Predicting Crime by Exploiting Supervised Learning on Heterogeneous Data

Úrsula R. M. Castro[a], Marcos W. Rodrigues[b] and Wladmir C. Brandão[c]
Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, Brazil

Keywords: Crime Analysis, Crime Prediction, Machine Learning, Supervised Learning, Supervised Classification.

Abstract: Crime analysis supports law-enforcing agencies in preventing and resolving crimes faster and efficiently by providing methods and techniques to understand criminal behavior patterns. Strategies for crime reduction rely on preventive actions, e.g., where perform street lighting and police patrol. The evaluation of these actions is paramount to establish new security strategies and to ensure its effectiveness. In this article, we propose a supervised learning approach that exploits heterogeneous criminal data sources, aiming to understand criminal behavior patterns and predicting crimes. Thus, we extract crime features from these data to predict the tendency of increase or decrease, and the number of occurrences of crimes types by geographic regions. To predict crimes, we exploit four learning techniques, as *k*-NN, SVM, Random Forest, and XGBoost. Experimental results show that the proposed approach achieves up to 89% of accuracy and 98% of precision for crime tendency, and up to 70% of accuracy and 79% of precision for crime occurrence. The results show that Random Forest and XGBoost usually perform better when trained with a short time window, while *k*-NN and SVM perform better with a longer time window. Moreover, the use of heterogeneous sources of data can be effectively used by supervised techniques to improve forecast performance.

1 INTRODUCTION

Crime is a violation of law and order that unbalances life in society and contributes to insecurity, promoting the increase of the sense of chaos and anarchy. According to the Universal Declaration of Human Rights, security is a guarantee, a fundamental right of citizens, allowing for the dignity of the human being (United Nations, 2015).

In contemporary society, public safety is one of the most critical problems, since recently the insecurity feeling has been increasing around the world (Oberwittler et al., 2010). Besides, crime control and prevention is not a trivial activity, demanding effective use of law-enforcing agencies resources. For example, the Brazilian Federal Constitution (Constitution, 1988) guarantees the security right for all citizens and assigns public safety as a duty of the State, and a right and responsibility of all. There is a general feeling that the government is unable to address the population's demand for public safety, especially in larger cities, both that violence is among the top major concerns of the Brazilians.

Misinformation on crimes is a problem that impairs the efforts against violence. Official statistics on violence in Brazil are far from reality, e.g., according to the National Secretary of Public Security (*SENASP*), only 30% of victims report crimes. The main reasons are the lack of trust in the police, the idea that the law-enforcing agencies do not take effective actions to solve crimes and punish the criminals, and the sense that they reporting have no positive impact on society. Thus, the actual number of crimes is higher than the officially reported ones. Nevertheless, some Web platforms, such as *WikiCrimes* and *Onde Fui Roubado*,[1] have been used to capture data on criminal actions. Also, social networks, such as *Facebook* and *Twitter*, can help to fill this information gap, since it is typical behavior among the victims to report crime incidents on their profile.

Thus, we propose a criminal prediction approach that exploits four different supervised learning techniques. We use *k*-NN (*k*-Nearest Neighbor), SVM (Support Vector Machine), RF (Random Forest), and

[a] https://orcid.org/0000-0002-0024-9362
[b] https://orcid.org/0000-0002-5329-3367
[c] https://orcid.org/0000-0002-1523-1616

[1] http://www.ondefuiroubado.com.br

Castro, Ú., Rodrigues, M. and Brandão, W.
Predicting Crime by Exploiting Supervised Learning on Heterogeneous Data.
DOI: 10.5220/0009392005240531
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 1*, pages 524-531
ISBN: 978-989-758-423-7
Copyright © 2020 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

XGBoost (eXtreme Gradient Boosting) to predict the tendency of increase or decrease and the number of occurrences of types of crimes by geographic regions. Further, we evaluate the effectiveness of the supervised techniques by contrasting their performance in three different crime datasets. The datasets are built from the OFF (*Official Records*) provided by Brazilian government, and UOF (*Unofficial Records*) from *Onde Fui Roubado* website. The third dataset, the CRIME, is a combination of OFF and UOF records.

Experiments attest to the effectiveness of our crime prediction approach, as well as of the features extracted from the heterogeneous data sources for prediction. Experimental results show that our approach achieves up to 84% of precision for crime tendency prediction in the UOF dataset, and 98% in the OFF dataset. Our approach achieves up to 79% of precision for a more difficult task of crime occurrence prediction in the UOF dataset, and 67% in the OFF dataset. Moreover, in the third dataset, some techniques such as XGBoost significantly improve their performance compared to the performance in the single datasets. Our main contributions are: a) We build a crime dataset (UOF) with public crime records extracted from the *Onde Fui Roubado* website; b) We characterize the OFF and UOF datasets, identifying relevant features about crime occurrences. Particularly, we exploit the gender and temporal aspects of crimes for crime prediction; c) We evaluate four different machine learning techniques for predicting trends and the crime's occurrence.

