. As such, the quantity and quality of the gathered data is governed by the available resources (e.g., time and funding availability) of said agencies [78]. The dataset contains records of flood events occurring across the United States between 1996 and 2019. The related damages, time, geographical center, month, and year for each recorded flood event are compiled within this database [78,79]. Within the dataset, the recorded damage was divided into property and crop damages, which were subsequently combined into a single variable within the analysis named *Monetary Damages*. It is worth noting that the damages recorded in this dataset pertain to only the direct damage resulting from the flooding water on the exposed assets and does not consider the indirect (cascade) damages (e.g., opportunity loss). Within the present dataset, the term "flood event" refers to only the flooding aspect of any natural disaster. Despite the aforementioned limitations, this dataset is still considered one of the best resources for flood damage records in the United States [30,79]. Figure 2a shows a temporal analysis, while Figure 2b shows a spatial analysis of the flood events occurring within the same period, where the numbers on each state are the number of recorded floods, and the colors are used to indicate the relative total monetary damage of each state. This analysis shows that the largest number of records and the largest monetary damage are within the state of Texas. This is attributed to the increased heat content over the western Gulf of Mexico, as it produces higher humidity and temperatures. This heat content is directly proportional to the precipitation resulting from different storms [80] and can also be attributed to the tropical weather region that Texas falls within, given that this region is susceptible to a large number of devastating hurricanes and extreme rainfall, coupled with the increased exposure caused by the increased urbanization rate [21].

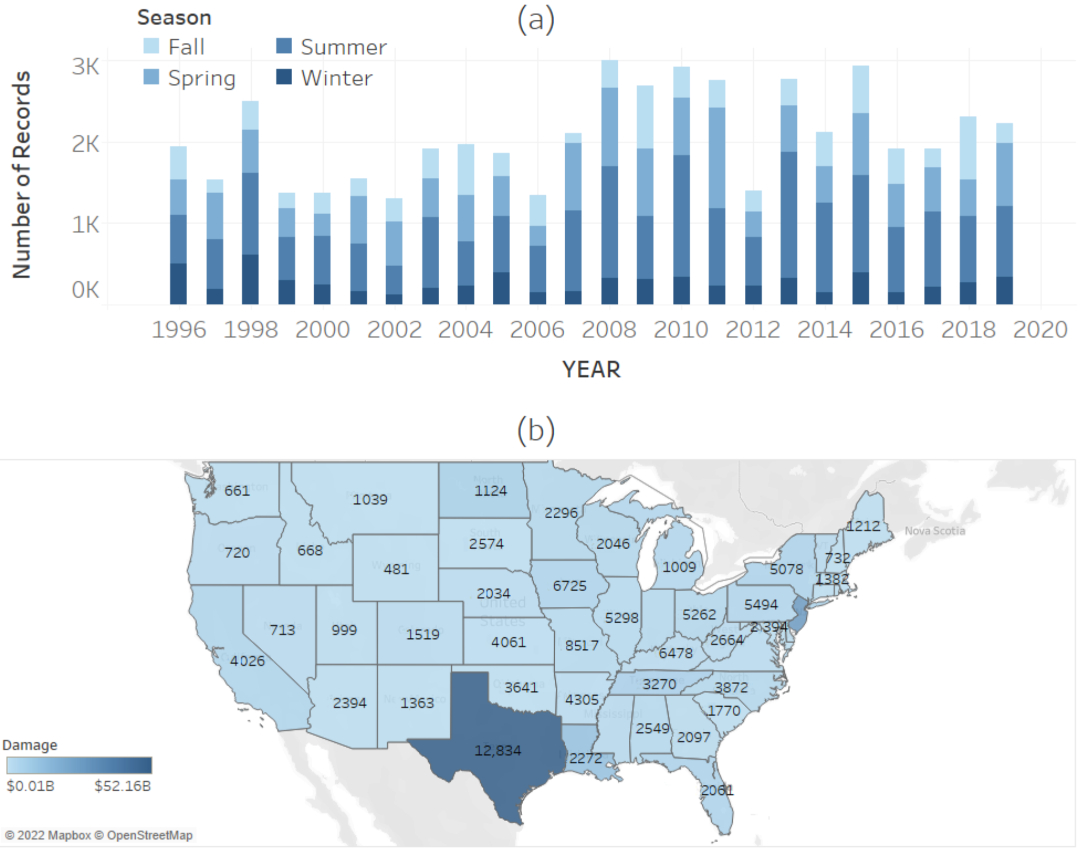

Figure 2. Descriptive spatio-temporal analysis of the employed dataset where (**a**) the annual number of floods between 1996 and 2019 indicated by season and (**b**) a multilayer spatial analysis of the dataset with the total number of records and the total damage in USD per state indicated by color.

Considering the objective of the current study, incorporating resilience metrics is key in identifying resilience-based categories. As such: (*i*) flood records that did not cause any monetary damage, injuries, or fatalities were excluded from the dataset, as they will not produce any resilience metrics to measure and will induce bias within the categorization model, and (*ii*) property and crop damage were summed up into a total monetary damage, and as mentioned earlier was adjusted to accommodate the inflation rate over the years using the Customer Price Index from the Bureau of Labor Statistics [81]. This monetary damage, along with the injured people and fatalities, represent the robustness of the exposed community, while the duration of the flood event represents downtime of the exposed community, which is a component of the rapidity metric.

The analysis showed that: (*i*) flood events that occurred during the spring were split into two categories based on their impacts, (*ii*) flood events causing longer disruptions were separated in a separate cluster, identifying a correlation between event duration and the impact of the flood event on the exposed community (i.e., relating robustness with rapidity and overall resilience), and (*iii*) flood events that resulted in the loss of human lives were clustered together. Events falling in Categories 1, 2, and 4 are more common than Categories 3 and 5 in terms of annual number of events. Given the multidimensional nature of resilience, more emphasis in the analysis was placed on the value of human injuries and fatalities than monetary loss. As such, although events in Category 3 follow those of Category 5 in terms of average damage per event, events falling in Category 4 follow those of Category 5 in terms of average affected people per event; hence, it was assigned a higher category than Category 3. It should be recalled that the event duration mentioned herein is the hazard's duration, which represents the down time of the community before the initiation of recovery efforts, representing a part of the total rapidity of the community. It is also worth noting that a longer flood duration corresponds to a less robust infrastructure system (e.g., drainage networks) to accommodate the hazard's capacity effectively, resulting in a lower overall resilience of the exposed community. The results were analyzed for the inherent features of each cluster, and each category was assigned a Flood Resilience Index (FRI) that increases gradually as the robustness decreases (i.e., functionality loss increases). As such, communities that are exposed to flood disasters with impacts falling in Category Ɱ are more resilient than those of Category Ɱ + 1, with Ɱ having values between one and four. A detailed description of the categories can be found in Table 1. It is worth noting that a community can be placed in a different category each time it is exposed to a flood disaster; however, by averaging all the resilience indices subsequent to the corresponding recorded flood disasters, an average index can be assigned to that community, comprehensively representing its overall resilience while accounting for all the previous disasters. The reader is referred to the study by Abdel-Mooty et al. (2021) for more details on the resilience-based categories employed herein.

Table 1. The community flood resilience-based categories.

