World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:15, No:2, 2021

An Application for Risk of Crime Prediction Using Machine Learning

Luis Fonseca, Filipe Cabral Pinto, Susana Sargento

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf

Abstract—The increase of the world population, especially in large urban centers, has resulted in new challenges particularly with the control and optimization of public safety. Thus, in the present work, a solution is proposed for the prediction of criminal occurrences in a city based on historical data of incidents and demographic information. The entire research and implementation will be presented start with the data collection from its original source, the treatment and transformations applied to them, choice and the evaluation and implementation of the Machine Learning model up to the application layer. Classification models will be implemented to predict criminal risk for a given time interval and location. Machine Learning algorithms such as Random Forest, Neural Networks, K-Nearest Neighbors and Logistic Regression will be used to predict occurrences, and their performance will be compared according to the data processing and transformation used. The results show that the use of Machine Learning techniques helps to anticipate criminal occurrences, which contributed to the reinforcement of public security. Finally, the models were implemented on a platform that will provide an API to enable other entities to make requests for predictions in real-time. An application will also be presented where it is possible to show criminal predictions visually.

Keywords—Crime prediction, machine learning, public safety, smart city.

I. Introduction

Public safety is one of the major concerns of the population around the world. Several factors have contributed to the increase of concern, such as the accelerated process of urbanization. It has been notorious in recent years the movement of people to cities and according to projections by the United Nations, in 2050 almost 70% of the population will live in urban areas [1]. Also, regarding the Global Terrorism Database, which defines ”acts of violence by non-state actors, perpetrated against civilian populations, intended to cause fear, in order to achieve a political objective”, the number of terrorist attacks that occurred during the last decade was the highest since they have been registered [2]. Machine learning techniques are crucial for applications involving smart cities and can be applied to crime prevention, since they assist in issues involving urban development which helps in extraction of value of the data retrieved [3].

This paper presents a solution that monitors and predicts criminal incidences in order to provide citizens and city authorities knowledge about the most dangerous areas; thus, adding value to the city for improving public safety. In this sense, this type of prediction can be useful in several ways, from more optimized and effective design of patrol routes as well as being useful for tourists who are unaware of the most dangerous areas of cities.

Moreover, once the data used is labeled, supervised models have been implemented, since all inputs and outputs of historic data are known but need to be predicted for other new instances. So, supervision learning has the purpose to find a general algorithm that maps the inputs to the outputs [4].

The prediction is performed through machine learning algorithms such as Random Forest, Neural Networks, K-Nearest Neighbors and Logistic Regression. These algorithms are assessed according to the data processing and transformation used. The results show that the use of Machine Learning techniques helps to anticipate criminal occurrences, which contributed to the reinforcement of public security. The models are implemented on a platform that will provide an API to enable other entities to make requests for predictions in real-time. Finally, an application is developed to show criminal predictions visually.

The remainder of this paper is organized as follows. Section II describes the related work, and Section III illustrates the study scenario. Then, the proposed approach is presented, starting with the data preparation in Section IV, and the machine learning algorithms in Section V. Finally, the platform deployment and application are depicted in Section VI, and the conclusions and future work are provided in Section VII.

II. Literature Review

Machine Learning is a field of Artificial Intelligence giving the ability for the machine to learn without explicit programming and aimed to solving more complex problems [5].

A. Researched Approaches

In this sense, some researches have been done with the aim of using machine learning techniques in order to promote public safety. Next, some investigated approaches will be presented.

1) Using Historical Data: One way that can help to combat crime is to use data from previous crimes with the aim to predict and prevent future incidences. An example of this was demonstrated by McClendon et al. in [6], where their research proved how effective and accurate machine learning algorithms can be at predicting violent crimes.

A comparative study was conducted between the violent crime patterns from two datasets with crime historical data. The focus of the research is towards analyzing the crime patterns of the four violent crime categories, which are

L. Fonseca* and F. Pinto are with the Altice Labs, Aveiro, Portugal (*e-mail: luismiguel.fonseca@ua.pt).

S. Sargento Doe is with Instituto de Telecomunicac¸ ˜oes, 3810-193 Aveiro, Portugal and University of Aveiro, 3810-193 Aveiro, Portugal.

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf

murders, rapes, robberies and assaults. The authors used the following algorithms: Linear Regression, Additive Regression, and Decision Stump algorithms using the same finite set of features. After the implementation of the algorithms, the result outputs five metrics that evaluate the effectiveness and efficiency of the algorithms: Correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and the root relative squared error. They observe the linear regression algorithm to be very effective and accurate in predicting the crime data based on the training set input for the three algorithms. They also stated that the relatively poor performance of the Decision Stump algorithm could be attributed to a certain factor of randomness in the various crimes and the associated features (exhibits a low correlation coefficient among the three algorithms). The branches of the decision trees are more rigid and give accurate results only if the test set follows the pattern modeled. On the other hand, the linear regression algorithm could handle randomness in the test samples to a certain extent.