This article is organized as follows: In Section 2, we review the related literature on supervised learning, as *k*-NN, SVM, RF and XGBoost. In Section 3, we present related works on crime prediction. In Section 4, we present our crime prediction approach. In Section 5, we introduce the heterogeneous data sources used in our methodology. In Sections 6 and 7, we present the experimental setup and the results of the evaluation of our crimes prediction in heterogeneous datasets. Finally, in Section 8 we conclude this article by showing the strengths and weaknesses of our approach, laying directions for future research.

2 BACKGROUND

There are several techniques reported in the scientific literature that can be used to predict and classify events. In this article, we focus on four well-known supervised techniques that usually provide effective results when used for classification tasks.

The *k*-Nearest Neighbor (*k*-NN) is a versatile and robust classifier often used as a reference for more complex classifiers, such as neural networks. In particular, it is a non-parametric supervised learning approach that classifies instances based on the similarity between them. Predictions are made for a new instance by searching the entire training set, to get the k most similar cases (neighbors) (Duda et al., 2001). The *k*-NN has a single hyper-parameter, the number of nearest neighbors (k) of an instance that one needs to classify. Some authors suggest $k = \sqrt{N}/2$, where N represents the number of samples in the training set. However, an improved solution is to estimate the k value through k-fold cross-validation, minimizing the validation error (Celisse and Mary-Huard, 2018)

The Support Vector Machines (SVM) is a supervised classifier that uses a hyperplane to split the class representation, maximizing the distance between the instances of different classes (Hsu et al., 2008). SVM is efficient for many classification tasks, but presents a high computational cost for high-dimensional datasets. SVM classifiers have some hyper-parameters, such as the kernel function, the regularization parameter (C) to avoid misclassification, and the *Gamma* parameter. The kernel function maps the original data to find the best separation of the mapped space (hyperplane), and the *Gamma* parameter manage the influence of the training instances, i.e., low *Gamma* considers distant instances of the hyperplane, while a high value only finds the close instances of the hyperplane (Syarif et al., 2016). For high C, one selects a small margin for the hyperplane, and for low C, one selects a large margin for the hyperplane. Commonly, the *Gamma* parameter is estimated as $\gamma = 1/p$, where p is the number of attributes (data dimensions).

Random Forest (RF) is another robust supervised bagging approach commonly used for classification tasks (Breiman, 2001). Notably, a random forest is composed of a set of decision trees (weak predictors), where each tree depends on the values of a random vector sampled and with equal distribution among the trees. The output of each tree in the set converges to a single result that matches with the most trees (Ho, 1995). RF approaches have two hyper-parameters that impact in the behavior of the classification model, the number of trees ($ntree$), and the number of attributes ($mtry$) used in each division (Kohavi and John, 1997). For classification tasks, the literature suggests $mtry = \sqrt{p}$, where p is the number of attributes in N samples.

The eXtreme Gradient Boosting (XGBoost) is a highly robust supervised classifier that has recently been dominating applied in machine learning (Chen and Guestrin, 2016). The popularity of XGBoost comes from its scalability in all scenarios, and the

ability to solve many data science problems in a fast and accurate way. XGBoost is an implementation of gradient-boosted decision trees (GBDT) designed for speed and performance. The basic idea of boosting is to combine hundreds of simple trees with low accuracy to build a more accurate model. Every iteration generates a new tree for the model. The Gradient Boosting machine utilizes the gradient descent to generate the new tree based on all previous trees, driving the objective function towards the minimum direction (Friedman, 2001). The XGBoost classifier has some hyper-parameters, such as the tree size (*stree*), the learning rate (*lr*) and *Gamma*. The choice of *lr* is not crucial but should be significantly small ($lr = 0.1$) (Bühlmann and Yu, 2010), while the *Gamma* parameter specifies the minimum loss reduction required to make a node split.

3 RELATED WORK

Different approaches to predict crimes using machine learning have been reported in the literature. In (Bogomolov et al., 2014), the authors propose an approach that uses human behavioral data from mobile network activity combined with demographic information to predict crimes. In especial, they evaluate Logistic Regression, SVM, Neural Networks, Decision Trees, and different implementations of ensembles of tree classifiers on three different datasets. They conclude that the RF outperforms the others, achieving 70% of accuracy. One of the datasets used has the geographic division of the London, the other is the *Criminal cases dataset*, including the geo-location of the reported crimes without temporal data, and the last is an official dataset with metrics about the population of each geographic area.

In the same line, another prediction approach exploits five classification algorithms to predict crime locations in Manila (Baculo et al., 2017). The authors use official crime records from 2012 to 2016, with nine attributes to evaluate the effectiveness of the BayesNet, Naive Bayes, J48, Decision Stump, and Random Forest algorithms. Their experimental results show that the RF classifier outperforms the other algorithms in precision when trained with crime instances composed by the following attributes: type of crime, day of the week, year, location and location category, latitude, longitude, isholiday, and raining.

