A mixed approach for urban flood prediction using Machine Learning and GIS

Marcel Motta, Miguel de Castro Neto, Pedro Sarmento

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal.

This is the accepted author manuscript of the following article published by Elsevier:

Motta, M., De Castro Neto, M., & Sarmento, P. (2021). A mixed approach for urban flood prediction using Machine Learning and GIS. International Journal of Disaster Risk Reduction, 102154. [Advanced online publication on 26 February 2021]. https://doi.org/10.1016/j.ijdrr.2021.102154

A mixed approach for urban flood prediction using Machine Learning and GIS

Marcel Motta[a,1], Miguel de Castro Neto[a], Pedro Sarmento[a]

[a] *NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal.*

ABSTRACT

Extreme weather conditions, as one of many effects of climate change, is expected to increase the magnitude and frequency of environmental disasters. In parallel, urban centres are also expected to grow significantly in the next years, making necessary to implement the adequate mechanisms to tackle such threats, more specifically flooding.

This project aims to develop a flood prediction system using a combination of Machine Learning classifiers along with GIS techniques to be used as an effective tool for urban management and resilience planning. This approach can establish sensible factors and risk indices for the occurrence of floods at the city level, which could be instrumental for outlining a long-term strategy for Smart Cities.

The most performant Machine Learning model was a Random Forest, with a Matthew's Correlation Coefficient of 0.77 and an Accuracy of 0.96.

To support and extend the capabilities of the Machine Learning model, a GIS model was developed to find areas with higher likelihood of being flooded under critical weather conditions. Therefore, hot spots were defined for the entire city given the observed flood history. The scores obtained from the Random Forest model and the Hot Spot analysis were then combined to create a flood risk index.

[1] Corresponding author, email: mmotta@novaims.unl.pt

KEYWORDS

Flood Prediction, Resilience Planning, Smart Cities, Machine Learning, GIS

1. INTRODUCTION

Over the course of time, environmental disasters have taken the centre stage in the political arena due to the vast impact on the economy and society as a whole, fuelled by the accelerated urban growth and climate change. According to the United Nation's report Habitat III [1], more than 50% of the global population is currently concentrated in urban areas and this number is expected to rise to over 70% by 2050. This outlook raises concerns regarding developmental challenges, such as adjusting urban infrastructure to improve emergency response and risk management, calling for the formulation of a broader urban management strategy for resilience against natural disasters. Furthermore, such strategies and policies contribute to economic growth, social development and environmental sustainability and resilience [1].

Within the context of natural disasters, floods are characterized as one of the most significant and common hazards, with a constantly rising frequency since 1960, being identified as the costliest natural disaster globally [2]. Between 2013 and 2018, 2,154 floods were reported in the city of Lisbon, leading to damages to property and general disruption in the local communities. Beyond the immediate impacts of such events, these disasters often exacerbate existing socioeconomic and environmental weaknesses in the urban system [1].

Dankers & Feyen [3] state that extreme precipitation events are expected to significantly increase in intensity and frequency due to the effects of climate change. This environmental shift raises concerns, as an increase in the frequency and magnitude of heavy precipitation events implies an increased risk of flooding. To face these new threats, risk management had to gradually evolve in order to adapt to the uncertainties brought in by climate change in urban

areas. In this context, it is necessary to prepare territories and population for the increasing hazards by introducing the concept of resilience [4]. As pointed out in the literature, resilience can be defined as "the capacity of a territory and its population to plan for, adapt, absorb, recover from, learn and evolve" [4] [5]. Additionally, this approach should not be limited to addressing systemic environmental disasters, but also framing resilience as a mean to realize opportunities for transformational development, economic growth and sustainability, being a driver to development itself [1]. However, the understanding and application of resilience still poses as a challenge to policy makers as it in fact remains poorly integrated into risk management strategies.

To tackle this challenge, Serre & Heinzlef [2] suggest the development of innovative strategies and decision support systems for new resilient urban environments. In that sense, several approaches have been developed to operationalize resilience, but they are mostly focused on introducing qualitative models model for mapping resilience of urban infrastructures [5] and proposing organizational and holistic tools to limit the impact of disruptive events [6] [7]. In contrast, flood modelling typically relies on quantitative models to enable reactive measures (i.e. emergency response and recovery) and/or proactive measures (i.e. risk analysis and mitigation) [8] [9] [10]. As described by Tingsanchali [11], these measures could be put into place through a strategic framework on integrated flood disaster management, consisting of four cyclic steps: 1) preparedness before flood impact, such as flood forecasting and warning; 2) readiness upon flood arrival; 3) emergency responses during flood impact and; 4) recovery and rehabilitation after flood impact.

As reported by Mosavi et al. [12], hydrological models have traditionally been developed for flood prediction, but new trends in Machine Learning (ML) techniques in the field of flood detection have shown promising results [9] [10]. Alternatively, Geographic Information Systems (GIS) have become recognized and utilized as major tools for monitoring and

analysing human crises and natural hazards in the past decades [13]. In that sense, several works have been developed about the application of GIS for flood analysis, by using geomorphological characteristics to estimate runoff/inundation models and vulnerability indices [14] [15] [16] [17].

With that in mind, the work herein developed aims to deliver a spatial prediction tool for flood events in Lisbon, capable of fulfilling this new paradigm for flood management and increasing resilience in two different dimensions: by supporting decision makers on implementing proactive measures in high-risk areas to mitigate damages caused by floods, and by allowing the optimization and the degree of preparedness of emergency and recovery services. By doing so, city managers should be able to understand which areas of the territory and communities are more vulnerable to floods, allowing them to outline a long-term risk management strategy while also dealing with short-term operational challenges in the response services. In the context of the local urban landscape, rainfall and runoff are assumed to be the root causes for flooding scenarios in Lisbon. Other typical causes for floods, such as infrastructure failures, fluvial behaviour, and indirect geological phenomena, are not representative at the city level and will not be considered in the scope of this study.

In this study, an advanced data-driven approach is proposed, which uses a combination of a ML classifier and GIS statistics. This two-step approach will be able to extend the predictions created with ML by providing geospatial analytics for the entire city, improving the spatial representation in the results provided by the predictive model.

2. MATERIALS & METHODS

In this section, methods and basic concepts used during the development of this paper are summarized and presented.

The project methodology was structured in accordance with the Cross Industry Standard Process for Data Mining (CRISP-DM). This framework provides general guidelines for iterative development and has been widely used in Data Science projects for several years [18], [19], [20]. As introduced by Wirth & Hipp [21], this model breaks data mining practices into six major phases:

1. Business Understanding: setting the objectives and requirements of the project from a business perspective. That is, a theoretical understanding the impact of floods, stakeholders and other relevant components, to lay the foundation for the work developed in the next steps. For this purpose, this phase will be addressed in chapter 1;
2. Data Understanding: getting familiar with the initial data set, discovering first insights and identifying data quality problems, using analytical techniques and data visualization;
3. Data Preparation: preprocessing and cleaning processes to construct the final data set for the modelling phase, such as feature engineering, coherence checking, imputation and outlier filtering;
4. Modelling: selecting modelling techniques and parameters. This way, informative features are selected so models can be effectively trained, parameters are defined and optimized, and performance measurements are benchmarked;
5. Evaluation: reviewing the development process, discussing the results obtained and verifying if the proposed solution is adequate to answer the core business problems, given a stated objective.
6. Deployment: organizing and presenting the generated knowledge to the customer/business. Ultimately, this paper attempts to offer a conceptual prototype that can be used by local authorities and policy makers to support the urban management and resilient planning for natural disasters.

The tools used for developing the ML models were implemented with Python (v. 3.7.4) using widely available Data Science modules, namely NumPy (v. 1.18.1), scipy (v. 1.4.1), pandas (v. 1.0.1), scikit-learn (v. 0.22.1) and imbalanced-learn (v. 0.6.2). As for the GIS layer, the tools used were implemented with the Spatial Analysis toolbox, available within ArcGIS Pro (v. 2.5). In Figure 1, it is presented the data pipeline for the proposed flood prediction system.