2) Spatial-temporal Analysis of Crimes: Lin et al. in [7] proposed a data-driven method based on "broken windows" theory and spatial analysis to analyze crime data using machine learning algorithms and thus, predict emerging crime hot spots for additional police attention. Based on this theory, they design a model that predicts the incidence of drug-related crime in the following month based on the incidence of drug-related crime, fraud, assault, intimidation, auto theft, and burglary in the current month. In this way, it is possible to extend the model with its spatial-temporal characteristics. Let each grid, which splitting from the map is regarded as a sample, and accumulates samples in the same time scale to construct matrices. Each matrix represents different spatial-temporal status; then, the matrix can be used to train and test by machine learning algorithms. The goal was to predict crime hot spots for the following month, which are not in the same temporal environment. Empty grids are removed to prevent model performance degradation. For the different time scales, they design seven sets of accumulated data over 1, 3, 6, 9, 12, 15, and 18 months. Then, they prepared the data frames to train the models. Experiments were run using different algorithms (Deep Learning, Random Forest, and Naïve Bayes) to compare prediction results against the proposed method.

Thereby, they demonstrate a machine learning method designed to provide improved prediction of future crime hot spots, with results validated by actual crime data. It was concluded that the model tuned using Deep Learning provides the best performance and also stated that visualizations of predicted hot spots can assist patrol planning and improve crime prevention.

3) Using Dynamic Features: Rumi et al. explored how dynamic features can significantly improve crime prediction in [8]. Their motivation is based on the fact that many studies are only based on demographic data as regional characteristics and not exploring human mobility through social media.

The main challenge of their research is that dynamic information is very sparse compared to the relatively static information. To address this issue, it was developed a matrix factorization based approach to estimate the missing dynamic features across the city. The authors stated that with dynamic features, in addition to crime prediction, it is feasible make a prediction of the category of crime, such as, Theft, Unlawful Entry, Drug Offence, Traffic Related Offence, Fraud and Assault.

In addition to historical, demographic and geographic features, authors extract the dynamic features from check-ins of Foursquare users. Foursquare is a geosocial and micro-blogging network that allows the user to indicate where they are, and search for their contacts who are close to that location. So, a location with visitors from diverse backgrounds in a time interval is highly correlated with some types of crime event such as theft; then, the authors concluded that monitoring the fluctuation of visitor diversity at locations provides useful information to the crime event prediction. In the research, it was used real datasets in Brisbane and New York City, and it was performed the following algorithms SVM, Random Forest, Neural Network, and Logistic Regression integrated with an ensemble based learning framework for crime event prediction.

The prediction performance before and after adding the proposed dynamic features have been compared. The test results demonstrate that the improvement of prediction performance after adding dynamic features is considerable and statistically significant. With this approach, the authors claim to have performance improvements in terms of precision and recall between 2% to 16% depending on the category.

B. Approaches Used in This Work

This work is a fusion of the approach of McClendon et al. [6] given that historical data is used to predict future criminal events, as well as that of Lin et al. [7], since the spatial-temporal approach is also employed, considering it is aimed to make a prediction for each neighborhood for a given time interval.

III. Scenario

This section describes the scenario chosen and the dataset used in this work. The scenario fits in the public safety area, more properly in the crime prediction. It was used a dataset of incidents that occurred in San Francisco, which is one of the most known cities of the United States of America. In addition to being a densely populated city, San Francisco is a large financial center and popular tourist destination. Thus, this city faces many challenges in the most diverse verticals, one of which is public safety. Data of police incident reports about this city were used to build the scenario.

In this order of ideas, the scenario of the work presented aims to make a prediction of the crime risk given a location and a period of time. As can be seen in Fig. 1, the prediction will be categorical; that is, it will predict a class, so, the output will be a category: **Reduced**, **Moderated** and **High**.

An example of a user story that can be applied in this scenario is: as a tourist, I want to know the likelihood of a crime occurring tomorrow night in Tenderloin, to find out if it is safe to go there.

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf

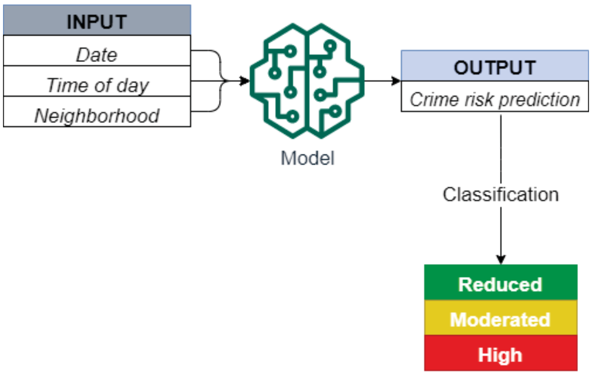


Fig. 1 Crime prediction scenario

A. Dataset

The dataset used is from a San Francisco open data platform that contains hundreds of datasets from the city and county of San Francisco [9], The dataset contains about 317 thousand rows and 29 columns, where each row is a police incident report from January 2018 to the present. These incidents can be filed by officers and by individuals through self-service online reporting for non-emergency cases.

There are some considerations to take into account on the current dataset, such as, an incident reported must be approved by a supervising officer. Once approved and electronically signed by a Sergeant or Lieutenant, no further information can be added to the initial report. A supplemental report will be generated if necessary for additional information or clarification. This means that an individual status will not change on an initial report but may be updated later through a supplemental report, which aims to provide additional incident information or to clarify a mistake in the initial report.