Community Flood Resilience Category	Title 2
1	Communities exposed to events that occur in the summer, causing disturbance less than 264 h (11 days) and/or causes up to 250 injuries and damage less than $2.5B without fatalities
2	Communities exposed to events that occur in the spring, causing any disturbance duration, causes up to 20 injuries and damage up to $1.5B without fatalities
3	Communities exposed to events occurring in any season, causing disturbance more than 264 h (11 days), and causing up to 250 injuries with any damage value and without fatalities
4	Communities exposed to events that occur in winter or fall, causing disturbance less than 264 h (11 days) causes up to 250 injuries and damage up to $2.5B without fatalities
5	Communities exposed to events occurring in any season, causing any disturbance duration that results in more than 250 injuries, causing damage more than $2.5B, with fatalities, and Communities exposed to events occurring in the spring that are not under class 2

3.2. Part (b): Resilience-Based Prediction

For this stage of the framework, a smaller geographical location needed to be identified such that the meteorological features of the dataset would be comparable, comprehensively representing the seasons and their respective hazard for said communities. This was also needed such that the built environment would match its respective hazard, given that different seasons (and subsequently the characteristics of the natural hazard) differ drastically across the United States (e.g., the winter in Michigan is drastically different than that of Florida and Texas). However, the framework is applicable on any location within the United States mainland as long as it is included in the development of the indices in part (a) of the framework. By inspecting Figure 2, as mentioned earlier, the state of Texas had the most recorded number of flood disasters between 1996 and 2019, and the most recorded damage as well. The high number of records is suitable for the development of the prediction model, as the model will need a large dataset for development, training, and testing. As such, the state of Texas was selected for the development of the prediction stage of the framework. The disaster database that recorded between 1996 and 2019 in the state of Texas was paired with the developed categories in Table 1 on a county level, where each event was assigned an index across the different counties, and the average index was calculated and assigned for each county. Figure 3 shows the spatial distribution of the total number of recorded disasters and average FRI across the counties. The spatial analysis shows a low correlation between the number of events and the FRI of a community, given that the more common flood events are those of low severity [11]. It is also worth mentioning that the spatial analysis shows a concentration of high FRI across the coastal area around the Gulf of Mexico. This can be attributed to high-tide flooding, which is becoming increasingly common in recent years as a result of relative increase in sea level [82]. According to NOAA, coastal communities are witnessing an increase in high-tide flooding, with some areas reporting a rapidly increasing rate [82,83]. This can also be attributed to the nature of the natural hazards affecting the area, where a damage of $6B was recorded in 2018, and the devastating Hurricane Harvey, which affected the entire state in 2017, causing an extreme rainfall event resulting in widespread devastation across different counties. The total damage from Hurricane Harvey reached $128.8B, leading to one of the most expensive natural disasters in modern history [82–84]. The spatial analysis presented in Figure 3 is also in line with the Cartographic Maps of Precipitation Frequency Estimates published by NOAA in Atlas 14 Volume 11 of Texas in 2018, showing an increased precipitation frequency and magnitude over the coastal area with the Gulf of Mexico [85].

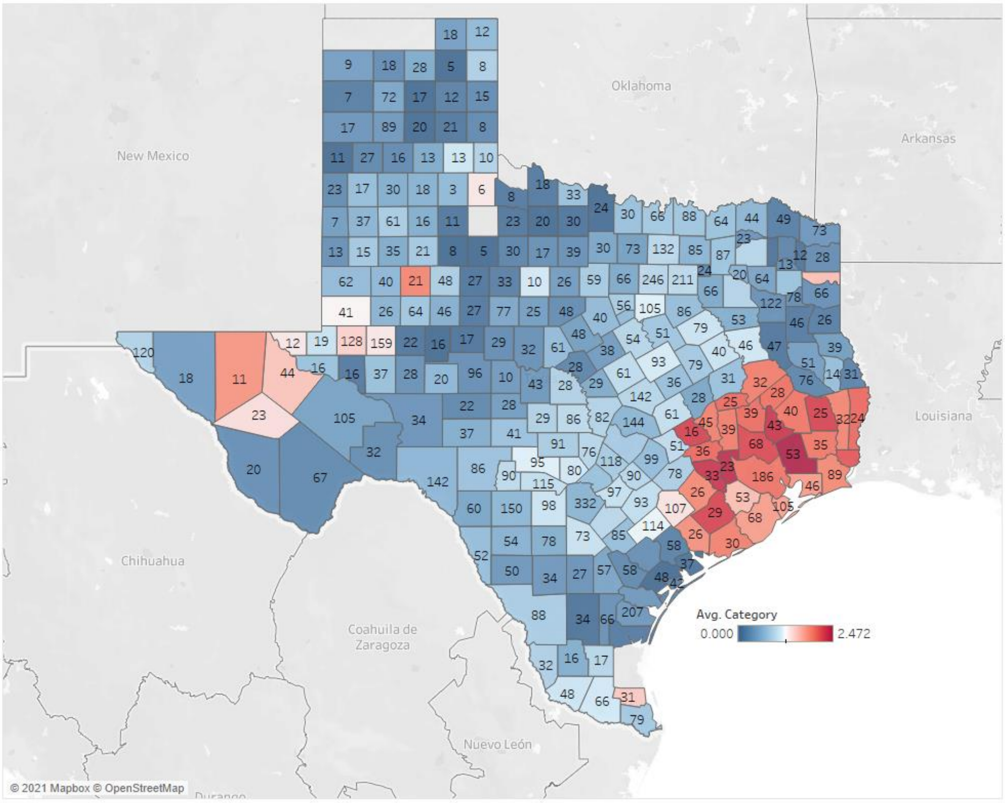

Figure 3. Spatial distribution of the number of records and the average FRI over different counties in the state of Texas.

3.3. Managerial Insights and Results

To complete the dataset for the prediction framework, climate information corresponding to each recorded flood event in each county was then extracted from the Global Historical Climatology Network (GHCN-Daily) under the National Center for Environmental Information [86,87]. To draw reliable insights from the proposed methodology, a comprehensive dataset must be present that includes all the pertinent variables with enough observations over the years to avoid biases. However, the present dataset implicitly presents this information through the spatio-temporal characteristics of the flood events when exposed to their relative communities.

The extracted climate data, as available, contained four variables for each recorded flood event: Maximum Daily Temperature, Minimum Daily Temperature, Average Daily Temperature, and Maximum Recorded Precipitation. These variables were then employed as predictors (dependent variables) for the FRI resulting from the recorded flood events (independent variable) to be used in the development of the prediction model. The dataset is subsequently divided into two subsets—Training and Testing (70% and 30%, respectively). The training subset was used in the development and training of the ML model, where the FRI implicitly contains information about the resilience (i.e., robustness and rapidity) of the exposed communities, and the climate variables contain information on the climatological features of the location, weather extremes, and different attributes, and causes, of the flood hazard. This comprehensive dataset was then inspected using exploratory data analysis and correlation plots, as shown in Figure 4. This figure presents a 5 × 5 matrix, in which the variables are labelled on the columns and rows. The matrix contains four information groups: (*i*) frequency scatter plots located at the lower triangle of the matrix, excluding the last column; (*ii*) smoothed frequency curves located at the diagonal of the matrix, where the last cell at the bottom right is a histogram for the categorical variable; (*iii*) correlation coefficients located at the upper triangle of the matrix, excluding the last column; and finally (*iv*) box plots located at the last column of the matrix. It is worth noting that this figure also presents statistical data analyses, as it shows the statistical distribution of the dataset within its variable space as well as the correlation between different variables. The box plots in Figure 4 show that the maximum, minimum, and average temperature variables are overlapping, evenly distributed and with a low range of outliers. This indicates that these variables are interdependent, which shows a consistency in the climatological features of the selected geographical study area. This is also supported by the correlation coefficients as the correlation between these variables is high across all the FRI categories. However, the precipitation variables contain heavy-tailed distribution with a larger range for the outliers, indicating an exceptionally large surge in the value of precipitation, which leads to the recorded flood events. The latteris supported by the correlation coefficient values between precipitation and other indices, especially at FRI-1, where the severity of the flood event is low, yet the frequency of occurrence is high [59]. This analysis supports the need to use ML models over traditional statistical learning models, as ML models are better equipped to deal with complex interdependent data for numerous applications [59,88].