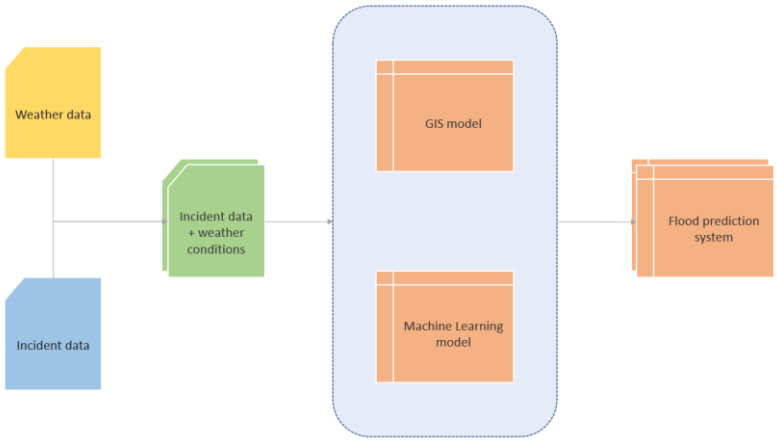

Figure 1. *Data pipeline for the proposed flood prediction system.*

The data used in this project was made available by the Lisbon city council, being comprised of local weather measurements and fire department emergency records. Using these two sources, ML models were trained to classify flood occurrences in the city. ML models utilize a supervised learning approach where, given a multidimensional projection of the explanatory/independent variables, a function or boundary is estimated to distinguish between classes defined by the dependent variable. In the context of this project, ML models utilized a combination of conditions (i.e. weather, time, season) to determine if a flood will take place, assuming a binary classification problem, where:

- 0, if no flood is observed/predicted, or;
- 1, if a flood is observed/predicted.

Based on previous work found in the literature, ML models can numerically formulate the flood nonlinearity, solely based on historical data without requiring knowledge about the underlying physical processes. Compared to traditional statistical models, ML models have greater accuracy and are also quicker to develop while requiring minimal inputs [12].

Using this approach, a set of binary classification algorithms were considered as, depending on their statistical assumptions (e.g. linear separability, normality, homoscedasticity), certain algorithms are more adequate than others, requiring some preliminary experimentation and model assessment. Therefore, to cover a broad set of heuristics applied for ML, a set of six classification models was deployed and evaluated using their corresponding scikit-learn implementations. These ML models were based on widely used algorithms suitable for flood prediction, as reported and reviewed by Mosavi et al. [12].

- Logistic Regression (LR): a linear, maximum-entropy classifier which computes the probability for the class variable by estimating a linear function;
- Support Vector Machine (SVM): a gaussian, kernel-based classifier which returns the predicted class by estimating a separating hyperplane between the two classes;
- Gaussian Naïve-Bayes (NB): a non-linear, Bayesian classifier which computes the conditional posterior probabilities of the class variable, assuming a gaussian distribution of the predictors;
- Random Forest (RF): a non-linear, ensemble classifier which computes the probability for the class variable using rule inference, obtained through an ensemble of decision trees;
- K-Nearest Neighbours (KNN): a non-linear, instance-based classifier which computes the probability for the class variable based on the k-nearest observed instances;

- Multi-Layer Perceptron (MLP): a non-linear, feedforward artificial neural network which computes the probability for the class variable by using layered inputs and backpropagation.

When evaluating these models, different performance metrics were used to better represent error and bias and to maximize robustness. As stated in previous works, the classification of imbalanced data sets can exert a major impact on the value and meaning of accuracy and on certain other well-known performance metrics [22]. For this purpose, Accuracy, Area Under Curve (AUC), Recall, F1 and Matthews Correlation Coefficient (MCC) were selected as metrics to provide a complete picture on model performance during the preliminary assessment. With exception of AUC, these metrics rely on factoring the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), as detailed in Table 1.

Metric	**Details**
Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$
Recall	$\frac{TP}{TP + FN}$
F1	$\frac{2 * TP}{2 * TP + FP + FN}$
MCC	$\frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

Table 1. *Metrics used for evaluating model performance, defined in [22].*

The ML pipeline proposed for this project presented an initial challenge as two critical spatial components could not be covered: the limited number of weather stations and the lack of representation of morphological factors that might contribute to an increased vulnerability. As mentioned by Chen et al. [15], the lack of high-resolution topographic and hydrologic data compromises the development and implementation of flooding models in urban areas. Due to this limitation, the resulting predictions from these models could be less reliable and would not be able to accurately determine the location where a flood will take place. It had been shown

in previous studies using advanced ML architectures that the catchment area and rainfall intensity are the two best predictors to develop a flood prediction model, while model's robustness drops when using a limited number of stations for fitting a model [9].

Therefore, to deal with the lack of spatial representation, additional tools were required to support the predictions obtained in the previous step. To fulfil this purpose a GIS model for the entire city was proposed to measure spatial relationships, considering that intrinsic geographic factors could increase the likelihood of flood in certain areas, as opposed to others.

Since each flood occurrence is associated to a specific spatial unit, the GIS model was used to identify spatial clusters with statistical significance, based on the Gi* statistic formulated by Getis & Ord [23]. To calculate the Gi* statistics, the Incremental Spatial Autocorrelation method defined the optimal distance threshold for maximizing the statistics score, creating clusters and assigning them as hot/cold spots [24]. The resulting clusters indicate which areas along the city are more likely to flood (hot spots) and which areas are less likely to flood (cold spots). For reference, Hot Spot analysis have been used in flood modelling as a tool for identifying spatial heterogeneity and to infer local vulnerability [25] [26].

While a Hot Spot analysis can highlight useful spatial patterns and ML models can be useful to find hidden patterns in data to predict future events, they are not independently sufficient to build a flood prediction system. This paper proposes a combination of both techniques by computing a risk index based on the scores provided by the GIS and ML models. To assess its performance and application, the combined model will be used to detect the floods occurred in the context of storms Elsa and Fabien back in December 2019 having its Recall (i.e. sensitivity) benchmarked.

The project described in this paper was developed in a partnership between the NOVA Cidade Urban Analytics Lab and the Lisbon city council. Through this partnership, weather data and

emergency records were collected from official sources and provided for the development of the project. The resulting solution was designed to be implemented and used by the Lisbon city council in the context of the *Plataforma de Gestão Inteligente de Lisboa* (PGIL), a smart cities platform for managing the city of Lisbon.

3. THEORY & CALCULATION

In this section, the practical development of the proposed flood prediction system will be presented into four subsections: Data understanding, Data preparation, Modelling, Evaluation, as described in chapter 2.

3.1. Data understanding

In order to fully understand the data, it is necessary to perform an initial exploratory analysis to identify the flood events, patterns, potential opportunities and limitations, so the data set can be handled and processed properly throughout the pipeline.

The required data was made available through by internal reports from the Lisbon city council, being originally collected by Lisbon fire department, *Regimento de Sapadores Bombeiros* (RSB)[2], and the national weather authority, *Instituto Português do Mar e Atmosfera* (IPMA)[3].

The emergency events described in the data set were reported throughout January 2013 and December 2018. For each one of the reported floods, the following variables were obtained:

- Geospatial coordinates using the WGS84 reference system, comprised of latitude and longitude (referred to as “lat” and “lon”);
- Timestamp (referred to as “datetime”);

[2] RSB website: https://www.lisboa.pt/cidade/seguranca-e-prevencao/regimento-de-sapadores-bombeiros
[3] IPMA website: https://www.ipma.pt/en/

- Event type (referred to as “Event_type”, where flood events are represented as codes “3500”, “3501” and “3502”);
- Event severity (referred to as “target”, where 1 stands for high severity, and 0 for low severity, given that all floods are labelled as 1).

Weather data consists of 52,584 observed measurements gathered from January 2013 until December 2018. The data set contains hourly measures of temperature, humidity, sun exposure, precipitation, wind speed and wind direction. These measures were obtained from three meteorological stations located in the city of Lisbon:

1. *Estação Aerológica Gago Coutinho de Lisboa* (referred to as “gago_coutinho”)
2. *Instituto Geofísico do Infante Dom Luís* (referred to as “geofisico”)
3. *Estação Meteorológica de Lisboa - Tapada da Ajuda* (referred to as “ajuda”)

As the entire city is covered by just three weather sensors, each incident reported by the RSB had to be assigned to one of these three sensors. This process consisted of measuring the L2 distance (i.e. shortest path from point A to point B) from the location of each incident to the three weather stations, where the closest one would be assigned as the reference station for that data point. This way, it was possible to establish a coverage area for each weather station and assign approximate weather measurements for every incident in the city. In Figure 2 it is presented the location of each weather station in the city of Lisbon along with their influence areas, based on the L2 distance for a 100 m^2 grid. The total area corresponds to the area of Lisbon municipality, situated in Portugal, with an area of about 100 km^2 in the north bank of the mouth of Tagus river, the longest river in Iberian Peninsula. Lisbon presents a typical Mediterranean climate with short and mild winters and warm summers.