B. Data Understanding

Data understanding is an important step in a machine learning work, since it is here where it must be evaluated the available data and its alignment with the the objective. Data are the fuel of machine learning and, if it is weak, then, it is often difficult to achieve the goals initially proposed. Thus, each dataset column was explored in order to understand its meaning and some trends.

After analyzing all columns, the main conclusions drawn are:

- After checking the distribution of incidents by months, it was verified that there is no significant variance between the different months; however, it can be noted that in the winter months there is a slight decrease in incidents compared to the summer months, which may make sense since in the summer the city will receive more tourists and the population itself spends more time away from home, which may influence these numbers;
- Also it was seen that there is no significant variance between days of the week, although on Friday and Wednesday there is a slightly higher frequency of incidents in comparison with the other days;
- On the other hand, there is a higher variance of incidents with regard to the time of its occurrence, since there is a greater number of incidents during the solar day (from 8:00 until 21:00) compared to night (from 21:00 until 8:00);
- Regarding the geographical areas, it was noticed that the northeastern zone is the most dangerous zone, as it has a higher criminal density as opposed to the west and south zones which appear to be more peaceful.

IV. Data Preparation

In the vast majority of times, raw data can contain errors, which impairs the data quality and in turn will influence the results. This process usually takes a considerable part of the machine learning process, but it is crucial for removing faulty data and filling in gaps. So, data preparation has as a main goal to catch errors and inconsistencies with the aim to enhance data quality that can be processed and analyzed more quickly and efficiently [10]. Taking into account the described scenario, a set of operations were carried out in order to optimize the quality of the data.

The performance of the machine learning algorithms depends a lot on the treatment that was given to the data previously. For that reason, the pre-processing is an iterative process. It is important to note that many of the decisions taken in this section were the result of many attempts to improve the quality of the data.

The dataset consists of 50 different categories of incidents; however, not all of these categories are crimes, for example, categories like Traffic Collision, Fire Report and Missing Person, etc. Hence, all records whose category is not considered a crime have been removed. After removing the incidents that were not considered a crime, the dataset was left with 254171 rows.

A. Handling Times

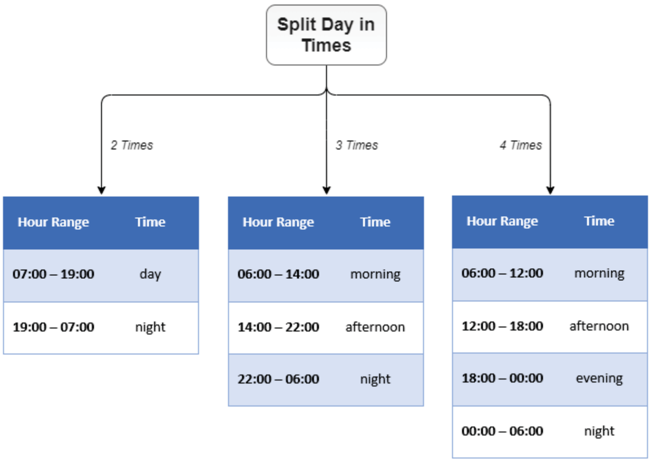
In the scenario described, the idea is to divide the day into parts. So, according to this requirement, a new column called *time_of_day* was created. Since the data pre-processing is not a straightforward step, three different approaches were followed to fill this column according to the population lifestyle and the daylight hours. In one approach the day was split into two equals parts, another into three equal parts and finally, into four equal parts. Fig. 2 shows the time interval in which each time fits.

B. Group Crimes by Neighborhood and Time of Day

As shown in Fig. 1, the location received as input will be the neighborhood as well as the time of the day. In this sense, a grouping of crimes was made taking into account the date, the neighborhood and time of day of samples.

Fig. 3 illustrates the final grouping, in which, the left side represents part of the original dataset and the right side are the results after grouping. The objective is to obtain a count of crimes that occurred on a certain date, in a certain neighborhood, at a certain time of the day. As can be seen, on

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf



Hour Range	Time
07:00 – 19:00	day
19:00 – 07:00	night

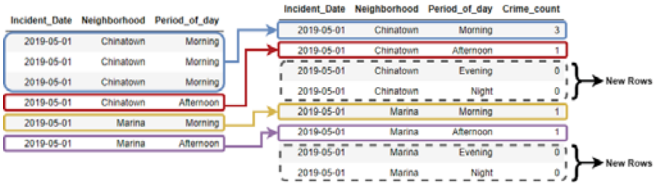
Hour Range	Time
06:00 – 14:00	morning
14:00 – 22:00	afternoon
22:00 – 06:00	night

Hour Range	Time
06:00 – 12:00	morning
12:00 – 18:00	afternoon
18:00 – 00:00	evening
00:00 – 06:00	night

Fig. 2 Splitting of day approaches

the 1 May 2019 (2019-05-01) in Chinatown, three crimes were committed in the morning; in this case, these three lines are grouped into one to create a new column called *Crime_count*, the number three is added to this column since there were three crimes. In Fig. 3 this operation is represent with a blue box.

When no crime occurs on a certain date, neighborhood and time of day then will be added a new row with *Crime_count* equals 0, as shown in Fig. 3 with a dashed line.