3.4. Model Performance and Discussion

For this analysis, multiple ML classification models were tested, namely, Bagged Decision Trees (DT), and Random Forest (RF) Techniques as ensemble-type models, and Naïve Bayes (NB) classification. The dataset was split as mentioned earlier to training and testing datasets, where the split was chosen randomly to ensure a homogenous distribution of the data in both subsets since the dataset is not evenly distributed along all FRI categories. In this analysis, (*i*) Bagged DT with 1000 bootstrap replications was used in as an ensemble method, with a minimum split of four; (*ii*) RF models with a wide range of trees up to 6000 was tested, and while all of them had similar performances, two models were highlighted in this study—RF with 300 trees and RF with 1000 trees—both with four variables randomly sampled at each split and a shrinkage parameter of 0.01 (referred to herein as RF 300 and RF 1000, respectively); and finally (*iii*) Naïve Bayes classification, as discussed earlier, with a 70–30% split between training and testing data subsets. Each of the aforementioned models have their own assessment measures for model performance (e.g., Gini impurity, entropy measure for DT, Mean Square Error, etc.). As such, other performance evaluation indices were utilized in this analysis to objectively compare the predictive performance in replicating the testing data subset of the employed algorithms. To that end, the precision, recall, and F10-score have been employed per Equations (3)–(5), respectively. The performance indices can be seen in Table 2; the accuracy and misclassification for all the models are compared, where it can be seen that the models perform adequately (for training subset: 53.8%, 97.8%, 98.2%, and 98.2% for NB, RF 300, RF 1000, and Bagged DT, respectively, and for the testing subset: 50.9%, 57.9. 57.8%, and 57.3% for the NB, RF 300, RF 1000, and Bagged DT, respectively). It can be concluded that the DT ensemble models are over-trained in the training dataset but perform better than the NB classifier in the testing dataset even if the results are comparable. This proves the need for a better performance measure for the class in each model—as seen in Table 2, the precision, recall, and F1-score for the training and testing subsets across all the classes. Figure 5 shows an enhanced visual inspection of the performance indices of the four models, where it can be concluded that the performance of the NB classification model is inferior to the ensemble techniques in terms of correctly classifying the data; this can be attributed to the fact that NB models perform better with smaller datasets, as they follow the laws of independent probabilities, indicating it does not perform well with correlated data [89]. In the training subset, the precision, recall, and F1-score for the ensemble models (i.e., Bagged DT, RF 300, and RF 1000) do not fall below 85% for all classes, which indicates a very good fit for the employed dataset. However, in the testing subset, the results vary for each category. While the results are overall satisfactory for all the ensemble models, the Bagged DT model had better performance when it came to Category 5 (RF models resembled 23% of the precision of the Bagged DT), where the data points falling in this category were scarce compared to the other categories. However, the RF models outperformed the Bagged DT in the precision of Category 3 (65% for the RF models compared to 20% for the Bagged DT model), indicating that random sampling for the variables in addition to the observations in the training algorithm yielded more favorable results than the Bagged DT. The results displayed in Table 2 and Figure 5 show that even though the models are comparable, given the importance of correctly classifying flood events falling in Category 5 due to its severity and impact, the Bagged DT is thus preferred over the RF models.

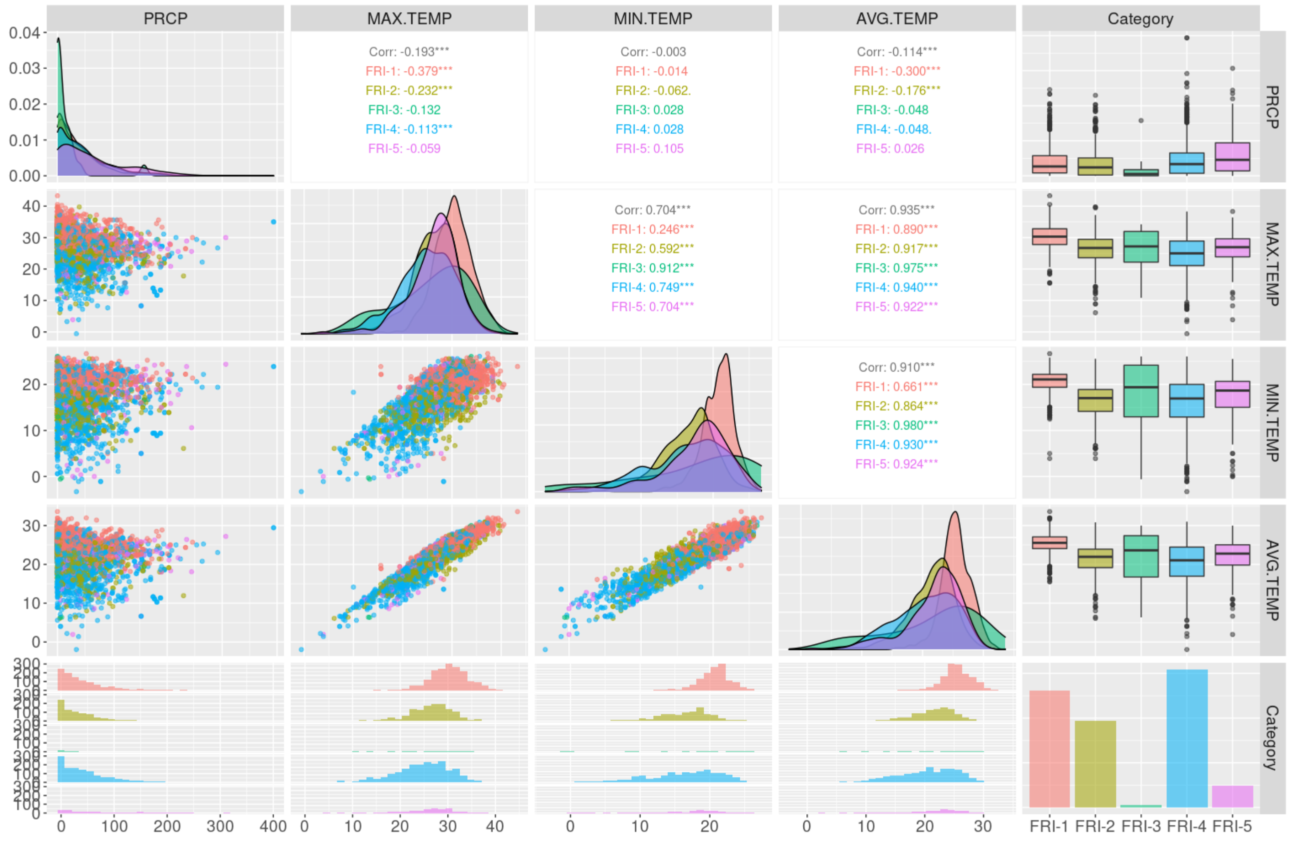

Figure 4. Exploratory and sensitivity data analysis of the climate information, and the FRI variables used in the prediction framework.