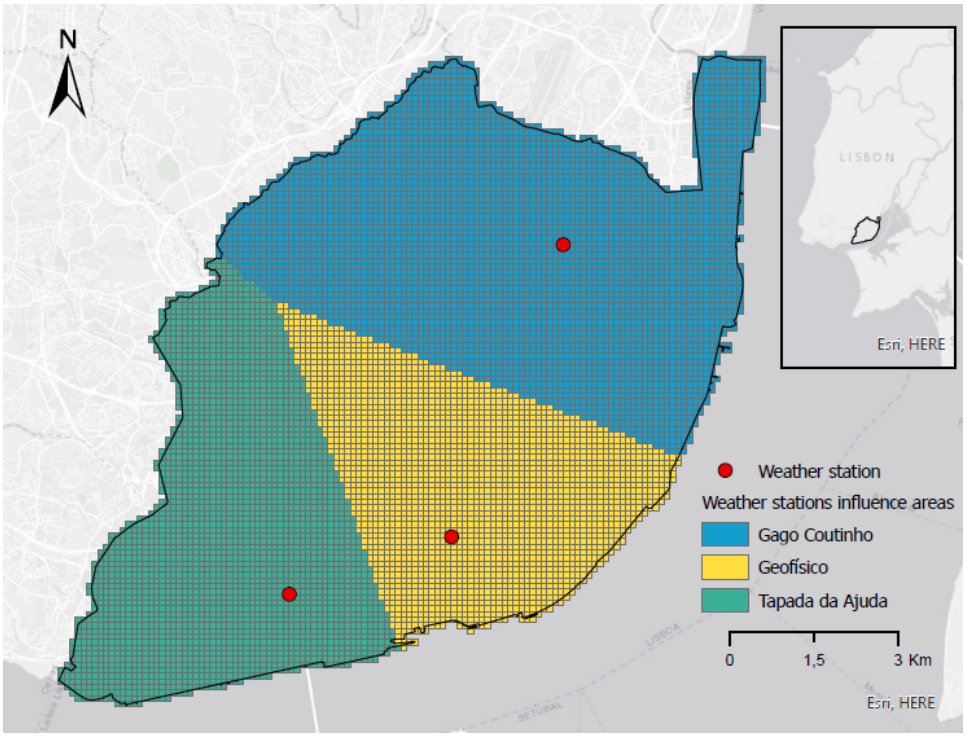

Figure 2. *Weather stations location and coverage.*

3.2. Data preparation

After merging the weather data and the flood reports data, the initial data set consisted of 9 input variables, as shown in Table 2:

Variable	Type	Description
temp	Continuous	Temperature in degrees Celsius.
hum	Continuous	Relative humidity ratio in %.
precip	Continuous	Precipitation in mm.
sun	Continuous	Sun exposure in W/m².
wind_speed	Continuous	Wind speed in m/s.
wind_dir	Continuous	Wind angular direction.
station	Categorical (nominal)	Name of the weather station.
datetime	Categorical (interval)	Timestamp.
target	Categorical (binary)	Flood occurrence flag.

Table 2. *Feature set before data preparation.*

In order to clean and prepare data for modelling, data quality issues, such as missing values, anomalies and outliers were identified and treated in two steps. Firstly, extreme values found in the data set were removed and replaced by null values. Secondly, for treating missing values

and minimizing noise created by imputation, the following strategy was used, considering the geospatial and temporal dimensions:

1. Use measurements from the nearby weather station based on the L2 distance (given a two-dimensional Euclidian space);
2. Use measurements from recent observations using a forward and backward time search;
3. Use k-nearest neighbours (KNN) for the two nearest neighbours (k=2);

Given that the variable for wind direction still had over 241 observations with missing data, one last step was implemented to avoid discarding useful data by using a random forest to impute the remaining data points. Doreswamy et al. [27] reported that using a Random Forest for imputation resulted in relatively low error compared to other typical ML methods.

After the data cleaning steps, all variables were filtered according to the flood occurrence flag (variable “target”), such that it is possible to compare behaviours and outcomes. By doing so, a bivariate analysis was used to identify patterns between each weather variable and the variable “target”, summarized in Figure 3.

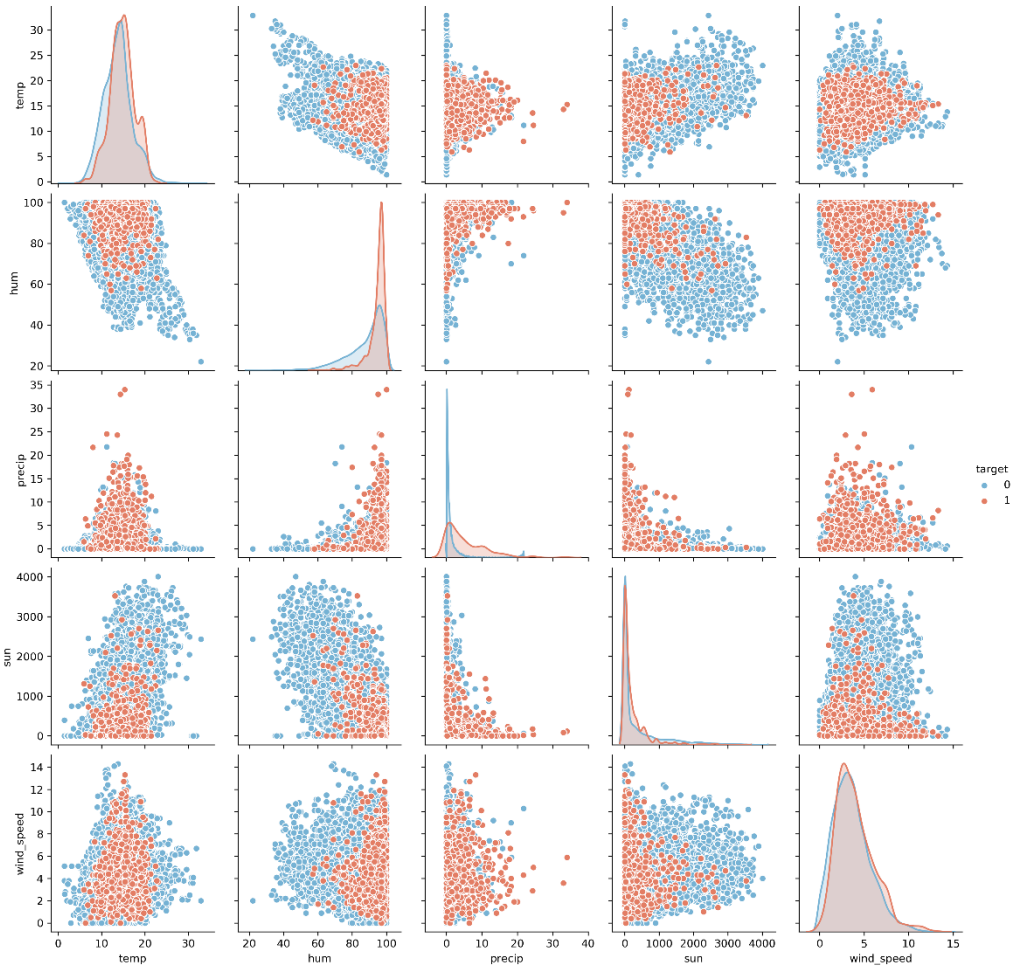

Figure 3. *Bivariate distribution of weather data and density estimation.*

Based on the initial findings obtained for the input variables, it was noticed that variables "wind_dir" and "datetime" required additional treatment to improve representation and provide more meaningful information. The values for wind direction ("wind_dir") represent angular measurements (i.e. azimuths) and should be interpreted as having equal magnitudes, such as categorical values. As for the timestamp ("datetime"), temporal data could be more discriminative if its components were extracted (i.e. day, month, year and hour). Therefore, proxies were created to represent timestamp and wind direction variables:

- Variable "weekday", representing the day of the week (e.g. "Sunday");
- Variable "period", representing the period of the day using intervals of 4 hours (e.g. "Late Night (0h-4h)", "Early Morning (4h-8h)");

- Variable “season”, representing the calendar seasons (e.g. “Winter”);
- Variable “wind_cardinal”, representing the wind direction using cardinal and intercardinal directions (e.g. “North”, “Northeast”).

In addition to these proxies, two new variable sets were proposed for the weather data:

1. City average: every measurement was averaged across all three weather stations for every hour in a way this feature could be used to perceive the general behaviour at the city level;
2. Moving average: every measure was accumulated using a 2-hour window prior to the observed hour. This feature could be used to identify a causality between the recorded flood events and prior weather conditions. Moving averages are commonly used to smooth time-series events by decreasing short-term fluctuations during the observation window. Based on the literature, the performance of a moving window method has been shown to depend on three parameters: the length of the time series, the length of the window, and the time step [8]. After some experiments, a 2-hour window for every hourly step was found to be the most robust setting (see Figure 4).