Incident_Date	Neighborhood	Period_of_day
2019-05-01	Chinatown	Morning
2019-05-01	Chinatown	Morning
2019-05-01	Chinatown	Morning
2019-05-01	Chinatown	Afternoon
2019-05-01	Marina	Morning
2019-05-01	Marina	Afternoon

Incident_Date	Neighborhood	Period_of_day	Crime_count
2019-05-01	Chinatown	Morning	3
2019-05-01	Chinatown	Afternoon	1
2019-05-01	Chinatown	Evening	0
2019-05-01	Chinatown	Night	0
2019-05-01	Marina	Morning	1
2019-05-01	Marina	Afternoon	1
2019-05-01	Marina	Evening	0
2019-05-01	Marina	Night	0

Fig. 3 Grouping crimes by neighborhood and time of day

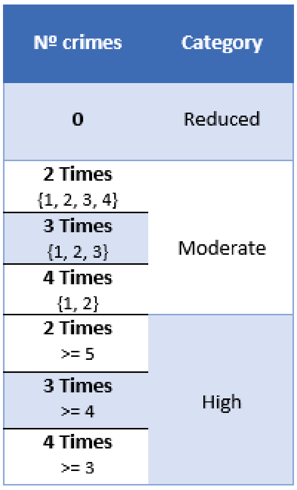
C. Risk Categorization

To apply classification algorithms, it was necessary to categorize the crime risk. To do this, a column called *Risk* was created that classifies the risk into three categories: *Reduced*, *Moderated* and *High*. The criteria for classification depends on the number of times that the day has been divided into. Considering that the day is divided into two times (day and night), each time will have a duration of 12 hours; while for days divided into four times (morning, afternoon, evening and night), each time has a duration of 6 hours.

For this reason, it was computed the mean of crimes per time for all approaches in the split. If the number of crimes is equal to zero, then, the category will be *Reduced* for all approaches. For the allocation of the remaining categories, the mean number of crimes per time in each approach was taken into account and has the following values:

- **2 times** (day and night): mean is 4
- **3 times** (morning, afternoon and night): mean is 3
- **4 times** (morning, afternoon, evening and night): mean is 2

Thus, the risk will be considered moderate if the number of crimes is less or equal to the mean for crimes. Fig. 4 demonstrates the two approaches to the risk categorization explained above.



Nº crimes	Category
0	Reduced
2 Times {1, 2, 3, 4}	Moderate
3 Times {1, 2, 3}	
4 Times {1, 2}	
2 Times >= 5	High
3 Times >= 4	
4 Times >= 3	

Fig. 4 Risk categorization according to the number of crimes

D. Merging Other Data Sources

In order to enrich the current dataset, a search was made for data that could be useful considering the scenario. However, it is a real challenge to find good data both in terms of quantity and quality. Since the location sent as an input will be the neighborhood, it was searched data that could add value to this attribute.

No information was found directly related to the neighborhood, but instead about census tracts. However, the neighborhoods were created by grouping 2010 Census tracts, using common real estate and resident definitions for the purpose of providing consistency in the analysis and reporting of socio-economic, demographic, and environmental data. So, through census tracts data, it is possible to infer data for a neighborhood by aggregating data from census tracts that belong to a particular neighborhood.

First of all, it is necessary to get the dataset that maps neighborhoods and census tracts. This dataset [11] contains all the census tracts and which neighborhood they belong to.

Then, it was sought for the median income [12], median age [13] and the total population [14] per census tract. All of these data have different sources, so there are three different datasets. In this sense, it was necessary to merge these three datasets to data with the mapping between neighborhood and census tract, resulting in a new dataset. This merging is highlighted in green in Fig. 5.

Nevertheless, the values of income, age and population are not grouped by neighborhood, but by census tract. For this reason, it was necessary to group these values. In this way, the income and age were aggregated by their mean per neighborhood, whereas in relation to the population a sum was applied. The next step is to merge this dataset which already aggregates four datasets (locations, income, age and

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf

populations) on the census tract, to the dataset that contains the crimes on the neighborhood column. This merging is highlighted in burgundy in Fig. 5.

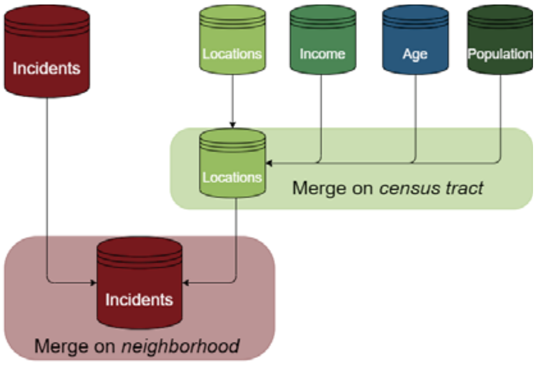


Fig. 5 Merge of dataframes

E. Relationship of Times of Day with Crimes Frequency

Bearing in mind that the times of the day is one of the key attributes to make the prediction, it is important to study in more detail the relation this feature with has the frequency of the crimes.

Since a day is divided into several periods, some differences are noted in the frequencies of the number of crimes per times. To have a better understanding of these differences, some visualizations were built for each type of division.