Recently, (Pradhan et al., 2019) proposes a prediction model to identify the type of crime that can occur in the city.They use an official crime dataset from 2003 to 2018 of San Francisco and propose a data preprocessing method to improve the prediction for the highly imbalanced dataset. To deal with this problem, the authors used three techniques: oversampling the minority classes, undersampling the majority class, and adjusting weights on the classifiers. Despite that, none of them showed a significant improvement in the recall or precision scores with *k*-NN, Multi-class Logistic Regression, Decision Tree, RF, and Naive Bayes algorithms.

The recent studies use time and space attributes for crime prediction, but the use of geographic features is under-exploited (Lin et al., 2018). The authors incorporated eighty-four types of geographic information related to vehicle theft in Taiwan. They use in their experiments a dataset from 2015 to 2018, and compare the performance of five approaches for crime prediction: *k*-NN, SVM, Random Forest, Deep Neural Networks. The results show that the Deep Neural Networks, with geographic features, outperforms the other approaches to predict vehicle theft occurrences.

In order to improve the crime prediction, (Belesiotis et al., 2018) studie how data from multiple heterogeneous sources can be used to make predictions about the spatial distribution of crime in large urban regions. The authors use six different datasets to learn models for predicting crime rates of fourteen types of crime. They exploit three techniques to investigate the impact of regression algorithms on their approach: Ridge regression, RF, and Support Vector Regression. Experiments show that Ridge and RF present a tendency to overestimate predictions, while SVM made balanced predictions. They argue that all datasets and types of data can potentially contribute to better prediction accuracy when appropriately combined.

DeepCrime (Huang et al., 2018) is a crime prediction framework that uncovers under-exploited dynamic crime patterns by using the evolving interdependencies between crimes with other present data in urban space. This framework enables to predict the crime occurrences of different types in each region of a city. The authors evaluate the framework using a dataset collected from New York in 2014, comparing the results against SVM, Auto-Regressive Integrated Moving Average, Logistic Regression, Multilayer Perceptron, Tensor Decomposition, Wide and Deep Learning, and Gated Recurrent Unit. Experimental results show that the *DeepCrime* framework outperforms the other baseline approaches.

Different from the previous related work reported in the literature, we exploit unofficial crime records collected from Web and an official crime record as a joint source of features about criminal patterns, testing and evaluating supervised learning techniques to predict the tendency and occurrence of different types of crimes in geographic locations. In the next sec-

tion, we present the architecture and components of our prediction approach.

4 CRIME PREDICTION

We present our approach to predict the tendency and occurrence of crimes which the workflow is illustrated in Figure 1. From Figure 1, we observe that, in the first step, the *Web Crawling* component collect unofficial crime records from the UOF website, storing them into the UOF dataset. Additionally, the OFF records are also collected and stored in the OFF dataset. In particular, the official crime reports come from the Brazilian government, which we obtained with the Department of State of Minas Gerais of Defense, and these reports build the OFF corpus.

In the second step, the *Feature Extraction* component filters crime records previously crawled, extracting from them crime-related features useful in prediction tasks. In particular, it discards incomplete records, e.g., records with no time or space information, as well as records outside a geographic region of interest. Additionally, a preprocessing procedure removes errors, including duplicated and noisy data, and performs text and encoding transformations. This procedure is paramount to improve the training and prediction effectiveness of the learning techniques used in the next step.

In the third step, the *Feature Transformation* phase performs transformations in the extracted features, deriving from the new features, such as the day of the week and the month in which the crime occurred, and whether the day is a holiday in the city. Moreover, it derives the neighborhood from the address, the period of the day from time, e.g., morning, afternoon, night, dawn, and it categorizes the city neighborhood in geographic regions. The goal of the new features is to characterize crime records better. Crimes that occur during the week may have different characteristics from crimes that occur on weekends, just as each month of the year and public holidays in the city have different frequencies and particulars related to the occurrence of crimes. We used the feature month to separate the training and test sets to make the predictions. About the period of the day, we believe that the crimes which happened in the same interval of time have a similar pattern. The idea of using the region as a feature is aiming to cluster the neighborhoods with the same characterizes and behavior patterns.

In the fourth step, our approach analyzes the features characterizing the OFF and UOF datasets by performing a complementarity analysis. Notably, it checks if both datasets are complementary, i.e., if the datasets are almost entirely different. In the end, our approach builds the CRIME dataset composed of the merging of the other two datasets. The attributes used as prediction features are presented in Table 1.

Table 1: Attributes used as prediction features.

Attribute	Description
Crime's type	Crime category, e.g., robbery and theft
Gender	Gender of the victim
Is Holiday	(1) if crime occurs in a holiday; (0) otherwise
Weekday	(1) if crime occurs in the weekend; (0) otherwise
Period	Period of the day the crime occurred
Region	Geographic location where the crime occurred

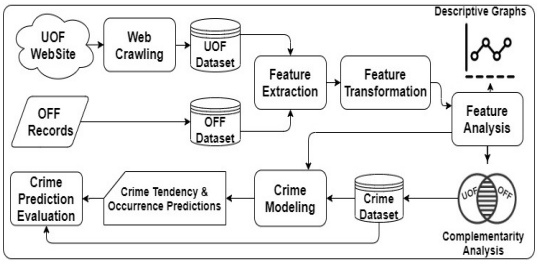

Figure 1: The workflow of our crime prediction.