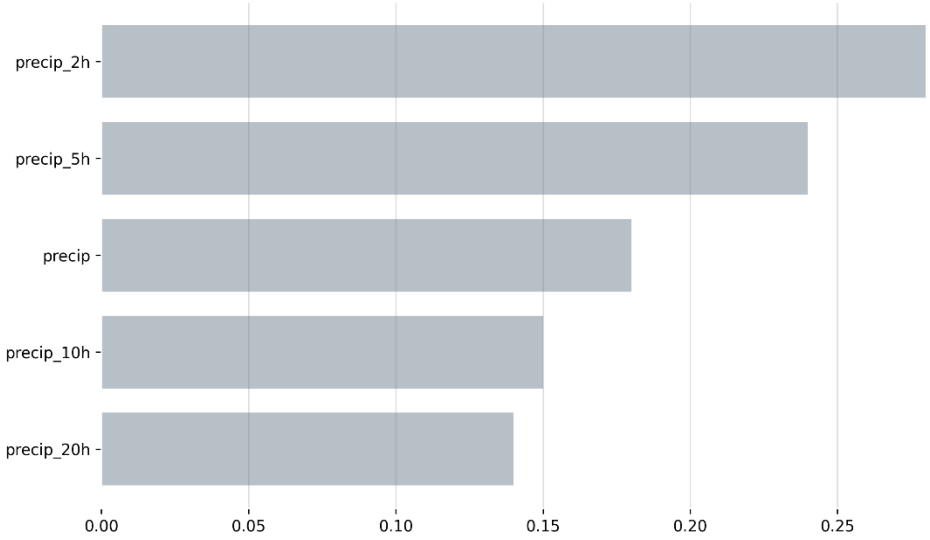

Figure 4. *Impurity-based feature importance using a Random Forest, given different window sizes.*

These new features were then used for deepening the exploratory analysis before being subject to modelling, along with the remaining features. The feature importance of these variables will be further discussed in section 3.3, during the modelling phase.

After merging the entire weather data history for all three weather stations and adding all flood records, a heavy class imbalance problem was noticed: 158,685 total hourly records for 2,154 flood occurrences (1.36%). To overcome this issue, the following criteria was used to remove data points with negative response:

- If no precipitation was recorded for all three weather stations, and;
- If no precipitation was recorded for the past 2 hours.

This strategy yielded only 20,943 total hourly records, of which 1,837 represented flood occurrences (8.77%). Despite a substantial loss of positive responses in the data set (over 15%), the removed observations could represent a potential coherence issue, as rainfall and floods are expected to have a causal relationship. Keeping these observations would only increase the noise in the model input and, in turn, decrease its predictive power.

Categorical features were dichotomized using “one-hot encoding”, allowing these features to be represented as binary vectors, which are more convenient to use throughout the modelling phase. Continuous features were scaled using the standardisation (z-score) method, as they were initially represented using different scales/measurement units. This method recomputes the mean to 0 and variance to 1, as it is a common requirement for many ML models.

At the end of the data preparation phase, the initial 8 variables were converted into 44 independent variables to be considered for the modelling phase.

3.3. Modelling

The modelling phase was conceived into three steps: preliminary assessment, validation and optimization. The first two steps will evaluate the performance of a set of six classification models, namely Logistic Regression (LR), Support Vector Machine (SVM), Gaussian Naïve-Bayes (NB), Random Forest (RF), K-Nearest Neighbours (KNN) and Multi-Layer Perceptron (MLP). The optimal model identified during the first two steps will be used in the optimization step, where hyperparameters will be fine-tuned for increasing performance. Additionally, the hyperparameters used in the preliminary assessment and validation will be based on the default settings implemented in scikit-learn.

For the preliminary assessment step, a single instance of each model was trained with all 44 variables. The input data set was split into two subsets, using a 75-25 sampling ratio, where 75% of the data was used for training and 25% was used for testing. In order to express the predictive power of each model, a selection of five performance measures was used, namely Accuracy, AUC, Recall, F1 and MCC, providing the following results (see Table 3).

	Measures				
Model	**Accuracy**	**AUC**	**Recall**	**F1**	**MCC**
Logistic Regression	0.95	0.92	0.54	0.66	0.64
Support Vector Machine	0.95	0.89	0.54	0.67	0.67
Gaussian Naïve-Bayes	0.87	0.90	**0.75**	0.51	0.48
Random Forest	**0.96**	**0.94**	0.67	**0.77**	**0.76**
K-Nearest Neighbours	0.95	0.90	0.61	0.69	0.67
Multi-Layer Perceptron	0.95	0.92	0.71	0.74	0.71

Table 3. *Preliminary model evaluation.*

In Figure 5, the Receiver Operating Characteristic (ROC) curve assessment was performed to compare the model's outcome given the true positive and false positive rates.

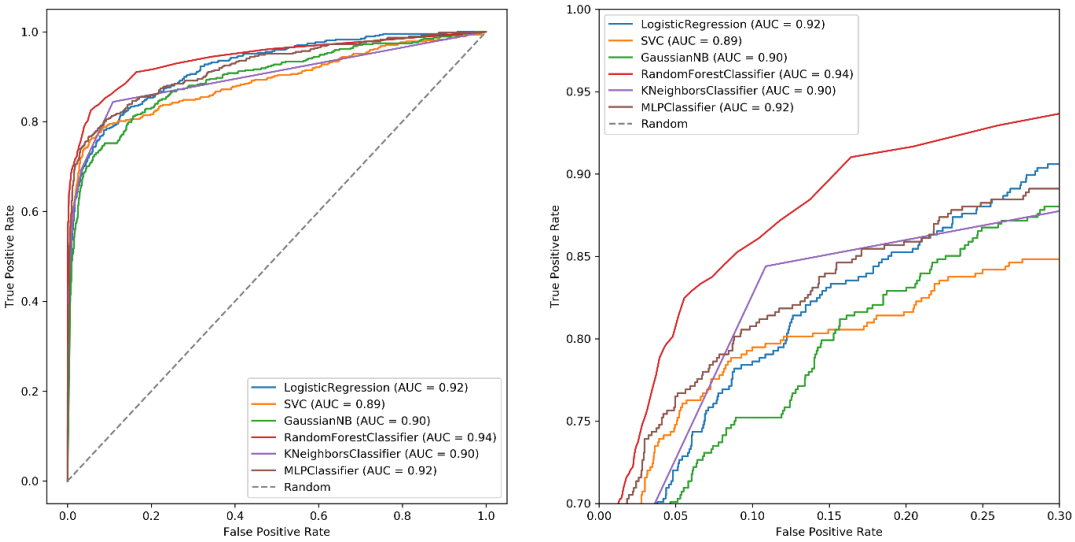

Figure 5. *Receiver Operating Characteristic (ROC) curve.*

The following remarks were considered after this preliminary assessment:

- Overall, RF was the best performing model for all indicators, except Recall;
- MLP had a slightly worse performance compared to RF, but showed a higher Recall (i.e. higher sensitivity to the relevant instances, being represented by the flood occurrences). Nonetheless, neural networks such as MLP could potentially be optimized and surpass RF;
- LR, SVM and KNN had subpar results. Despite using different heuristics, they had a similar predictive power;
- Among all models, NB had the worst performance. However, it presented an interesting trade-off between Accuracy (worst) and Recall (best).
- Considering the low prevalence described during the data preparation phase, where only 8.77% of the observations represented flood events, measures like Accuracy and AUC are less capable of effectively measuring the predictive power for the events to be detected, represented in the minority class. Measures like Recall, F1 score and MCC are more suitable for measuring the performance for imbalanced data sets, when relevant instances are poorly represented.

The findings described in the preliminary assessment still lack statistical significance, given that the selected sample could not be representative of the entire population. Additionally, the lack of statistical significance represents a major concern regarding the ability of a model being able to predict “unseen data” and minimize the risk of “overfitting”, as using different subsets of data and/or initial state while fitting these models could greatly impact the estimated outcome.

To overcome these concerns, the same models were re-trained in the validation step, but this time using k-fold cross-validation. In this specific case, ten subsets (i.e. folds) were drawn from the training set and used to train k instances of the selected model using combinations of k-1 subsets. The remaining subset is used to validate the trained model. To better represent the class distribution for the target variable, stratified sampling was used during cross-validation to maintain the imbalanced nature of the target variable for each k subset.

To benchmark their performance, only MCC was considered among all metrics. As mentioned by Luque et al. [22], certain classification performance metrics can be heavily impacted by class imbalance. In the literature, the MCC is reported as an adequate option for representing classification error and minimizing bias when compared to other metrics, such as Accuracy and F1 score.

Scores were then aggregated to compute the final score for each model. Based on the final scores (shown in Table 4 and Figure 6), the previous findings were corroborated: RF model showed a superior predictive power, while keeping the deviation to a minimum.