1) Day and Night: In this approach, the day is divided into two parts, each representing **12 hours**. In Fig. 6, each point represents one crime, and it is clear that there is a higher number of crimes recorded during the daytime period compared to night. Also, there is a higher dispersion in the count of crimes during the day. While at night the number of crimes is usually lower, during the day this number can grow to higher values. In short, we can notice that during a day there is a greater tendency to happen a crime.

2) Morning, Afternoon and Night: In this approach, the day is divided into three parts, each consisting of **8 hours**. Fig. 6 shows that the night period has the lowest mean crime rate, while the morning and afternoon have mean crime rates of 2.6 and 3.6, respectively. It can be seen that there is a larger number distribution of crimes during the morning or afternoon, while during the night the number of crimes is usually lower.

3) Morning, Afternoon, Evening and Night: In this approach, the day is divided into four parts, with each part consisting of **6 hours**. In Fig. 6 it can be noted that, once again, night and morning have the lowest crime rates, that is, less than 2 in both cases, whereas during the evening and afternoon the mean is equal to or greater than 2.5. It is interesting to note that, although the average crime rate is lower in the morning compared to evening, there is a greater variance in relation to the number of crimes.

V. Modeling

In this section, the distinct approaches presented in the previous section will be tested with the different algorithms. Fig. 7 shows all combinations between different approaches and different algorithms. So, the main objective is to compare the results of the different algorithms for the different approaches.

There are several categorical variables. For this reason, it is necessary to convert them to numeric form. For that, these variables were converted variable into ”dummy” variables where it was created a binary column (dummy column) for each category and assigns 1 to the column if the feature belongs or 0 otherwise.

In relation to the numerical columns, tests were also performed with and without the scaling of the features. For example, the minimum and maximum values of the $mean\_age$ column are 25 and 65 years, respectively, which makes a range of 40 years. On the other hand, the minimum and maximum values of $mean\_income$ are 16016 and 195375, respectively, which makes a range of 179359 dollars. As a result, the values of these two columns are on completely different scales. The aim is to prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges [15]. To perform the scaling, it was used the $MinMaxScaler$ [16] from $scikit-learn$. This estimator scales and translates each feature individually between zero and one. For each value in a feature, it will subtract the minimum value in the feature and then divides it by the range; and is presented in (1). $MinMaxScaler$ maintains the shape of the original distribution. It does not meaningfully change the information embedded in the original data and outliers are preserved.

$$x' = \frac{x - xMin}{xMax - xMin} \tag{1}$$

In some tested approaches, the data were relatively unbalanced. To address this issue there are two techniques: **undersampling**: this technique randomly removes samples from the majority class, with or without replacement; and **oversampling**: based on the samples already present in the data set, this technique creates new samples from the minority class. The oversampling technique that presented the best results was SMOTE (Synthetic Minority Over-sampling TEchnique) [17].

A. Defining Model Evaluation Rules

Model evaluation metrics are required to quantify model performance, To evaluate the model, it is necessary to divide the data into training and testing. The classic approach to do a simple division of data, where 75% is used for training and 25% for testing.

However, in order to provide ample data for training the model and to leave ample data for validation, it was used $KFoldcrossvalidation$ [18]. With this method, the data are divided into k subsets (folds). The holdout method is repeated k times, such that each time, one of the k folds is used as the test data and $k - 1$ of the folds as training data. This significantly reduces bias and also significantly reduces variance, as most of the data are also being used in the validation set. Interchanging the training and test sets also adds to the effectiveness of this method. The data was divided into