Table 2. Predictive model performance comparison for different class predictions in the different ML models.

	Training Precision					**Training Recall (Sensitivity)**					**Training F1-Score**					**Training**	
	FRI-1	**FRI-2**	**FRI-3**	**FRI-4**	**FRI-5**	**FRI-1**	**FRI-2**	**FRI-3**	**FRI-4**	**FRI-5**	**FRI-1**	**FRI-2**	**FRI-3**	**FRI-4**	**FRI-5**	**Accuracy**	**Misclass.**
Naïve Bayes	86.9%	38.7%	33.3%	40.5%	23.0%	52.1%	42.9%	75.0%	67.6%	52.5%	65.2%	40.7%	46.2%	50.7%	32.0%	53.8%	46.2%
RF 300	99.6%	98.3%	94.4%	99.1%	85.2%	98.2%	97.3%	100.0%	98.6%	99.1%	98.9%	97.8%	97.1%	98.8%	91.6%	97.8%	2.2%
RF 1000	99.6%	98.7%	94.4%	99.0%	84.4%	98.2%	96.9%	100.0%	98.7%	100.0%	98.9%	97.8%	97.1%	98.8%	91.6%	98.2%	1.8%
Bagged DT	98.4%	97.2%	94.1%	98.5%	93.3%	99.2%	98.2%	94.1%	97.8%	89.3%	98.8%	97.7%	94.1%	98.1%	91.2%	98.2%	1.8%
	Testing Precision					**Testing Recall (Sensitivity)**					**Testing F1-Score**					**Testing**	
	FRI-1	**FRI-2**	**FRI-3**	**FRI-4**	**FRI-5**	**FRI-1**	**FRI-2**	**FRI-3**	**FRI-4**	**FRI-5**	**FRI-1**	**FRI-2**	**FRI-3**	**FRI-4**	**FRI-5**	**Accuracy**	**Misclass.**
Naïve Bayes	89.0%	36.8%	50.0%	33.6%	8.5%	52.0%	42.2%	40.0%	58.5%	27.8%	65.7%	39.3%	44.4%	42.7%	13.0%	50.9%	49.1%
RF 300	72.7%	48.5%	66.7%	58.5%	8.5%	64.8%	50.2%	25.0%	57.8%	27.8%	68.5%	49.3%	36.4%	58.1%	13.0%	57.9%	42.1%
RF 1000	73.4%	50.4%	66.7%	55.6%	8.5%	64.1%	50.0%	28.6%	57.4%	26.3%	68.4%	50.2%	40.0%	56.5%	12.8%	57.8%	42.2%
Bagged DT	66.9%	48.1%	20.0%	58.4%	36.7%	70.1%	50.9%	33.3%	58.5%	18.0%	68.5%	49.4%	25.0%	58.5%	24.2%	57.3%	42.7%

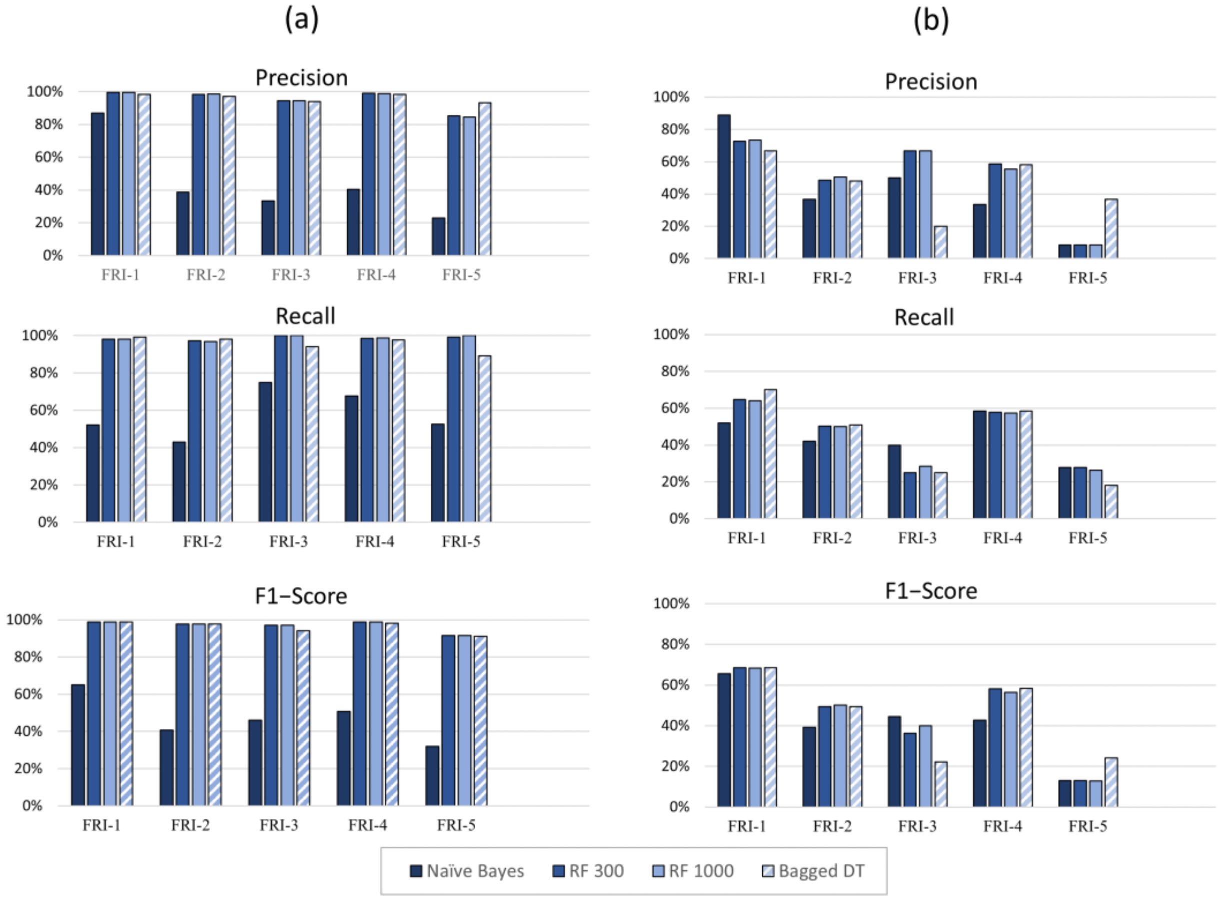

Figure 5. Prediction performance indices for the four utilized models where: (**a**) is the training subset performance, and (**b**) is the testing subset performance.

Further investigation of the RF and Bagged DT models shows that the variables used as predictors in the current study influence the behavior of the predictive analysis at each class. This influence indicates the need for more comprehensive and climatologically representative variables to be used as predictors. In data-driven studies, model performance depends heavily on the available dataset; as such, the authors were constrained by the available data to use in the validation of the developed methodology. A comprehensive dataset would include as much observations as possible over a wider time span, with numerous variables (e.g., atmospheric pressure, wind speed, wind direction, humidity, topology exposure, etc.). To assess the importance of the individual variables in the analysis, the mean decrease Gini (MDG) was employed in the RF ensemble models. Figure 6 shows the MDG and the mean decrease accuracy for the RF with 300 and 1000 tree models, the MDG indicates that the average temperature is the most important variable in both models, followed by the precipitation in the RF 1000 models, and the minimum temperature in the RF 300 model, albeit with a very small difference with the precipitation in the RF 300 model. This supports that the Average temperature (correlated with the minimum temperature) and the Precipitation are key variables when predicting the community-flood resilience in exposed communities.