In the fifth step, the prediction features are used to train supervised classifiers. The *Crime Modeling* component provides a classification model used to predict the tendency and occurrence of crimes by type of crime and geographic location. To tackle crime prediction as a classification problem, we follow the general framework of discriminative learning (Liu, 2009). Crimes reported in cities can vary significantly from month to month, and it is common for law-enforcing agencies to conduct crime analysis and comparison for months, and an interval of months. So, in particular, our goal is to learn an optimal hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$, mapping the input space $\mathcal{X}$ to the output space $\mathcal{Y}$. To this end, a plethora of learning algorithms could be deployed. In this work, we used k-NN, SVM, RF, and XGBoost. Finally, in the sixth step, the *Crime Prediction Evaluation* phase evaluates the effectiveness of the prediction models.

5 CRIME DATA SOURCES

We present the crime datasets for the criminal prediction, cover a complementarity analysis, and preprocessing procedures for effective supervised training.

5.1 The OFF Dataset

The OFF dataset has 520,378 records regarding January 2012 to November 2017, which is presented in

Figure 2 by over time and type of crime.

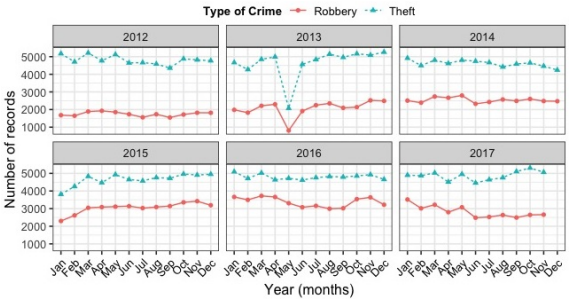

Figure 2: The composition of the OFF dataset.

From Figure 2, we can observe that theft records are the majority, accounting for more than 64% of the total number of records, and robbery has been increasing over the years, while thefts remain almost stable. In 2017, the robbery records dropped significantly. Additionally, more than 57% of victims are men, the crimes occur mostly on Mondays, their occurrences increase on holidays, one geographic region concentrates the most significant number of criminal records, and the criminals usually act in the afternoon and night.

5.2 UOF Dataset

The UOF dataset[2] has 6,529 regarding January 2012 to December 2017, which is presented in Figure 3 by over time and type of crime.

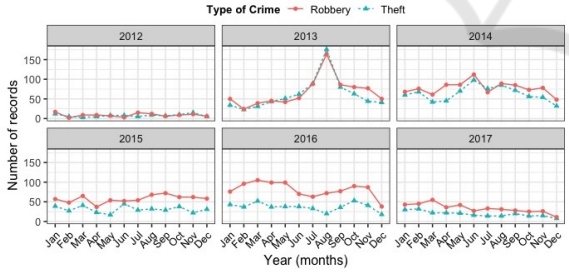

Figure 3: The composition of the OFF dataset.

The 2014 year concentrates the highest number of records, while the lowest number was in 2012. From Figure 3, we observe that similarly to the OFF dataset, robbery records are the majority, accounting for more than 59% of the total number of records, and robbery has been increasing over the years, while thefts remain almost stable, except for two peaks in August 2013 and June 2014. Also, more than 61% of victims are men, the crimes occur mostly on Wednesdays, their occurrences increase on holidays, one geographic region concentrates the most significant number of criminal records, and the criminals were usually acting at night.

[2]Available at *http://dx.doi.org/10.5281/zenodo.3673834*.

5.3 Complementarity Analysis

We provide a complementarity analysis of the crime datasets. In particular, we check how the datasets complement each other by checking if there is an intersection between their records. To perform this analysis, we use the following record attributes: latitude, longitude, date, and the period of the day, which of the crime occurred, victim's sex and, crime type.

The address entered in different systems, and the way the systems handle each address may result in the same criminal record with different latitudes and longitudes. To determine the existence of records referencing the same crime in the two datasets, we use the latitude and longitude of the criminal record with the precision of one geographical block.

We performed the complementarity analysis using the following attributes: date, the period of the day, victim's gender, type of crime, and latitude and longitude with the precision of one geographical block. The result of this analysis shows that the UOF and OFF datasets have only thirty-six common records, meaning they are significantly complementary to each other. Therefore, from the combination of the records of the two datasets, we obtained a combined dataset that we call the CRIME dataset. By combination, we mean merging the two datasets from the union of their records by removing the duplicate records.

5.4 Preprocessing Procedures

Supervised learning techniques often require input data transformation for effective training. To train each one of the learning techniques used in our crime prediction approach we perform the following data transformations in the three datasets: i) encode the day of the week attribute in one of the weekdays (Monday, Tuesday, Wednesday, Thursday and Friday) or weekend days (Saturday and Sunday); ii) encode the regions attribute according to the density of each geographic region provided by IBGE (Census 2010); iii) converts the attributes period, day of the week, type of crime and region into numerical data.