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf

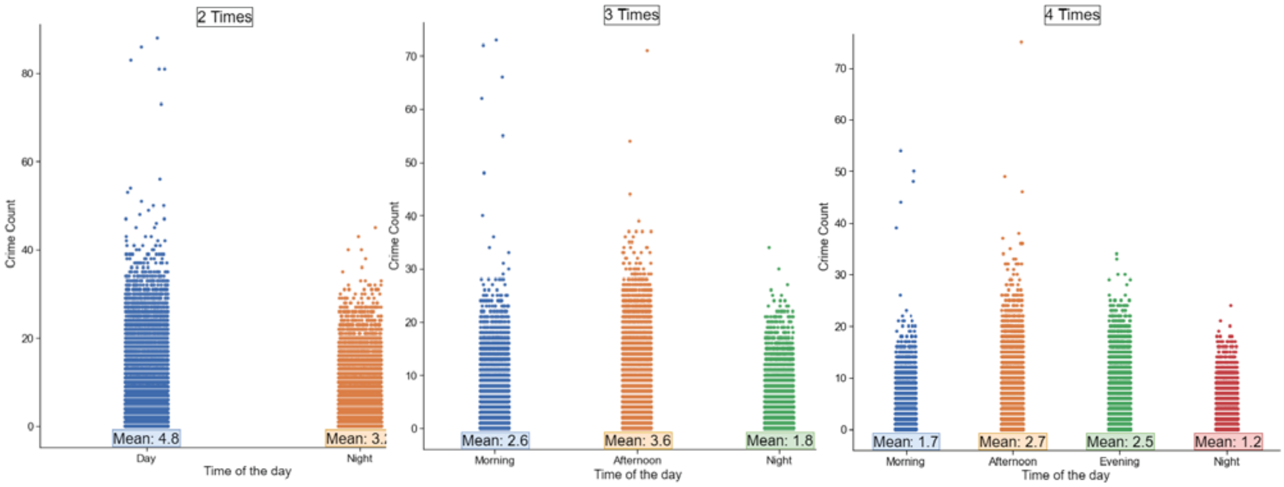


Fig. 6 Comparison of crime counts between between periods

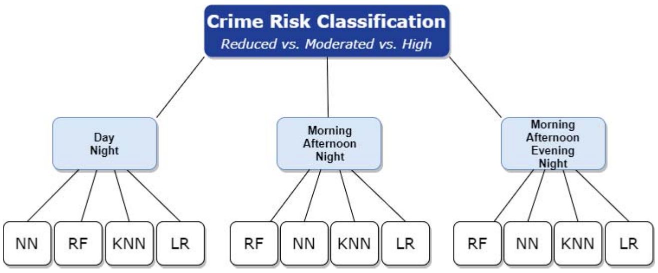


Fig. 7 Combining the different approaches with the various algorithms

four folds, which means that each fold contains 25% of data. The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop; in this case, with four iterations.

The following evaluation metrics were used:

1) Precision: Out of all the positive classes that have been predicted correctly, the precision gives how many are actually positive. Precision is a good measure to determine when the costs of False Positive is high. The formula of precision is presented in (2).

$$precision = \frac{TP}{TP + FP} \tag{2}$$

2) Recall: Out of all the positive classes, the recall gives how much has been predicted correctly. Recall calculates how many of the actual positives the model captures through labeling it as Positive (True Positive). The formula of recall is presented in (3).

$$recall = \frac{TP}{TP + FN} \tag{3}$$

3) F1 Score: Helps to measure *Recall* and *Precision* at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more. The formula for the F1 Score is presented in (4).

$$F1Score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

The results of each metric of each fold will be the average weighted by support (the number of true instances for each label). The final results will be the mean of the results of the folds.

B. Multiclass Classification - Reduced vs. Moderated vs. High

Multiclass classification is the task of classifying the elements of a given dataset into three or more groups. As already stated, these groups are *Reduced*, *Moderated* and *High*. All algorithms will be executed for each of the approaches according to each division in a day. The results will be presented in tables to analyze.

C. Day and Night

With this approach, as depicted in Table I, the most frequent class is Moderated with 30036 rows (47%), followed by High with 17764 rows (28%) and finally, Reduced with 15668 rows (25%).

Table I presents the results of the four algorithms executed for the *day and night* approach. The results are quite similar across all algorithms. The increase of performance is notorious when data have been scaled on the Neural Network, about 11% of improvement for all metrics. In the remaining algorithms, there were no significant differences.

D. Morning, Afternoon and Night

In this day division, depicted in Table II, the most frequent class is also Moderated with 38709 rows (41%), but this time, followed by class Reduce with 33181 rows (25%), and finally High with 23312 rows (25%).

Table II presents all results of four algorithms executed for *morning, afternoon and night* approach. Once again, the results are quite similar across all algorithms and the performance of Neural Network had about 15% of improvement with scaling. In the remaining algorithms, there were no significant differences.

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf

TABLE I
Results for Day and Night Approach

	Without Scaling								With Scaling							
	Unbalanced				Banlaced				Unbalanced				Banlaced			
	RF	NN	KNN	LR	RF	NN	KNN	LR	RF	NN	KNN	LR	RF	NN	KNN	LR
Precision	74%	67%	74%	74%	73%	65%	73%	73%	74%	74%	74%	74%	73%	73%	73%	73%
Recall	72%	63%	72%	72%	72%	63%	71%	71%	72%	73%	72%	72%	72%	75%	71%	71%
F1-Score	72%	64%	72%	71%	72%	64%	72%	72%	72%	72%	72%	71%	72%	72%	72%	72%
Time (s)	1.42	4.25	0.32	19.85	2.90	4.25	0.62	31.45	1.50	7.20	1.83	4.06	2.50	11.80	11.80	6.09

TABLE II
Results for Morning, Afternoon and Night Approach

	Without Scaling								With Scaling							
	Unbalanced				Banlaced				Unbalanced				Banlaced			
	RF	NN	KNN	LR	RF	NN	KNN	LR	RF	NN	KNN	LR	RF	NN	KNN	LR
Precision	71%	63%	72%	72%	70%	60%	71%	70%	71%	71%	72%	72%	70%	71%	71%	70%
Recall	69%	60%	71%	69%	69%	58%	71%	68%	69%	69%	71%	69%	69%	69%	68%	69%
F1-Score	69%	60%	69%	69%	70%	60%	69%	69%	69%	69%	69%	69%	70%	69%	69%	69%
Time (s)	2.44	4.61	0.63	31.00	3.34	4.81	0.84	36.88	2.59	12.27	3.74	7.01	3.56	16.10	6.74	8.49