	MCC Score (k=10)	
Model	**Mean**	**Standard Deviation**
Logistic Regression	0.65	0.03
Support Vector Machine	0.68	**0.01**
Gaussian Naïve-Bayes	0.48	0.02
Random Forest	**0.77**	0.02
K-Nearest Neighbours	0.69	0.03
Multi-Layer Perceptron	0.71	0.03

Table 4. *K-fold validation average results (k=10).*

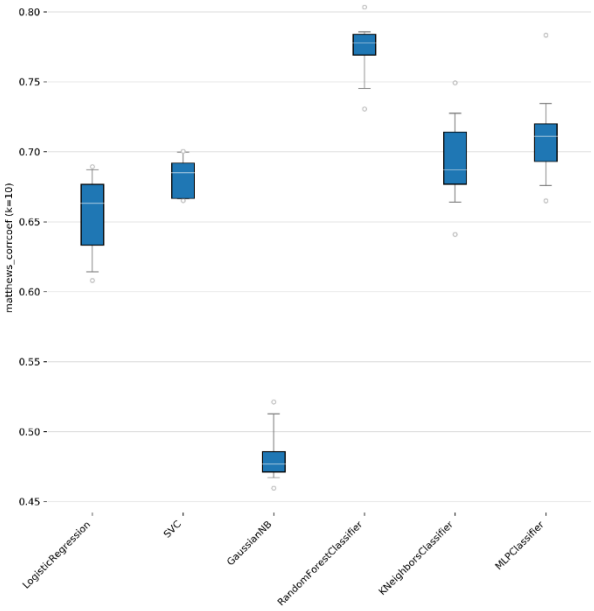

Figure 6. *K-fold validation results using Matthew's Correlation Coefficient (k=10).*

As proposed in the literature [28], a cross-validation strategy using ten multiple runs with ten folds each (i.e. 10x10 cross-validation) was used to provide more reliable performance measures. To obtain these measures, a Student's t-distribution was considered for 100 observations (10 runs of 10 folds, resulting in 100 instances) with a confidence level of 95% (α=0.05) and 10 degrees of freedom. The results found are in line with the previous findings and according to Bouckaert & Frank [28] should possess higher significance, better reproducibility and consistency, as shown in Table 6.

Model	MCC score (k=10 x 10)			
	2.5th quantile (α/2)	**Mean (μ)**	**97.5th quantile (1 – α/2)**	**Standard Deviation (σ)**
Logistic Regression	0.58	0.66	0.73	0.04
Support Vector Machine	0.61	0.68	0.76	0.04
Gaussian Naïve-Bayes	0.42	0.48	0.54	0.03
Random Forest	**0.71**	**0.77**	**0.84**	0.03
K-Nearest Neighbours	0.63	0.70	0.76	0.03
Multi-Layer Perceptron	0.64	0.71	0.77	0.03

Table 5. *10x10 cross-validation results for a Student's t-distribution (α=0.05, df=10, n=100).*

Finally, the optimization step was put into place to improve model performance for the RF model, being found the most performant in the previous steps. A 75-25 sampling split was used, where 75% of the data was allocated for training and validation, using k-fold cross-validation with stratification. The remaining 25% was allocated for testing. Using these settings, new instances of the RF model were trained using three strategies: dimensionality reduction, data resampling and hyperparameter tuning.

Dimensionality reduction aimed to remove less informative and redundant variables which could potentially decrease model performance, filter and wrapper methods were used. Previous studies indicate these methods, as dimensionality reduction techniques, should be able to remove irrelevant, redundant or noisy features, usually leading to better learning performance, i.e., higher learning accuracy, lower computational cost, and better model interpretability [29]. As a filter method, the correlation matrix for all quantitative variables was used to identify redundant variables. As shown in Figure 7, it was noticed that the original weather variables showed high redundancy compared to the engineered weather features (i.e. city average and cumulative moving average).

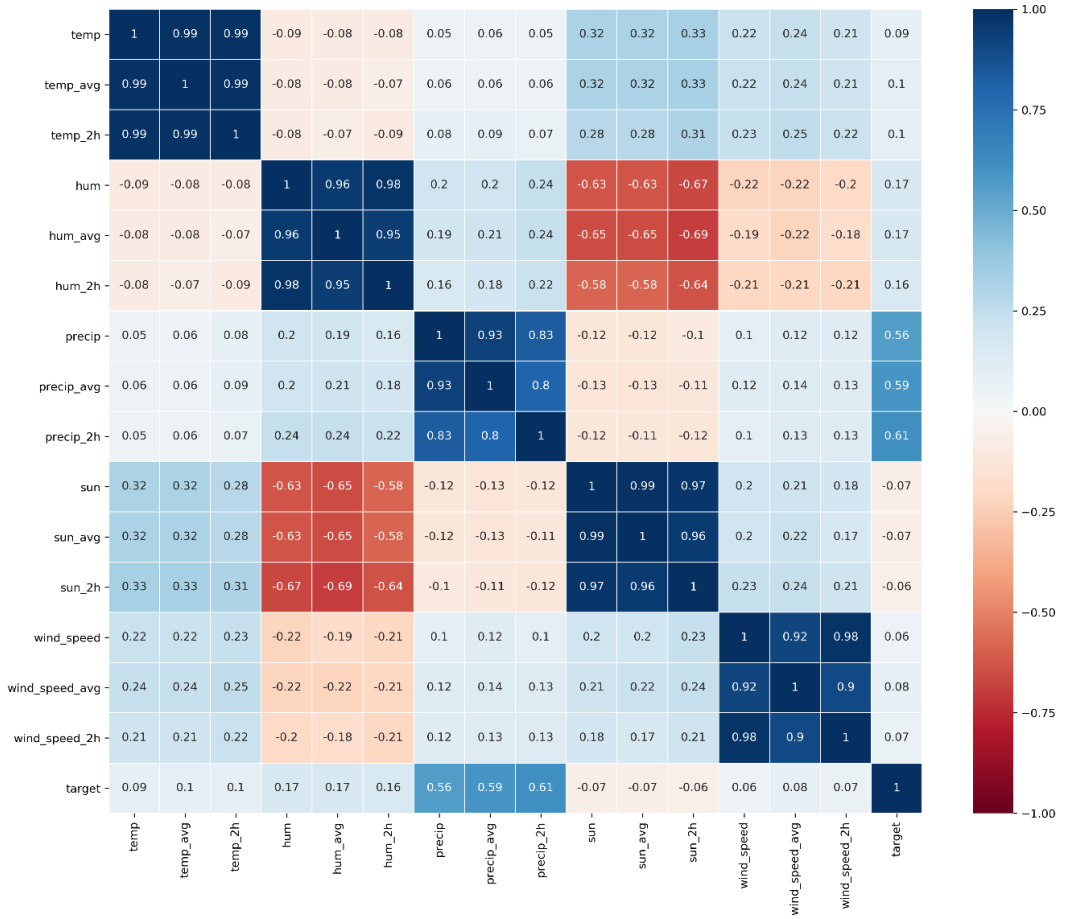

Figure 7. *Correlation matrix for weather variables after feature engineering.*

As a wrapper method, backward feature elimination was used to evaluate the discriminative power for each variable, where all features had their importance measured and the least important features were eliminated. This method recursively evaluates the entire feature set and eliminates less informative features until the optimal number of features is achieved. As stated in the literature, this process may capture interacting features more easily than other methods [30].

Both methods led to the removal of 12 out of 44 variables, however, regardless of the feature selection method, no significant improvement was observed in model performance.

Data resampling was used to deal with the inherent class imbalance between flood and non-flood records, a step using alternative sampling methods was proposed to reduce the imbalance between the two classes. Imbalanced data poses a difficulty for learning algorithms, as they

will be biased towards the majority group. At the same time usually the minority class is the one more important from the data mining perspective, as despite its rareness it may carry important and useful knowledge [31]. Some experimentation was performed with over-sampling the minority class, under-sampling the majority class and a combination of both methods (i.e. SMOTE+ENN), as introduced by Batista et al. [32]. As shown in Figure 8, all three methods resulted in degrading the MCC in exchange of improving the Recall. That is, the resulting model would show a higher classification error in favour of being more sensitive to predicting flood occurrences.

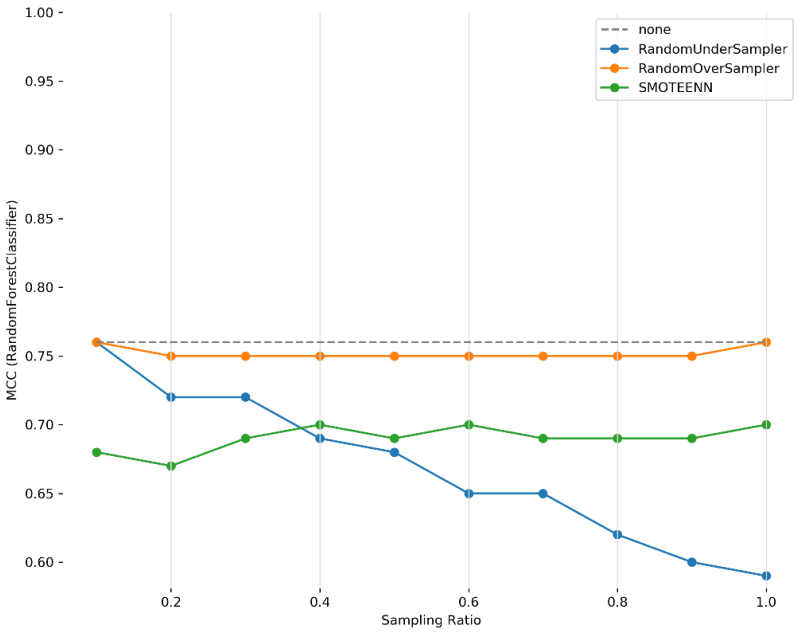

Figure 8. *Resampling performance using different methods and ratios (at Sampling Ratio 1.0, classes are balanced).*

Hyperparameter tuning was implemented using a grid search, cross-validation algorithm was in order to optimize the hyperparameters used for training the RF model. During this process, a slight gain in performance was observed when setting the number of estimators to 500 and the information gain criterion to “gini”. However, these optimal settings did not translate into any significant improvement in performance when evaluated with unseen data (test sample), being later disregarded. As explained by Mantovani et al. [33], it is important to consider that

the tuning that yields the model with highest performance in a data set might not perform so well on other data sets.