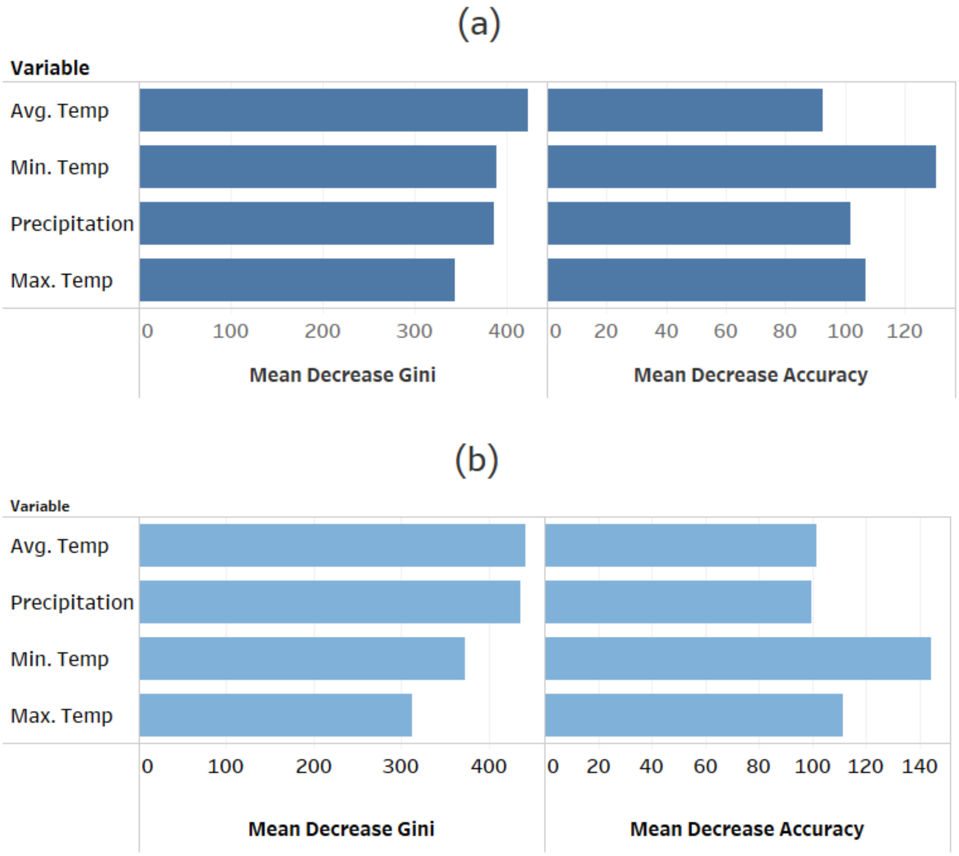

Figure 6. Mean decrease Gini and mean decrease accuracy in (**a**) Random Forest model with 300 trees and (**b**) Random Forest model with 1000 trees.

The results of the analysis displayed in the current study shows that the framework and methodology presented herein are applicable in flood resilience prediction studies. This framework informs decision-making process through developing an early-warning system that can be continuously updated by including new, and more accurate, climate data. The framework presented herein can also be coupled with global climate models to study the temporal changes in flood resilience and the climate impact on infrastructure resilience. This coupling would enable informed decisions and policies for a better utilization of resilience means (i.e., resourcefulness and redundancy) to enhance the community's climate resilience. It is worth noting that these predictions and projections will be subject to the uncertainty associated with the climate models; as such, a reliable ensemble from multiple models needs to be used in order to reduce the effect of this uncertainty and reduce the variability between these different models.

The framework presented herein can also be applicable in different data-driven studies, where the purpose is to investigate the spatio-temporal vulnerability of a system facing an external disruption (e.g., vulnerability-based evacuations).

4. Discussion and Conclusions

As the IPCC 2021 report stated, extreme rainfall events are expected to increase in frequency and intensity over the next decade, with an increase of over 2.0 m in the average sea level by the end of the current century. Numerous studies were developed to assess community resilience, mostly considering the feature of the hazard rather than the features of the exposed system at risk. The current work aims to: (1) identify specific variables to represent resilience means across a specific time-span to develop an comprehensive dataset for data-driven models, (2) develop resilience indices using unbiased data-driven methods under different weather conditions across a specific region, (3) develop a comparative spatial analysis to identify at-risk communities and assess their vulnerabilities to further enhance their resilience [59], (4) couple the indices with climate information to develop a well synchronized dataset to be used with future climate models for accurate resilience prediction, and finally (5) test the framework using the NWS disaster records to develop flood resilience indices. The output of said categorization is then coupled with the historic climate information from NOAA corresponding to the disaster records from 1996 to 2019. The resulting dataset is used to develop, train, and test the prediction ML model.

The demonstration application of the developed framework was developed using unsupervised ML techniques in Part (a) and supervised ML in Part (b). In Part (a), the model was applied to the NWS's historical disaster database, collected across the United States from 1996 to 2019. This dataset included variables with information regarding the damage, duration, indirect/direct injuries, and fatalities, and these variables were used to extract the resilience information correspondence to each recorded disaster (i.e., robustness and rapidity) so that the developed categorization would capture the resilience of the exposed community, resulting in five categories (i.e., indices). For the second part of the framework, the state of Texas was chosen as a test location, given the uniformity of the meteorological conditions over the state and the uniformity of the built environment (with few acceptable exceptions). A spatial analysis within the state of Texas was conducted using the developed indices in Part (a), highlighting the more vulnerable counties within the state. This spatial analysis concluded that the coastal areas around the Gulf of Mexico are subjected to flood events that result in a higher index than other counties, resulting in a larger impact on the robustness of said communities. This highlights the need for an accurate methodology to predict future impact on said communities to be able to develop proactive flood risk management strategies and enhance their overall resilience.

The second part of the application utilized numerous ensemble prediction techniques (i.e., Random Forest (RF) with 300 and 1000 trees, Bagged Decision Trees (DT), and Naïve Bayes (NB) classification). The output of this stage demonstrated the applicability of the developed framework, with comparable results across the different models. While the Bagged DT outperformed the RF models in categories where the data were scarce, they performed similarly in other categories. To objectively assess the performance of all the models, precision, recall, and F-1 Score were employed across different categories, in training and testing datasets, resulting in a comprehensive conclusion that the prediction framework is employable in resilience-guided studies. However, to objectively develop a data-driven method, a comprehensive enough dataset with variable across different regions and across the years, with enough variables should be employed. In the current framework demonstration study, the authors were limited by the available data; however, the prediction performance of the framework can be improved given more climate information (i.e., wind speed, humidity, and air pressure, etc.). These variables would increase the correlation with the developed resilience indices, resulting in a more robust dataset for the training and testing of the prediction model. A limitation of the work presented herein is that future climate projections were not considered in the demonstration application. Provided the availability of said projections, the trajectory of the resilience of the exposed community can be determined, and the vulnerability and resilience can be evaluated ahead of projected extreme events, giving policy makers the opportunity to develop mitigation and resilience enhancement plans to avoid future disasters. The framework can be adapted to account for the uncertainty induced by the climate projections' nature and the probabilistic nature of the hazard as well as the response of the community and the resulting resilience. This can be carried out through accumulating probabilities resulting from Monte Carlo simulations to determine the response to the hazard itself and include it in the prediction framework.