6 EXPERIMENTAL SETUP

To evaluate our approach, we run experiments aiming to answer the research questions: i) how effective are the supervised learning techniques to predict crime

occurrences, i.e., the number of occurrences of different types of crimes by geographic region? ii) how effective are the supervised learning techniques to predict crime tendency, i.e., a particular type of crime tends to increase or decrease by geographic region?

We use four different techniques to generate prediction models: *k*-NN, SVM, RF, and XGBoost. Besides, we report effectiveness in terms of precision and accuracy, i.e., the percentage of true positives for all positive predictions, and the hit rate or the number of instances correctly classified, respectively. To realize the experiments, we filter the crime dataset to get the year with records in all geographic regions, i.e., 2016 for the UOF dataset and 2017 for the OFF dataset. Furthermore, we use seven configurations for training and test sets for each dataset. In the case of the CRIME dataset, we replicate the settings of UOF and OFF datasets due to the difference of the year chosen in the others datasets. Besides, we set the hyper-parameters of each machine learning technique in each training and test configuration schema presented in Table 2.

For *k*-NN, we set the Manhattan distance. Also, we performed 10-fold cross-validation (Jain, 1991) to set the *k*-NN and RF hyper-parameters. The *k*-fold cross-validation divides the training set (resampling) into *k* subsets. From these subsets, the hold-out method is repeated *k* times, so that each time, one of the *k* subsets is used for the test, while the remainder (*k-1*) is used for model training. The hold-out method stores the accuracy metric at each iteration, returning the best parameters with the best accuracy. For XGBoost, we set the learning rate as $lr = 0.1$, and use grid-search method (Hsu et al., 2008) to set *stree* and *Gamma* hyper-parameters. Also, we use the grid-search method to set the *C* and *Gamma* parameters for SVM, and the $ntree$ and $mtry$ parameters for RF. The grid-search method obtains the best parameters set by combining a given range of parameters for each model. It stores the accuracy metric of each iteration, returning the best parameters with the best accuracy.

7 EXPERIMENTAL RESULTS

As we mentioned in Section 6, for each dataset, we evaluate different configuration schemes of training and test, reporting the accuracy and precision metrics, which the results are in Table 3. The outperforming results for each configuration are highlighted.

From Table 3, we observe that all learning techniques perform well to predict occurrence and the tendency in all datasets. Referring to the occurrence prediction, the *k*-NN and RF achieve up to 70% of accuracy in the UOF dataset, and the *k*-NN and XGBoost achieve up to 69% of accuracy in the OFF dataset. In the CRIME dataset, all techniques achieve up to 68% of accuracy in 2017, and the *k*-NN and the XGBoost perform better in 2016, achieving up to 64% of accuracy. Likewise, concerning the precision, the *k*-NN and SVM achieve up to 79% of precision in the UOF dataset, and the *k*-NN and XGBoost achieve up to 67% of precision in the OFF dataset. In the CRIME dataset, the XGBoost achieve up to 65% of precision in 2016 and, together with *k*-NN and SVM, up to 67% of precision in 2017.

About the tendency prediction, they achieve up to 84% of precision in the UOF dataset and up to 98% in the OFF dataset. In the CRIME dataset, they achieve up to 98% of precision (2017), and the *k*-NN technique achieves up to 92% of precision (2016). Likewise, concerning the accuracy, they achieve up to 78% in the UOF dataset and up to 89% in the OFF dataset. In the CRIME dataset, the *k*-NN, RF and XGBoost achieve up to 78% and 89% of tendency's accuracy in 2016 and 2017, respectively, outperforming the SVM approach in up to 23.61%.

In summary, for crime occurrence prediction, the *k*-NN usually performs better than the other three techniques in accuracy and precision, except for XGBoost precision in the *CRIME* dataset. Additionally, for crime tendency prediction, the *k*-NN sometimes outperforms the other three techniques in precision in UOF and OFF datasets, and in CRIME dataset, the *k*-NN and the RF mostly outperform the other two techniques. Moreover, the performance of all the techniques in the CRIME dataset was many times better and at least closer to the performance in the separate datasets, attesting that the use of heterogeneous sources of data can help to improve prediction performance. Recalling our research questions presented in Section 6, the previously observations attest the effectiveness of our prediction approach to predict crime occurrence and tendency.

The experimental results show that the four predictive techniques usually perform better in the OFF dataset, which is almost 80 times bigger than the UOF dataset. The same behavior is observed for the year 2017 of the CRIME dataset compared to the year 2016 of the same dataset. Thus, with more records for training, the supervised learning techniques can better adjust the learning bias, providing better predictions. Furthermore, we observe that the performance of the predictive approaches decreases as the ratio of the amount of training and test data decreases, showing some temporal and density relationship of training data with performance.

According to the results, our prediction approach

Table 2: Estimated hyper-parameters for our predictive models of crime.