Overall, the strategies described in the optimization step were able to provide marginal improvement in model performance (less than 1%). While this improvement is not really significant, the optimized model was able to substantially reduce the number of required inputs for predicting floods, greatly decreasing the computational cost (i.e. training time).

Regarding feature importance, Figure 9 shows the weights attributed to each one of the 21 variables during the model training. As reported during the optimization step, the city averages (identified by the suffix “avg”) and cumulative moving averages (identified by the suffix “2h”) showed higher importance than the original weather variables when being used as predictor for the occurrence of floods.

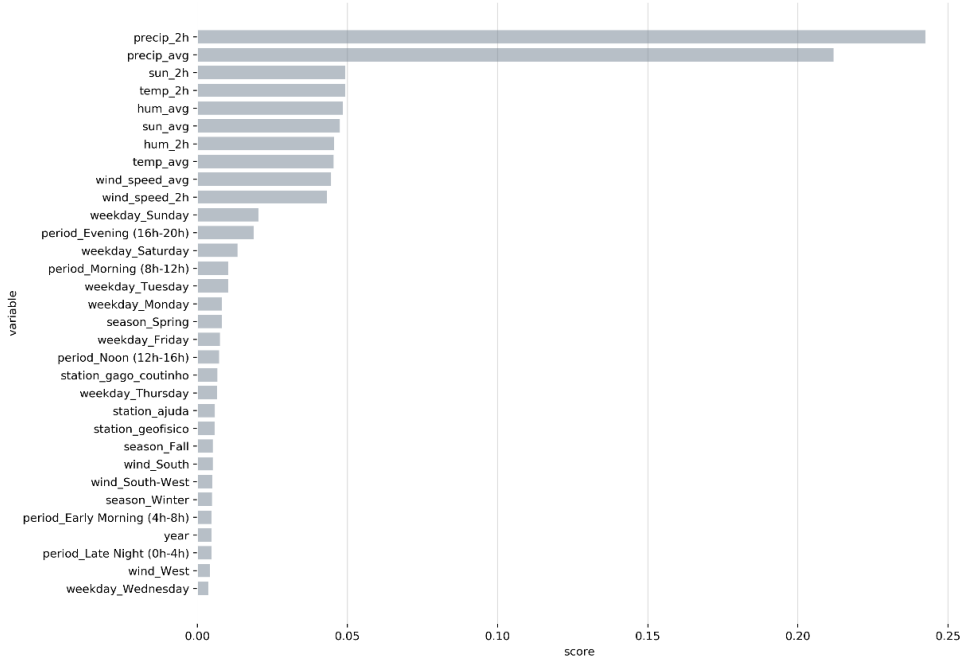

Figure 9. Feature importance estimated by a Random Forest.

3.4. Evaluation

Once the optimal classifier was found, additional steps were required to accurately establish where a flood will occur, if the required conditions are met. As the classifier was trained

exclusively on weather data sourced by three stations, it has a limited representation of the spatial dimension, as the predictions are defined at the station level. Additionally, it was observed that some areas are more prone to flooding than others, which might imply that other underlying factors and causes could have been dismissed from the ML model. For this purpose, the application of GIS is proposed to identify spatial heterogeneity and to design a spatial vulnerability indicator, as suggested in the literature [13] [26]. Finally, this indicator could be used to fine-tune the predictions obtained in the previous step by determining the locations most subject to floods.

The Hot Spot analysis uses historical georeferenced data to find contiguous areas with a higher prevalence of floods based on their spatial relationship, objectively measured by the Gi* statistic. The spatial unit used for this analysis was defined on a grid at the city level, comprised of 100x100m cells, having in total 8,986 cells.

The measured spatial relationship is conditioned to a distance parameter which determines how geospatial data points will be clustered. This parameter can be optimized using the Incremental Spatial Autocorrelation method, which is able to select the optimal distance based on its ability to maximize the z-score for the Gi* statistic. By using this method, the optimal distance was found at 324 meters, resulting in the map shown in Figure 10.

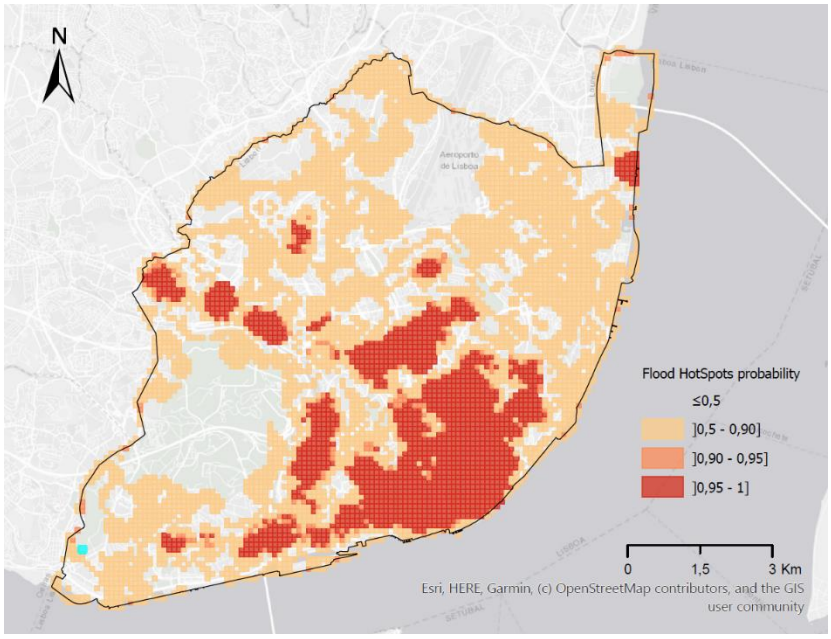

Figure 10. *Hot spots analysis for observed floods from 2013 to 2018.*

By performing the Hot Spot analysis, the spatial prevalence of floods could be identified and quantified and, along with the flood prediction scores provided by the ML classifier, it could be used for factoring the risk of flood for critical areas. Therefore, the p-value was used to represent the statistical significance for each grid cell to be characterized as a hot spot, which was then used alongside the predicted scores from the classification model so each grid cell could have a single metric to represent a risk index. This function computes the weighted average between both scores (ranging between 0 and 1), which can be adjusted to favour either the ML model (RF) or the GIS model (HS), as expressed below:

$$Flood\ Risk\ Index\ =\ weight_{RF} * score_{RF} + weight_{HS} * (1\ -\ pvalue_{HS})$$

For the experiments performed in this paper, the weights were not adjusted (i.e. set to 0.5), returning a simple average between these scores.

4. RESULTS & DISCUSSION

In this section, the results obtained by the flood risk index are presented and its significance is discussed.

After using the modelling strategy described in the previous chapter, the following settings yielded the best results when training the ML model:

- Random Forest with 100 estimators and “entropy” as the information gain criterion;
- From the 44 input features, 32 were used;
- No resampling, class imbalance was kept at 8.77%.

As stated by Aziz et al. [9], most of the hydrologic processes are non-linear and exhibit a high degree of spatial and temporal variability. This statement is in line with the findings obtained while training and testing each predictive model, where non-linear models showed better overall performance, especially when it comes to minimizing the sensitivity for the minority class (i.e. Recall). Moreover, the methods used in this project yielded better results for the RF model, while other non-linear models, such as MLP and KNN, provided sub-optimal results. Linear models, such as LR and SVM, seemed to maximize Accuracy while lacking Recall. Conversely, the NB model had the worst results due to the lack of Accuracy, despite the highest Recall.

Given the implications of working with an imbalanced dataset, all models were objectively evaluated using MCC, as other measures would lead to higher bias and poor representation of errors. MCC was found to be the most conservative measure in comparison to Accuracy, AUC, Recall and F1.