To that end, further research can be implemented to advance this framework through (1) incorporating more variables within the utilized datasets, (2) combining the results of the different ensemble ML models used in this study to further enhance the prediction performance, and (3) applying the framework to future climate projections to predict the expected change in the resilience of the exposed communities.

Author Contributions: M.N.A.-M.: Conceptualization, Data curation, Investigation, Methodology, Software, Formal analysis, Validation, Visualization, Writing—original draft, Writing—review and editing, and Funding acquisition. W.E.-D.: Writing—review and editing, Visualization, Supervision, Methodology, Funding acquisition, Formal analysis, and Conceptualization. P.C.: Conceptualization, Funding acquisition, Supervision, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council: Grant No. [CREATE/482707-2016], and the Vanier Canada Graduate Scholarship (Vanier—CGS) awarded to the Corresponding Author.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this article are publicly available. The meteorological disaster database used to generate the resilience-based categories is provided by the NWS a sub-agency under the National Oceanic and Atmospheric Administration (NOAA), available at (https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/) (accessed on 1 May 2021). The historical climate data used as dependent variables in the ML model is provided by the Global Historical Climatology Network, a sub-agency under NOAA, and is available at (https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND) (accessed on 6 June 2021).

Acknowledgments: The work presented herein is supported by the Vanier Canada Graduate Scholarship (Vanier-CGS) awarded to the corresponding author and the Natural Science and Engineering Research Council (NSERC) through the CaNRisk—Collaborative Research and Training Experience (CREATE) program. Additional support through the INViSiONLab and the INTERFACE Institute at McMaster University is also acknowledged.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Dawod, G.M.; Mirza, M.N.; Al-Ghamdi, K.A.; Elzahrany, R.A. Projected Impacts of Land Use and Road Network Changes on Increasing Flood Hazards Using a 4D GIS: A Case Study in Makkah Metropolitan Area, Saudi Arabia. *Arab. J. Geosci.* **2014**, *7*, 1139–1156. [CrossRef]
2. Lian, J.; Xu, H.; Xu, K.; Ma, C. Optimal Management of the Flooding Risk Caused by the Joint Occurrence of Extreme Rainfall and High Tide Level in a Coastal City. *Nat. Hazards* **2017**, *89*, 183–200. [CrossRef]
3. Wilby, R.L.; Beven, K.J.; Reynard, N.S. Climate Change and Fluvial Flood Risk in the UK: More of the Same? *Hydrol. Process.* **2007**, *2309*, 2300–2309. [CrossRef]
4. Stocker, T.F.; Dahe, Q.; Plattner, G.-K.; Tignor, M.M.B.; Allen, S.K.; Boschung, J.; Nauels, A.; Xia, Y.; Bex, V.; Vincent, P.M. *Climate Change 2013: The Physical Science Basis*; IPCC: New York, NY, USA, 2013; ISBN 9789291691388.
5. Linkov, I.; Bridges, T.; Creutzig, F.; Decker, J.; Fox-Lent, C.; Kröger, W.; Lambert, J.H.; Levermann, A.; Montreuil, B.; Nathwani, J.; et al. Changing the Resilience Paradigm. *Nat. Clim. Chang.* **2014**, *4*, 407–409. [CrossRef]
6. Bertilsson, L.; Wiklund, K.; de Moura Tebaldi, I.; Rezende, O.M.; Veról, A.P.; Miguez, M.G. Urban Flood Resilience—A Multi-Criteria Index to Integrate Flood Resilience into Urban Planning. *J. Hydrol.* **2019**, *573*, 970–982. [CrossRef]
7. da Silva, J.; Kernaghan, S.; Luque, A. A Systems Approach to Meeting the Challenges of Urban Climate Change. *Int. J. Urban Sustain. Dev.* **2012**, *4*, 125–145. [CrossRef]
8. NOAA National Climate Report—Annual 2018 | State of the Climate | National Centers for Environmental Information (NCEI). Available online: https://www.ncdc.noaa.gov/sotc/national/201813#over (accessed on 5 May 2020).
9. de Moel, H.; Aerts, J.C.J.H. Effect of Uncertainty in Land Use, Damage Models and Inundation Depth on Flood Damage Estimates. *Nat. Hazards* **2011**, *58*, 407–425. [CrossRef]
10. World Economic Forum. *The Global Risks Report 2019*, 14th ed.; World Economic Forum: Cologny, Switzerland, 2019.
11. Kron, W. Flood Risk = Hazard • Values • Vulnerability. *Water Int.* **2005**, *30*, 58–68. [CrossRef]
12. Salem, S.; Siam, A.; El-Dakhakhni, W.; Tait, M. Probabilistic Resilience-Guided Infrastructure Risk Management. *J. Manag. Eng.* **2020**, *36*, 04020073. [CrossRef]
13. Nofal, O.M.; van de Lindt, J.W. Understanding Flood Risk in the Context of Community Resilience Modeling for the Built Environment: Research Needs and Trends. *Sustain. Resilient Infrastruct.* **2020**, *7*, 1–17. [CrossRef]
14. *Community Resilience Planning Guide for Buildings and Infrastructure Systems: A Playbook*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020.