	Configuration Schema		k-NN	SVM		RF		XGBoost	
	Training	Test	k	C	*Gamma*	*ntree*	*mtry*	*Gamma*	*stree*
UOF	Nov/2016	Dec/2016	6	1.00	0.07	2	100	0.50	3
	Sep - Nov/2016	Dec/2016	8	0.10	0.0005	2	7	0.30	4
	Jun - Nov/2016	Dec/2016	11	0.07	0.0005	17	8	0.50	2
	Jan - Nov/2016	Dec/2016	13	0.07	0.0005	2	50	0.50	2
	Jan - Sep/2016	Oct - Dec/2016	8	2.50	0.09	2	115	0.30	2
	Jan - Jun/2016	Jul - Dec/2016	29	1.00	0.0007	25	5	0.50	4
	Jan - Mar/2016	Apr - Dec/2016	16	2.50	0.10	15	25	0.30	3
OFF	Oct/2017	Nov/2017	146	9.30	0.01	2	5	0.50	2
	Aug - Oct/2017	Nov/2017	119	3.30	0.01	2	5	0.50	2
	May - Oct/2017	Nov/2017	70	0.30	0.30	2	40	0.50	4
	Jan - Oct/2017	Nov/2017	134	0.30	0.30	9	96	0.30	3
	Jan - Sep/2017	Oct - Nov/2017	12	35.00	0.005	2	7	0.50	3
	Jan - Jun/2017	Jul - Nov/2017	2	1.50	0.03	2	5	0.50	4
	Jan - Mar/2017	Apr - Nov/2017	2	4.30	0.03	13	73	0.30	2
CRIME	Nov/2016	Dec/2016	215	0.50	0.09	11	5	0.30	2
	Sep - Nov/2016	Dec/2016	118	0.10	0.0001	2	5	0.40	3
	Jun - Nov/2016	Dec/2016	6	0.10	0.20	55	9	0.50	3
	Jan - Nov/2016	Dec/2016	11	6.90	0.10	2	2	0.40	2
	Jan - Sep/2016	Oct - Dec/2016	383	6.20	0.10	11	5	0.50	2
	Jan - Jun/2016	Jul - Dec/2016	114	90.00	0.05	25	3	0.50	2
	Jan - Mar/2016	Apr - Dec/2016	50	50.00	0.03	2	5	0.30	2
	Oct/2017	Nov/2017	144	6.30	0.10	11	80	0.50	2
	Aug - Oct/2017	Nov/2017	110	4.70	0.07	2	98	0.30	2
	May - Oct/2017	Nov/2017	185	20.00	0.005	3	95	0.50	3
	Jan - Oct/2017	Nov/2017	193	17.00	0.05	70	5	0.30	2
	Jan - Sep/2017	Oct - Nov/2017	4	8.40	0.03	2	55	0.50	4
	Jan - Jun/2017	Jul - Nov/2017	453	68.00	0.01	23	4	0.50	4
	Jan - Mar/2017	Apr - Nov/2017	383	0.30	0.0007	15	30	0.30	3

Table 3: The effectiveness of the predictive approaches.