In terms of the features used throughout the study, rainfall was found to be the best predictor for flood detection among all weather variables used in this study, in accordance with Aziz et

al. [9]. Furthermore, experiments performed in this paper showed that the predictive power for this variable could be further enhanced by engineering aggregated features, especially when using moving averages with a short-length window.

The flood hazard map generated by the Hot Spot analysis provided optimal results using a distance threshold (i.e. radius) of 324 meters, when considering a 100x100m spatial unit. The p-values obtained from the hot spots were used in combination with the scores provided by the ML model and tested against the recent history of storms in the city.

In December 2019, storms Elsa and Fabien were responsible for major disruptions in urban areas due to intense precipitation, as a significant number of floods were reported. During this period, 116 flood occurrences were reported, of which 73 occurrences were correctly identified by the ML model, representing a sensitivity of 0.63. After applying the results obtained from the Hot Spot analysis to compute the Flood Risk Index, it was possible to better represent the spatial dimension and further boost the ML predictions. This time 97 occurrences were correctly identified, improving the sensitivity to 0.84.

As shown in Figure 11, at 22h on December 19th (the day the storms hit the city), several events were reported throughout the city, in accordance with the high Flood Risk Index reaching up to 0.85 in the highlighted areas.

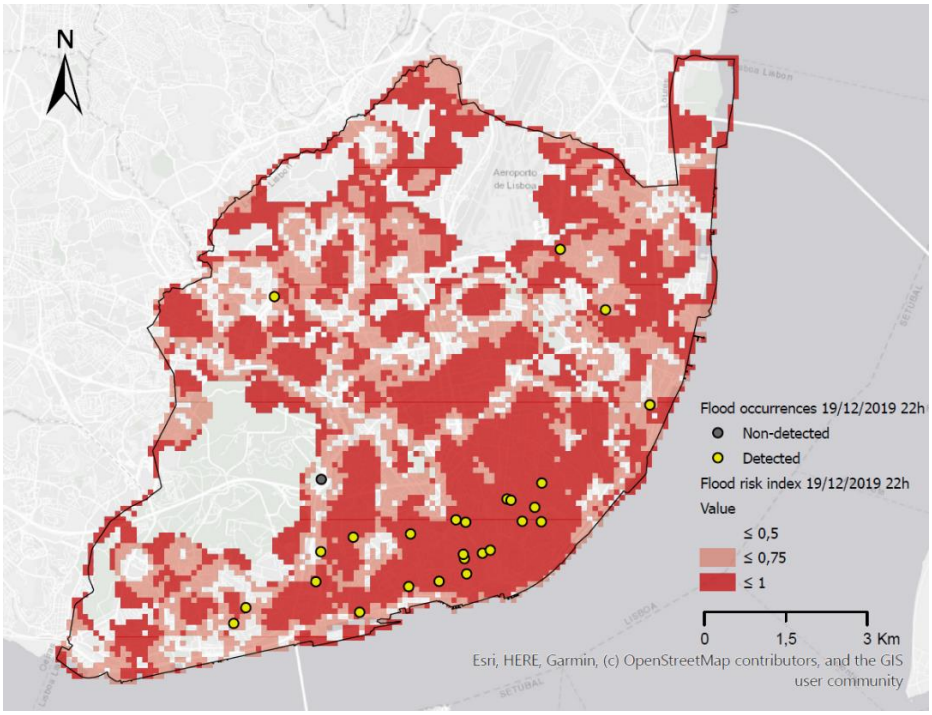

Figure 11. *Summary results for flood detection using the Flood Risk Index.*

However, for a particular hour on December 16th (3 days before the storms), no significant events were reported (as shown in Figure 12). Still, the Flood Risk Index was significantly high, reaching up to 0.65 in the highlighted areas. This behaviour indicates that the model is prone to classifying false positive occurrences.

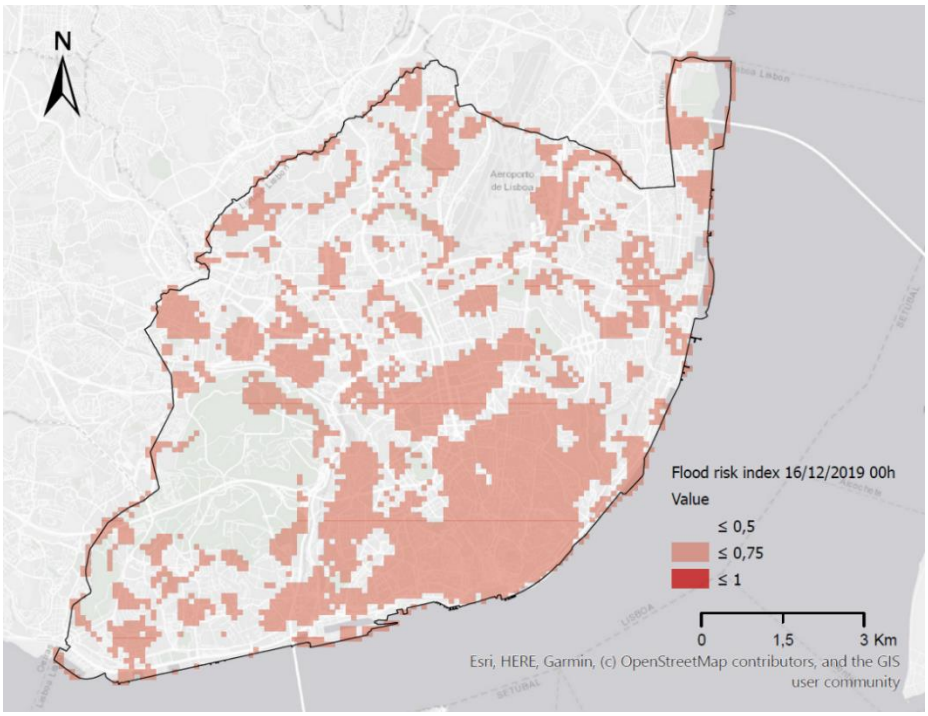

Figure 12. *Summary results for flood detection using the Flood Risk Index.*

This approach led to increasing the model sensitivity based on the likelihood inferred from the historical data and the observed hot spots. It is understood that the number of false positives could increase due to the higher sensitivity, but this behaviour can be controlled by adjusting the weights attributed to each score.

While using a combination of ML and Hot Spot analysis can highlight useful spatial patterns for predicting floods, these patterns are still not able to provide a priori theoretical explanation for their occurrence and prevalence in certain areas. For this purpose, additional morphological and hydrological data would be necessary to determine vulnerability factors and establish a causal relationship between these factors and the likelihood of floods [15].

As shown by Mosavi et al. [12], flood modelling works in the literature are mostly constrained to either short-term or long-term predictions. The hybrid approach herein proposed is able to integrate long-term patterns via the spatial modelling of floods into short-term predictions given by observed weather conditions, attempting to achieve a better representation of underlying spatial characteristics and inherent local vulnerabilities. Moreover, by performing spatial modelling with GIS statistics it is possible to determine whether or not identifiable spatial patterns exist (e.g. spatial heterogeneity, dependence) , as postulated in [23] [24]. In this study it's possible to infer that floods should be understood as a heterogeneous process (i.e. features present different relationships across space) These findings reinforce the need for spatially explicit models and suggests adjustments to general ML models before they can be properly used for flood modelling.

From a risk management standpoint, the proposed solution could be effectively integrated by local authorities into their resilience strategy for the city of Lisbon, according to three main capacities presented by Serre [5]: resistance, absorption, and recovery. That is, this solution

could be used to (1) identifying high-risk areas that might require proactive measures to mitigate vulnerabilities, (2) improving the allocation and readiness of response services when flooding conditions are observed and (3) eventually, minimizing recovery times, in line with Tingsanchali [11]. However, the implementation of prototype models to serve as leading indicators and early warning systems in disaster management are still in early stages of development and must consider an interdisciplinary approach to investigate these phenomena [13].

5. CONCLUSION

The approach developed throughout this project showed promising results for predicting floods with limited sensing data. This paper demonstrates that, by combining ML and GIS, it is possible to determine the key predictors and conditions for a flooding scenario by using a limited amount of data.

From the ML standpoint, non-linear models were more capable of detecting floods and, among those, the most performant model was a Random Forest, yielding a Matthews Correlation Coefficient of 0.77. As identified in the modelling phase, the 2-hour window moving average for rainfall was found to be the most important flood predictor.

The GIS model was able to adjust the sensitivity of the predictions obtained from the ML model based on the hot spots observed in the history. Areas where hot spots were observed were considered to have an increased likelihood of flooding, thus having a higher Flood Risk Index. When using equivalent weights for both models, the model sensitivity reached 0.84. However, the increased sensitivity led to a higher false positive rate, indicating the need for further adjustments in the model threshold and/or score weights.

Additionally, the GIS model also improved the spatial representation; as the ML model was able to compute predictions at the weather station level, the combination of both models allowed to compute a risk index for every 100 m^2 cell in the entire city.