15. Netherton, M.D.; Stewart, M.G. Risk-Based Blast-Load Modelling: Techniques, Models and Benefits. *Int. J. Prot. Struct.* **2016**, *7*, 430–451. [CrossRef]
16. Cimellaro, G.P.; Fumo, C.; Reinhorn, A.M.; Bruneau, M. *Quantification of Disaster Resilience of Health Care Facilities*; Earthquake Engineering to Extreme Events University at Buffalo: Buffalo, NY, USA, 2009.
17. Bruneau, M.; Chang, S.E.; Eguchi, R.T.; Lee, G.C.; O'Rourke, T.D.; Reinhorn, A.M.; Shinozuka, M.; Tierney, K.; Wallace, W.A.; Von Winterfeldt, D. A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities. *Earthq. Spectra* **2003**, *19*, 733–752. [CrossRef]
18. Murdock, H.J. *Resilience of Critical Infrastructure to Flooding: Quantifying the Resilience of Critical Infrastructure to Flooding in Toronto, Canada*; UNESCO-IHE: Delft, The Netherlands, 2017.
19. Minsker, B.; Baldwin, L.; Crittenden, J.; Kabbes, K.; Karamouz, M.; Lansey, K.; Malinowski, P.; Nzewi, E.; Pandit, A.; Parker, J.; et al. Progress and Recommendations for Advancing Performance-Based Sustainable and Resilient Infrastructure Design. *J. Water Resour. Plan. Manag.* **2015**, *141*, A4015006. [CrossRef]
20. Australian Institute for Disaster Resilience. *Flood Emergency Response: Classification of the Floodplain*; Guideline 7-2; Australian Institute for Disaster Resilience: Melborne, Australia, 2017.
21. Federal Emergency Management Agency (FEMA). *Definitions of FEMA Flood Zone Designations*; FEMA: Washington, DC, USA, 2012; pp. 1–2.
22. Ragini, J.R.; Anand, P.M.R.; Bhaskar, V. Big Data Analytics for Disaster Response and Recovery through Sentiment Analysis. *Int. J. Inf. Manage.* **2018**, *42*, 13–24. [CrossRef]
23. Turkington, T.; Breinl, K.; Ettema, J.; Alkema, D.; Jetten, V. A New Flood Type Classification Method for Use in Climate Change Impact Studies. *Weather Clim. Extrem.* **2016**, *14*, 1–16. [CrossRef]
24. Ganguli, P.; Paprotny, D.; Hasan, M.; Güntner, A.; Merz, B. Projected Changes in Compound Flood Hazard From Riverine and Coastal Floods in Northwestern Europe. *Earth's Futur.* **2020**, *8*, e2020EF001752. [CrossRef]
25. Ganguly, K.K.; Nahar, N.; Hossain, B.M. A Machine Learning-Based Prediction and Analysis of Flood Affected Households: A Case Study of Floods in Bangladesh. *Int. J. Disaster Risk Reduct.* **2019**, *34*, 283–294. [CrossRef]
26. Hemmati, M.; Ellingwood, B.R.; Mahmoud, H.N. The Role of Urban Growth in Resilience of Communities Under Flood Risk. *Earth's Futur.* **2020**, *8*, e2019EF001382. [CrossRef]
27. Murnane, R.J.; Daniell, J.E.; Schäfer, A.M.; Ward, P.J.; Winsemius, H.C.; Simpson, A.; Tijssen, A.; Toro, J. Future Scenarios for Earthquake and Flood Risk in Eastern Europe and Central Asia. *Earth's Futur.* **2017**, *5*, 693–714. [CrossRef]
28. Rözer, V.; Peche, A.; Berkhahn, S.; Feng, Y.; Fuchs, L.; Graf, T.; Haberlandt, U.; Kreibich, H.; Sämann, R.; Sester, M.; et al. Impact-Based Forecasting for Pluvial Floods. *Earth's Futur.* **2021**, *9*, 2020EF001851. [CrossRef]
29. Swain, D.L.; Wing, O.E.J.; Bates, P.D.; Done, J.M.; Johnson, K.A.; Cameron, D.R. Increased Flood Exposure Due to Climate Change and Population Growth in the United States. *Earth's Futur.* **2020**, *8*, e2020EF001778. [CrossRef]
30. Downton, M.W.; Pielke, R.A. How Accurate Are Disaster Loss Data? The Case of U.S. Flood Damage. *Nat. Hazards* **2005**, *35*, 211–228. [CrossRef]
31. Patil, B.M.; Joshi, R.C.; Toshniwal, D. Missing Value Imputation Based on K-Mean Clustering with Weighted Distance. In *Contemporary Computing. IC3 2010. Communications in Computer and Information Science*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 600–609.
32. Yagci, K.; Dolinskaya, I.S.; Smilowitz, K.; Bank, R. Incomplete Information Imputation in Limited Data Environments with Application to Disaster Response. *Eur. J. Oper. Res.* **2018**, *269*, 466–485. [CrossRef]
33. Haggag, M.; Yorsi, A.; El-dakhakhni, W.; Hassini, E. Infrastructure Performance Prediction under Climate-Induced Disasters Using Data Analytics. *Int. J. Disaster Risk Reduct.* **2021**, *56*, 102121. [CrossRef]
34. Bose, I.; Mahapatra, R.K. Business Data Mining—A Machine Learning Perspective. *Inf. Manag.* **2001**, *39*, 211–225. [CrossRef]
35. Goos, G.; Hartmanis, J.; Van, J.; Board, L.E.; Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J.M.; Mattern, F.; Zurich, E.; et al. *Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006.
36. King, R.D.; Muggletont, S.; Lewiso, R.A.; Sternberg, M.J.E. Drug Design by Machine Learning: The Use of Inductive Logic Programming to Model the Structure-Activity Relationships of Trimethoprim Analogues Binding to Dihydrofolate Reductase (Arfcl Integence/Ee Acv/Prote l/Active Sites). *Proc. Natd. Acad. Sci. USA* **1992**, *89*, 11322–11326. [CrossRef]
37. Mckinney, B.A.; Reif, D.M.; Ritchie, M.D.; Moore, J.H. Biomedical Genomics And Proteomics Machine Learning for Detecting Gene-Gene Interactions A Review. *Appl. Bioinform.* **2006**, *5*, 77–88. [CrossRef]
38. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009; Volume 27, ISBN 9780387848570.
39. Gentleman, R.; Hornik, K.; Parmigiani, G. *Bicondutor Case Studies*; Springer: Berlin/Heidelberg, Germany, 2008.
40. Mojaddadi, H.; Pradhan, B.; Nampak, H.; Ahmad, N.; Ghazali, A.H. Bin Ensemble Machine-Learning-Based Geospatial Approach for Flood Risk Assessment Using Multi-Sensor Remote-Sensing Data and GIS. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1080–1102. [CrossRef]
41. Mosavi, A.; Ozturk, P.; Chau, K.W. Flood Prediction Using Machine Learning Models: Literature Review. *Water* **2018**, *10*, 1536. [CrossRef]
42. Shafizadeh-Moghadam, H.; Valavi, R.; Shahabi, H.; Chapi, K.; Shirzadi, A. Novel Forecasting Approaches Using Combination of Machine Learning and Statistical Models for Flood Susceptibility Mapping. *J. Environ. Manage.* **2018**, *217*, 1–11. [CrossRef]