TABLE III
Results for Morning, Afternoon, Evening and Night Approach

	Without Scaling								With Scaling							
	Unbalanced				Banlaced				Unbalanced				Banlaced			
	RF	NN	KNN	LR	RF	NN	KNN	LR	RF	NN	KNN	LR	RF	NN	KNN	LR
Precision	67%	56%	66%	66%	67%	53%	66%	66%	66%	66%	65%	66%	66%	66%	66%	66%
Recall	67%	55%	66%	67%	66%	53%	65%	66%	66%	66%	65%	66%	66%	66%	65%	66%
F1-Score	67%	54%	66%	66%	66%	52%	66%	66%	66%	66%	65%	65%	66%	65%	66%	66%
Time (s)	3.60	6.26	1.08	45.10	5.05	6.34	1.51	57.10	3.55	20.09	8.20	9.41	5.76	28.50	14.00	12.87

E. Morning, Afternoon, Evening and Night

In this 4-division day, depicted in Table III, the most frequent class is also Reduced with 53769 rows (42%), followed by class Moderated with 40054 rows (32%), and finally High with 33113 rows (26%).

Table III presents all results of the four algorithms executed for the *morning, afternoon, evening and night* approach. As with the two previous approaches, the neural network improves the results with scaling and in the remaining algorithms, there were no significant differences.

F. Chosen Model and Results Discussion

After examining the results, the approach chosen was *morning, afternoon, evening and night*. Although the results for this approach have shown some decrease in the performance compared to the other approaches, it has a great advantage, which is the increase of detail for the period of time chosen, which in turn will be more useful for the final user. As stated by Rumi et al. in [8], crime prediction in finer temporal grain will help the police to design their patrol strategy dynamically, and increase the probability to reduce the crime rate more effectively.

Although all algorithms have demonstrated similar performances, random forest algorithm was chosen, because it performed slightly better. Overall, in the selected model, data will not be scaled since the random forest is not sensitive to feature scaling. Also, the data will be not balanced since in the chosen *morning, afternoon, evening and night* approach, all classes are reasonably balanced and with the balancing, there were no differences in performances as noted in Table III.

Considering the frequency of each crime risk as a baseline, a comparison can be made with the results obtained by the chosen model and thus, verify the improvement achieved. For example, 42% of records belong to Reduced class, then, if the model always predicts Reduced, the precision will be 42%, since out of all the predicted classes, the Reduced is just 42% actually positive. Meanwhile, the model had a precision of 69% regarding the Reduced class, so, it was achieved an improvement of 65% over the baseline. As can be seen in Table IV, considering all categories with the machine learning model, an improvement of 110% was obtained.

It is also important to note that the High category contained the most marked improvement at about 178%. From the user's point of view, this can be very appropriate as the main concern is to anticipate locations with a high criminal risk.

TABLE IV
Improvements Achieved with the ML Model Compared to the Baseline

Risk	Baseline	Precision	ML Model Improvement
Reduced	42%	69%	65%
Moderated	32%	60%	88%
High	26%	72%	178%
Overall	33%	67%	110%

It was also examined the importance of each feature in the construction of random forest, and it was noted that the population and income were among the most important features as well as the time of the day. On the other hand, the features related to the day of the week proved to be less important.

VI. Deployment and Application

This section describes the last step in a machine learning project, the model deployment. Thus, the model is applied in a system where several entities can interact with it.

After finding the best model that meets the initial requirements, it is necessary to deploy it. First, it is necessary to save the model in a binary object on disk; then, it is possible to load and use the model in an application.

A back-end application that contains the model was developed that also retrieves daily incidents of the San Francisco Open Data, through the REST API provided, and stores it in a big data warehouse. This application also provides an API that allows other entities to send a request and obtain a prediction. The high-level architecture of the service logic explained is shown in Fig. 8.

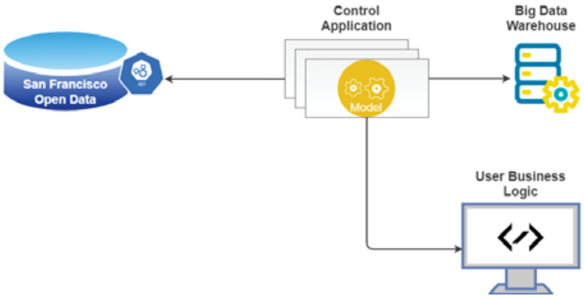


Fig. 8 High-level architecture of service logic

A. Client Application

In order to have a graphical interface in which any user could easily consult the crime risk prediction for San Francisco, a front-end application was integrated that makes requests to the API provided by the back-end application and shows the results iteratively on a map.

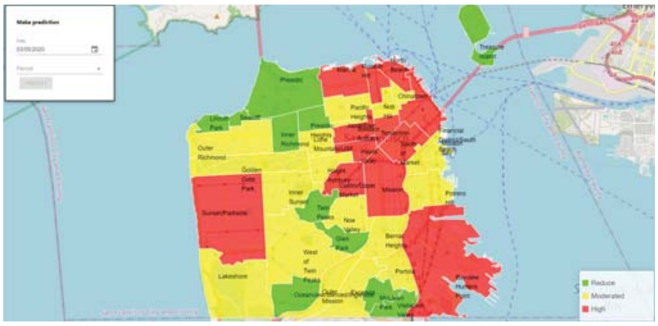


Fig. 9 Front-end application that shows the prediction for San Francisco on the map

When loading the page shown in Fig. 9, the Client application fetches the current prediction from the back-end and shows it through the Map. To give an overview of the panorama, each neighborhood is given a color according to the risk of crime for the current moment. As can be seen on the label in the lower right corner, if the crime risk is Reduced, the assigned color is **green**; if the crime risk is Moderate, the assigned color is **yellow**; and if it is High, the assigned color is **red**.