	Configuration Schema		Prediction's Accuracy				Prediction's Precision				Tendency's Accuracy				Tendency's Precision			
	Training	Test	k-NN	SVM	RF	XGB	k-NN	SVM	RF	XGB	k-NN	SVM	RF	XGB	k-NN	SVM	RF	XGB
UOF	Nov/2016	Dec/2016	0.66	0.68	**0.70**	0.66	0.57	0.46	**0.69**	0.61	**0.78**	**0.78**	**0.78**	**0.78**	**0.84**	**0.84**	**0.84**	**0.84**
	Sep - Nov/2016	Dec/2016	**0.68**	**0.68**	**0.68**	0.66	**0.63**	0.46	0.46	0.59	**0.78**	**0.78**	**0.78**	0.72	**0.84**	**0.84**	**0.84**	0.72
	Jun - Nov/2016	Dec/2016	**0.68**	**0.68**	**0.68**	**0.68**	**0.63**	0.46	0.46	0.46	**0.78**	**0.78**	**0.78**	**0.78**	**0.84**	**0.84**	**0.84**	**0.84**
	Jan - Nov/2016	Dec/2016	**0.70**	0.68	0.68	0.68	**0.79**	0.46	0.46	0.46	**0.78**	**0.78**	**0.78**	**0.78**	**0.84**	**0.84**	**0.84**	**0.84**
	Jan - Sep/2016	Oct - Dec/2016	**0.67**	**0.67**	0.64	0.64	0.65	**0.68**	0.57	0.58	**0.48**	0.44	0.46	0.45	0.34	0.44	**0.49**	**0.49**
	Jan - Jun/2016	Jul - Dec/2016	**0.69**	0.68	0.68	0.67	**0.69**	0.46	0.60	0.62	0.54	0.50	0.49	**0.57**	**0.61**	0.50	0.51	**0.61**
	Jan - Mar/2016	Apr - Dec/2016	**0.70**	0.69	0.69	0.67	0.66	**0.79**	0.47	0.59	**0.60**	0.53	0.52	0.56	**0.71**	0.57	0.54	0.65
OFF	Oct/2017	Nov/2017	**0.69**	0.68	0.68	0.68	**0.67**	0.66	0.66	0.66	**0.89**	0.78	0.78	**0.89**	**0.97**	0.96	0.96	**0.97**
	Aug - Oct/2017	Nov/2017	0.68	0.68	0.68	**0.69**	**0.67**	0.66	0.66	**0.67**	**0.89**	**0.89**	**0.89**	**0.89**	**0.97**	**0.97**	**0.97**	**0.97**
	May - Oct/2017	Nov/2017	**0.68**	**0.68**	**0.68**	**0.68**	0.66	0.66	0.66	**0.67**	0.78	**0.83**	**0.83**	**0.83**	**0.98**	**0.98**	**0.98**	**0.98**
	Jan - Oct/2017	Nov/2017	**0.69**	0.68	0.68	**0.69**	**0.67**	0.66	0.66	**0.67**	**0.83**	**0.83**	**0.83**	**0.83**	**0.98**	**0.98**	**0.98**	**0.98**
	Jan - Sep/2017	Oct - Nov/2017	**0.66**	**0.66**	**0.66**	**0.66**	0.63	**0.65**	0.43	0.43	0.64	0.61	0.61	**0.78**	**0.96**	0.95	0.95	0.51
	Jan - Jun/2017	Jul - Nov/2017	0.65	**0.67**	0.66	0.66	0.60	**0.63**	0.61	0.44	**0.69**	0.56	0.56	0.48	0.82	**0.84**	**0.84**	0.48
	Jan - Mar/2017	Apr - Nov/2017	**0.66**	**0.66**	**0.66**	**0.66**	0.60	**0.63**	0.44	0.44	**0.69**	0.61	0.61	0.58	**0.78**	**0.78**	**0.78**	0.52
CRIME	Nov/2016	Dec/2016	**0.62**	**0.62**	**0.62**	0.61	**0.62**	**0.62**	**0.62**	0.60	0.72	0.72	**0.78**	**0.78**	0.81	0.81	**0.83**	0.78
	Sep - Nov/2016	Dec/2016	0.59	**0.60**	0.59	0.59	0.56	0.59	**0.60**	0.56	**0.72**	0.67	**0.72**	0.67	**0.92**	0.82	0.77	0.82
	Jun - Nov/2016	Dec/2016	0.63	**0.64**	0.63	**0.64**	0.63	**0.65**	0.64	**0.65**	**0.78**	0.72	0.72	0.72	**0.84**	0.81	0.81	0.81
	Jan - Nov/2016	Dec/2016	**0.59**	**0.59**	**0.59**	**0.59**	**0.59**	**0.59**	**0.59**	0.58	**0.78**	0.67	0.56	0.67	**0.84**	0.81	0.70	0.81
	Jan - Sep/2016	Oct - Dec/2016	0.61	**0.62**	0.61	0.61	0.60	**0.62**	0.61	0.61	0.37	0.50	**0.56**	0.43	0.47	0.61	**0.64**	0.51
	Jan - Jun/2016	Jul - Dec/2016	**0.62**	**0.62**	**0.62**	**0.62**	0.61	**0.62**	**0.62**	0.61	0.48	0.56	**0.63**	0.56	0.61	0.59	**0.66**	0.65
	Jan - Mar/2016	Apr - Dec/2016	**0.63**	**0.63**	0.61	0.61	**0.63**	**0.63**	0.61	0.61	0.56	**0.58**	0.57	0.54	**0.72**	0.70	0.57	0.65
	Oct/2017	Nov/2017	**0.68**	**0.68**	**0.68**	**0.68**	**0.66**	**0.66**	**0.66**	**0.66**	**0.89**	0.84	**0.89**	**0.89**	**0.97**	0.96	**0.97**	**0.97**
	Aug - Oct/2017	Nov/2017	**0.68**	**0.68**	**0.68**	**0.68**	0.66	0.66	0.66	**0.67**	**0.83**	0.72	**0.83**	**0.83**	**0.98**	**0.98**	**0.98**	**0.98**
	May - Oct/2017	Nov/2017	**0.68**	**0.68**	**0.68**	**0.68**	0.66	0.66	**0.67**	**0.67**	**0.83**	0.67	**0.83**	**0.83**	**0.86**	0.85	**0.86**	**0.86**
	Jan - Oct/2017	Nov/2017	**0.68**	**0.68**	**0.68**	**0.68**	**0.67**	0.66	0.65	0.66	**0.78**	0.72	**0.78**	**0.78**	**0.86**	0.85	**0.86**	**0.86**
	Jan - Sep/2017	Oct - Nov/2017	0.65	0.65	**0.66**	0.65	0.62	0.62	**0.63**	0.62	**0.89**	0.72	0.81	0.75	0.94	0.97	**0.98**	**0.98**
	Jan - Jun/2017	Jul - Nov/2017	0.63	**0.64**	**0.64**	0.59	0.60	0.59	0.61	**0.65**	0.50	0.48	**0.65**	0.51	0.67	0.63	0.69	**0.71**
	Jan - Mar/2017	Apr - Nov/2017	**0.59**	**0.59**	0.54	0.48	0.59	0.59	**0.60**	0.55	0.47	0.47	**0.70**	0.65	0.67	0.67	**0.76**	**0.76**

can exploit supervised learning techniques and a set of spatial and temporal features extracted from heterogeneous sources of crime records to provide effective models for crime occurrence and tendency prediction.