As for the limitations found in the data set, despite relying exclusively on weather conditions and lacking spatial variability for these conditions (as this data was collected from only three stations), this approach was consistently capable of predicting floods with a high degree of confidence. Moreover, its robustness indicates it could be replicated in diverse scenarios and its performance could be improved when using high-resolution spatial data, supported by a larger number of weather stations.

Suggestions for future work associated with the proposed model:

- This paper did not consider a cost function for misclassification, which could be developed in future works to optimize the allocation of resources in the context of a flood management strategy. Misclassification of disaster events is a serious concern as it could result in a substantial economic and social impact.
- The weights used for calculating the Flood Risk Index could be further studied to obtain the optimal settings for boosting the model's predictive ability.

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Habitat III, U. N. (2017). Issue papers 21 - Smart Cities. In United Nations Conference on Housing and Sustainable Urban Development, 142-149. https://habitat3.org/wp-content/uploads/Habitat-III-Issue-Papers-report.pdf.

[2] D. Serre & C. Heinzlef (2018). Assessing and mapping urban resilience to floods with respect to cascading effects through critical infrastructure networks. International Journal of Disaster Risk Reduction, 30, 235-243. https://doi.org/10.1016/j.ijdrr.2018.02.018.

[3] R. Dankers & L. Feyen (2008). Climate change impact on flood hazard in Europe: An assessment based on high-resolution climate simulations. Journal of Geophysical Research: Atmospheres, 113(D19). https://doi.org/10.1029/2007JD009719.

[4] C. Heinzlef & D. Serre (2020). Urban resilience: From a limited urban engineering vision to a more global comprehensive and long-term implementation. Water Security, 11, 100075. https://doi.org/10.1016/j.wasec.2020.100075.

[5] D. Serre (2018). DS3 Model Testing: Assessing Critical Infrastructure Network Flood Resilience at the Neighbourhood Scale. In Urban Disaster Resilience and Security (pp. 207-220). Springer, Cham. https://doi.org/10.1007/978-3-319-68606-6_13.

[6] C. Heinzlef, B. Robert, Y. Hémond & D. Serre (2020). Operating urban resilience strategies to face climate change and associated risks: some advances from theory to application in Canada and France. Cities, 104, 102762. https://doi.org/10.1016/j.cities.2020.102762.

[7] C. Heinzlef, V. Becue & D. Serre (2019). Operationalizing urban resilience to floods in embanked territories – Application in Avignon, Provence Alpes Côte d'azur region. Safety Science, 118, 181–193. https://doi.org/10.1016/j.ssci.2019.05.003.

[8] J.M. Cunderlik & D.H. Burn (2003). Non-stationary pooled flood frequency analysis. Journal of Hydrology, 276(1-4), 210-223. https://doi.org/10.1016/S0022-1694(03)00062-3.

[9] K. Aziz, A. Rahman, G. Fang & S. Shrestha (2014). Application of artificial neural networks in regional flood frequency analysis: A case study for Australia. Stochastic environmental research and risk assessment, 28(3), 541-554. https://doi.org/10.1007/s00477-013-0771-5.

[10] H. Mojaddadi, B. Pradhan, H. Nampak, N. Ahmad & A.H.B. Ghazali (2017)., Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. Geomatics, Natural Hazards and Risk, 8(2), 1080-1102. https://doi.org/10.1080/19475705.2017.1294113.

[11] T. Tingsanchali (2012). Urban flood disaster management. Procedia Engeneering 32, 25–37. https://doi.org/10.1016/j.proeng.2012.01.1233.

[12] A. Mosavi, P. Ozturk & K.W. Chau (2018). Flood prediction using machine learning models: Literature review. Water, 10(11), 1536. https://doi.org/10.3390/w10111536.

[13] A. Fekete, K. Tzavella, I. Armas, J. Binner, M. Garschagen, C. Giupponi, V. Mojtahed, M. Pettita, S. Schneiderbauer & D. Serre (2015). Critical data source; Tool or even infrastructure? Challenges of geographic information systems and remote sensing for disaster risk governance. ISPRS International Journal of Geo-Information, 4(4), 1848-1869. https://doi.org/10.3390/ijgi4041848.

[14] S.K. Jain, R.D. Singh & S.M. Seth (2000). Design flood estimation using GIS supported GIUHApproach, Water Resources Management, 14(5), 369-376. https://doi.org/10.1023/A:1011147623014.

[15] J. Chen, A.A. Hill & L.D. Urbano (2009). A GIS-based model for urban flood inundation, Journal of Hydrology, 373(1-2), 184-192. https://doi.org/10.1016/j.jhydrol.2009.04.021.

[16] H.M. Lyu, W.J. Sun, S.L. Shen & A. Arulrajah (2018). Flood risk assessment in metro systems of mega-cities using a GIS-based modeling approach Science of the Total Environment, 626, 1012-1025. https://doi.org/10.1016/j.scitotenv.2018.01.138.

[17] S.A. Mohamed & M.E. El-Raey (2020). Vulnerability assessment for flash floods using GIS spatial modeling and remotely sensed data in El-Arish City, North Sinai, Egypt. Natural Hazards, 102(2), 707-728. https://doi.org/10.1007/s11069-019-03571-x.

[18] A. Nadali, E.N. Kakhky & H.E. Nosratabadi (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. In 2011 3rd International Conference on Electronics Computer Technology (Vol. 6, pp. 161-165). IEEE. https://doi.org/10.1109/ICECTECH.2011.5942073.

[19] F. Martinez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernandez Orallo, M. Kull, N. Lachiche, M.J. Ramirez Quintana & P.A. Flach (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/tkde.2019.2962680.

[20] E. Kristoffersen, O.O. Aremu, F. Blomsma, P. Mikalef & J. Li (2019). Exploring the relationship between data science and circular economy: An enhanced CRISP-DM Process Model. In Conference on e-Business, e-Services and e-Society (pp. 177-189). Springer, Cham. https://doi.org/10.1007/978-3-030-29374-1_15.

[21] R. Wirth & J. Hipp (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1). Springer-Verlag.

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

[23] A. Getis & J.K. Ord (1992). The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis, 24(3), 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

[24] J.K. Ord & A. Getis (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical analysis, 27(4), 286-306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x.

[25] G. Blöschl, J. Hall, A. Viglione, R.A. Perdigão, J. Parajka, B. Merz, D. Lun, B. Arheimer, G.T. Aronica, A. Bilibashi, M. Boháč, O. Bonacci, M. Borga, I. Čanjevac, A. Castellarin, G.B. Chirico, P. Claps, N. Frolova, D. Ganora, L. Gorbachova, A. Gül, J. Hannaford, S. Harrigan, M. Kireeva, A. Kiss, T.R. Kjeldsen, S. Kohnová, J.J. Koskela, O. Ledvinka, N. Macdonald, M. Mavrova-Guirguinova, L. Mediero, R. Merz, P. Molnar, A. Montanari, C. Murphy, M. Osuch, V. Ovcharuk, I. Radevski, J.L. Salinas, E. Sauquet, M. Šraj, J. Szolgay, E. Volpi, D. Wilson, K. Zaimi & N. Živković (2019). Changing climate both increases and decreases European river floods. Nature, 573(7772), 108-111. https://doi.org/10.1038/s41586-019-1495-6.

[26] J.L. Leis & S. Kienberger (2020). Climate risk and vulnerability assessment of floods in Austria: Mapping homogenous regions, hotspots and typologies. Sustainability, 12(16), 6458. https://doi.org/10.3390/su12166458.

[27] Doreswamy, I. Gad & B.R. Manjunatha (2017). Performance evaluation of predictive models for missing data imputation in weather data. In 2017 International Conference on

Advances in Computing, Communications and Informatics (ICACCI) (pp. 1327-1334). IEEE. https://doi.org/10.1109/ICACCI.2017.8126025.

[28] R.R. Bouckaert & E. Frank (2004). Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 3-12). Springer. https://doi.org/10.1007/978-3-540-24775-3_3.

[29] J. Miao & L. Niu (2016). A Survey on Feature Selection. Procedia Computer Science, 91, 919-926. https://doi.org/10.1016/j.procs.2016.07.111.

[30] R. Kohavi & G.H. John (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324. https://doi.org/10.1016/s0004-3702(97)00043-x.

[31] B. Krawczyk (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0.

[32] G.E. Batista, R.C. Prati & M.C. Monard (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29. https://doi.org/10.1145/1007730.1007735.

[33] R.G. Mantovani, T. Horvath, R. Cerri, J. Vanschoren & A.C. Carvalho (2016). Hyper-Parameter Tuning of a Decision Tree Induction Algorithm. In 2016 5th Brazilian Conference on Intelligent Systems (BRACIS) (pp. 37-42). IEEE. https://doi.org/10.1109/BRACIS.2016.018.