43. Rodrigues, M.; De la Riva, J. An Insight into Machine-Learning Algorithms to Model Human-Caused Wildfire Occurrence. *Environ. Model. Softw.* **2014**, *57*, 192–201. [CrossRef]
44. Haggag, M.; Siam, A.S.; El-Dakhakhni, W.; Coulibaly, P.; Hassini, E. A Deep Learning Model for Predicting Climate-Induced Disasters. *Nat. Hazards* **2021**, *196*, 227–243. [CrossRef]
45. Hanewinkel, M.; Zhou, W.; Schill, C. A Neural Network Approach to Identify Forest Stands Susceptible to Wind Damage. *For. Ecol. Manage.* **2004**, *196*, 227–243. [CrossRef]
46. Abdel-Mooty, M.N.; El-Dakhakhni, W.; Coulibaly, P. Community Resilience Classification Under Climate Change Challenges. In Proceedings of the Canadian Society of Civil Engineering Annual Conference, Montreal, QC, Canada, May 2021; Springer: Singapore, 2021; pp. 227–237. [CrossRef]
47. Otterbach, J.S.; Manenti, R.; Alidoust, N.; Bestwick, A.; Block, M.; Bloom, B.; Caldwell, S.; Didier, N.; Fried, E.S.; Hong, S.; et al. Unsupervised Machine Learning on a Hybrid Quantum Computer. *arXiv* **2017**, arXiv:1712.05771.
48. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data Clustering: A Review. *ACM Comput. Surv.* **2000**, *31*, 264–323.
49. Seyed Shirkhorshidi, A.; Aghabozorgi, S.; Wah, Y. A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PloS ONE* **2015**, *10*, e0144059. [CrossRef]
50. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 18–21 June 1965; University of California: Berkeley, CA, USA, 1967; pp. 281–297.
51. Alsabti, K.; Ranka, S.; Singh, V. An efficient k-means clustering algorithm. In *Electrical Engineering and Computer Science*; Syracuse University: Syracuse, NY, USA, 2000; Available online: https://surface.syr.edu/eecs/43 (accessed on 10 June 2021).
52. Hartigan, J.A.; Wong, M.A. A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108.
53. Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained K-Means Clustering with Background Knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, 28 June–1 July 2001; pp. 577–584.
54. Mitra, P.; Ray, R.; Chatterjee, R.; Basu, R.; Saha, P.; Raha, S.; Barman, R.; Patra, S.; Biswas, S.S.; Saha, S. Flood Forecasting Using Internet of Things and Artificial Neural Networks. In Proceedings of the 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE IEMCON 2016, Vancouver, BC, Canada, 13–15 October 2016; pp. 1–5.
55. Park, D.C. Centroid Neural Network for Unsupervised Competitive Learning. *IEEE Trans. Neural Netw.* **2000**, *11*, 520–528. [CrossRef]
56. Khajwal, A.B.; Noshadravan, A. Probabilistic Hurricane Wind-Induced Loss Model for Risk Assessment on a Regional Scale. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng.* **2020**, *6*, 1–9. [CrossRef]
57. Kwayu, K.M.; Kwigizile, V.; Zhang, J.; Oh, J.S. Semantic N-Gram Feature Analysis and Machine Learning-Based Classification of Drivers' Hazardous Actions at Signal-Controlled Intersections. *J. Comput. Civ. Eng.* **2020**, *34*. [CrossRef]
58. Gnanaprakkasam, S.; Ganapathy, G.P. Evaluation of Regional Flood Quantiles at Ungauged Sites by Employing Nonlinearity-Based Clustering Approaches. *Environ. Sci. Pollut. Res.* **2019**, *26*, 22856–22877. [CrossRef]
59. Abdel-Mooty, M.N.; Yosri, A.; El-Dakhakhni, W.; Coulibaly, P. Community Flood Resilience Categorization Framework. *Int. J. Disaster Risk Reduct.* **2021**, *61*, 102349. [CrossRef]
60. Khalaf, M.; Hussain, A.J.; Al-Jumeily, D.; Baker, T.; Keight, R.; Lisboa, P.; Fergus, P.; Al Kafri, A.S. A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction. In Proceedings of the 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8. [CrossRef]
61. Zumel, N.; Mount, J. *Practical Data Science with R*, 2nd ed.; Manning Publication: Shelter Island, NY, USA, 2020; ISBN 9781617295874.
62. Singh, H. Understanding Gradient Boosting Machines | by Harshdeep Singh | Towards Data Science. Available online: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab (accessed on 12 May 2021).
63. Nagpal, A. Decision Tree Ensembles- Bagging and Boosting | by Anuja Nagpal | Towards Data Science. Available online: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9 (accessed on 12 May 2021).
64. Boehmke, B.; Greenwell, B.M. *Hands-On Machine Learning with R*, 1st ed.; Taylor & Francis: London, UK, 2019; ISBN 9781138495685.
65. Gong, J.; Caldas, C.H.; Gordon, C. Learning and Classifying Actions of Construction Workers and Equipment Using Bag-of-Video-Feature-Words and Bayesian Network Models. *Adv. Eng. Inform.* **2011**, *25*, 771–782. [CrossRef]
66. Wu, T.F.; Lin, C.J.; Weng, R.C. Probability Estimates for Multi-Class Classification by Pairwise Coupling. *J. Mach. Learn. Res.* **2004**, *5*, 975–1005.
67. Gondia, A.; Siam, A.; El-Dakhakhni, W.; Nassar, A.H. Machine Learning Algorithms for Construction Projects Delay Risk Prediction. *J. Constr. Eng. Manag.* **2020**, *146*, 1–16. [CrossRef]
68. Mccallum, A.; Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. In Proceedings of the Annual AAAI National Conference, Palo Alto, CA, USA, 26–30 July 1998.
69. Zhang, H. The Optimality of Naive Bayes. In Proceedings of the annual AAAI National Conference, Palo Alto, CA, USA, 25–29 July 2004.
70. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: London, UK, 1984; ISBN 9781315139470.
71. Efron, B.; Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat. Sci.* **1986**, *1*, 54–75. [CrossRef]

72. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
73. Brownlee, J. Bagging and Random Forest Ensemble Algorithms for Machine Learning. Available online: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/ (accessed on 12 May 2021).
74. Feofilovs, M.; Romagnoli, F. Resilience of Critical Infrastructures: Probabilistic Case Study of a District Heating Pipeline Network in Municipality of Latvia. *Energy Procedia* **2017**, *128*, 17–23. [CrossRef]
75. Liaw, A.; Wiener, M. Classification and Regression by Randomforest. *R News* **2002**, *2*, 18–22.
76. Fielding, A.H. Introduction to Classification. In *Cluster and Classification Techniques for the Biosciences*; Cambridge University Press: Cambridge, UK, 2006; pp. 78–96.
77. Brownlee, J. How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification. Available online: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/ (accessed on 16 June 2021).
78. Murphy, J.D. NWSI 10-1605, Storm Data Preparation. 2018. Available online: https://www.nws.noaa.gov/directives/sym/pd01016005curr.pdf (accessed on 10 June 2021).
79. Downton, M.W.; Miller, J.Z.B.; Pielke, R.A. Reanalysis of U.S. National Weather Service Flood Loss Database. *Nat. Hazards Rev.* **2005**, *6*, 13–22. [CrossRef]
80. Trenberth, K.E.; Cheng, L.; Jacobs, P.; Zhang, Y.; Fasullo, J. Hurricane Harvey Links to Ocean Heat Content and Climate Change Adaptation. *Earth's Futur.* **2018**, *6*, 730–744. [CrossRef]
81. *Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL]*; U.S. Bureau of Labor Statistics: Washington DC, USA, 2021. Available online: https://fred.stlouisfed.org/series/CPIAUCSL (accessed on 5 May 2020).
82. Sweet, W.V.; Dusek, G.; Carbin, G.; Marra, J.; Marcy, D.; Simon, S. 2019 State of U.S. High Tide Flooding and a 2020 Outlook. 2020. Available online: https://tidesandcurrents.noaa.gov/publications/Techrpt_092_2019_State_of_US_High_Tide_Flooding_with_a_2020_Outlook_30June2020.pdf (accessed on 10 June 2021).
83. NOAA, U.S. High-Tide Flooding Continues to Increase | National Oceanic and Atmospheric Administration. Available online: https://www.noaa.gov/media-release/us-high-tide-flooding-continues-to-increase (accessed on 10 June 2021).
84. NOAA Office for Coastal Management Texas. Available online: https://coast.noaa.gov/states/texas.html (accessed on 10 June 2021).
85. Perica, S.; Pavlovic, S.; Laurent, M.S.; Trypaluk, C.; Unruh, D.; Wilhite, O. *Precipitation-Frequency Atlas of the United States Volume 11 Version 2.0: Texas*; NOA: Washington DC, USA, 2018; Volume 11.
86. Menne, M.J.; Durre, I.; Vose, R.S.; Gleason, B.E.; Houston, T.G. An Overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Ocean. Technol.* **2012**, *29*, 897–910. [CrossRef]
87. Menne, M.J.; Durre, I.; Vose, R.S.; Gleason, B.E.; Houston, T.G. Global Historical Climatology Network-Daily (GHCN-Daily), Version 3. Available online: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00861 (accessed on 10 June 2021).
88. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining Practical Machine Learning Tools and Techniques*, 4th ed.; Elsevier: Amsterdam, The Netherlands, 2017; ISBN 9780128042915.
89. Ashari, A. Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *Int. J. Adv. Comput. Sci. Appl.* **2013**, *4*, 33–39. [CrossRef]