8 CONCLUSIONS

We exploited supervised learning on heterogeneous data extracted form official and unofficial crime reports for crime prediction. Notably, we proposed a predictive approach that collects crime reports from the Web and used them as a source of features useful to predict the tendency of increase or decrease and the number of occurrences of types of crimes by geographic regions. Additionally, we thoroughly evaluated the techniques used by our approach, and the results of this evaluation showed that all methods perform well, with a slight advantage to k-NN and XGBoost in some cases. They provide an accuracy of up to 89% and a precision of up to 98% for crime tendency prediction and, for occurrences prediction, the accuracy of up to 70% and a precision of up to 79%.

Experimental results showed that there are quantitative, temporal, and density relationship between training data and performance, i.e., with more records for training, the supervised learning techniques can better adjust the learning bias, providing better predictions.

As strengths of this article, we highlight: i) we used heterogeneous data sources, official and unofficial crime records, to predict crimes; ii) we provided a complementarity analysis showing the feasibility of using different data sources by combining them into a single and bigger dataset; iii) we proposed a predictive approach capable of predicting the tendency and occurrence of different types of crimes in different geographic regions; iv) we evaluated four different machine learning techniques used by our crime prediction approach; v) the authorities can use our approach to plan crime preventive actions, such deciding where to perform street lighting and police patrol.

Contrasting the predictive techniques presented in this article and other recent crime prediction approaches reported in the literature, we observed that there is still room for further improvements. Therefore, for future work we plan to evaluate other learning techniques, such as latent factor models and neural networks for crime prediction. Also, we want to exploit different geographic properties of crimes as features within our predictive approach, and extend the datasets to cover more types of crimes. Finally, we also intent to use more sources of heterogeneous data, specially crime records with judgment data.

ACKNOWLEDGEMENTS

The present work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Níıvel Superior - Brazil (CAPES) - Financing Code 001. The authors thank CNPq, FAPEMIG, PUC Minas and SESP-MG (Secretaria de Estado de Segurança Pública de Minas Gerais) for the partial support in the execution of this work.

REFERENCES

Baculo, M. J. C., Marzan, C. S., de Dios Bulos, R., and Ruiz, C. (2017). Geospatial-temporal analysis and classification of criminal data in manila. In *International Conference on Computational Intelligence and Applications*, ICCIA'17, pages 6–11. IEEE.

Belesiotis, A., Papadakis, G., and Skoutas, D. (2018). Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems*, 3(4):12:1–12:31.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. In *International Conference on Multimodal Interaction*, ICMI'14, pages 427–434. ACM.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bühlmann, P. and Yu, B. (2010). Boosting. *Wiley Interdiscipl. Reviews: Computational Statistics*, 2(1):69–74.

Celisse, A. and Mary-Huard, T. (2018). Theoretical analysis of cross-validation for estimating the risk of the k-nearest neighbor classifier. *Journal of Machine Learning Research*, 18:1–54.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 785–794. ACM.

Constitution, B. (1988). Constitution of the Federative Republic of Brazil.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley-Interscience, 2ed edition.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Ho, T. K. (1995). Random decision forests. In *International Conference on Document Analysis and Recognition*, ICDAR'95, pages 278–282. IEEE.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2008). A practical guide to support vector classication. Technical report, National Taiwan University, Taipei, Taiwan.

Huang, C., Zhang, J., Zheng, Y., and Chawla, N. V. (2018). Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In *ACM International Conference on Information and Knowledge Management*, CIKM'18, pages 1423–1432. ACM.

Jain, R. (1991). *The art of computer systems performance analysis*. Wiley-Interscience.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324.

Lin, Y.-L., Yen, M.-F., and Yu, L.-C. (2018). Grid-based crime prediction using geographical features. *International Journal of Geo-Information*, 7(8).

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Oberwittler, D., Hummelsheim, D., Hirtenlehner, H., and Jackson, J. (2010). Social insecurities and fear of crime: A cross-national study on the impact of welfare state policies on crime-related anxieties. *European Sociological Review*, 27(3):327–345.

Pradhan, I., Potika, K., Eirinaki, M., and Potikas, P. (2019). Exploratory data analysis and crime prediction for smart cities. In *International Database Applications & Engineering Symposium*, IDEAS'19, pages 4:1–4:9. ACM.

Syarif, I., Prugel-Bennett, A., and Wills, G. (2016). Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *Telecommunication Computing Electronics and Control*, 14:1502–1509.

United Nations (2015). Universal declaration of human rights.