To select another date to make a prediction, the user can use the widget in the upper left corner that allows to choose the data and time of the day. The user cannot choose a date greater than a week, since the model will be retrained weekly with the most recent data. If the user chooses a date before today, it will return the real result instead of prediction.

VII. Conclusion

This paper proposes a solution based on machine learning techniques to predict crime events, which can be a great contribution to the improvement of public safety in a city, since it is a major concern in cities around the world.

It was interesting to note how the pre-processing and transformation can influence the results of the model, namely, in the dividing the day into several time periods.

Bearing in mind that a multiclass classification task was performed, and since this type of task is more demanding than a binary classification, then taking into account the results presented in Table IV, it can be said that good results have been achieved. However, it is expected that the results can still be improved by adding more variety of data, especially dynamic data.

This solution is developed for the city of San Francisco, due to the origin of the data. However, the solution can be adapted to other cities if similar data are made available.

With respect to future work, it would be interesting to test other types of algorithms, give greater importance to the most recent crime events, that is, events that occurred in the last month will have a more significant weight in training the model, and extend this solution to other cities to assess if criminal trends change depending on the city.

References

[1] U. Nations, “68% of the world population projected to live in urban areas by 2050, says un — un desa — united nations department of economic and social affairs,” https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html, May 2018, (Accessed on 26/10/2019).

[2] G. T. Database, “Incidents over time,” https://www.start.umd.edu/gtd/, 12 2018, (Accessed on 04/05/2020).

[3] Y. Wu, W. Zhang, J. Shen, Z. Mo, and Y. Peng, “Smart city with Chinese characteristics against the background of big data: Idea, action and risk,” *Journal of Cleaner Production*, vol. 173, pp. 60–66, 2018. [Online]. Available: https://doi.org/10.1016/j.jclepro.2017.01.047

[4] M. Mohammadi and A. Al-Fuqaha, “Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges,” *IEEE Communications Magazine*, vol. 56, no. 2, pp. 94–101, 2018.

[5] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, “Machine learning for internet of things data analysis: a survey,” *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, 2018.

[6] L. McClendon and N. Meghanathan, “Using Machine Learning Algorithms to Analyze Crime Data,” *Machine Learning and Applications: An International Journal*, vol. 2, no. 1, pp. 1–12, 2015.

[7] Y. L. Lin, T. Y. Chen, and L. C. Yu, “Using Machine Learning to Assist Crime Prevention,” *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*, pp. 1029–1030, 2017.

[8] S. K. Rumi, K. Deng, and F. D. Salim, “Crime event prediction with dynamic features,” *EPJ Data Science*, vol. 7, no. 1, 2018. [Online]. Available: http://dx.doi.org/10.1140/epjds/s13688-018-0171-7

[9] DataSF, “Datasf — office of the chief data officer — city and county of san francisco,” https://datasf.org/, 10 2020, (Accessed on 10/12/2020).

[10] M. R. Berthold and K. P. Huber, “MISSING VALUES AND LEARNING OF FUZZY RULES,” vol. 6, no. 1998, pp. 171–178, 1998.

[11] DataSF, “Analysis neighborhoods - 2010 census tracts assigned to neighborhoods — datasf — city and county of san francisco,” https://data.sfgov.org/Geographic-Locations-and-Boundaries/Analysis-Neighborhoods-2010-census-tracts-assigned/bwbp-wk3r/, 10 2020, (Accessed on 10/12/2020).

[12] U. S. Census, “https://data.census.gov/cedsci/table?q=san francisco income,” https://data.census.gov/cedsci/table?q=san%20francisco%20income, 10 2020, (Accessed on 10/12/2020).

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf

[13] ——, “https://data.census.gov/cedsci/table?q=san francisco age&tid=acsst1y2019.s0101,” https://data.census.gov/cedsci/table?q=san%20francisco%20age&tid=ACSST1Y2019.S0101, 10 2020, (Accessed on 10/12/2020).

[14] ——, “https://data.census.gov/cedsci/table?q=san francisco population&tid=acsdp1y2019.dp05,” https://data.census.gov/cedsci/table?q=san%20francisco%20population&tid=ACSDP1Y2019.DP05, 10 2020, (Accessed on 10/12/2020).

[15] L. A. Shalabi, R. Mahmod, A. Azim, A. Ghani, and Y. M. Saman, “A New Model for Extracting a Classifactory Knowledge from Large Datasets Using Rough Set Approach A New Model For Extracting A Classifactory Knowledge From Large Datasets Using Rough Set Approach,” no. January 1999, 1999.

[16] S. Learn, “6.3. preprocessing data — scikit-learn 0.23.2 documentation,” https://scikit-learn.org/stable/modules/preprocessing.html, 10 2020, (Accessed on 10/12/2020).

[17] imbalanced-learn API, “imbalanced-learn api — imbalanced-learn 0.5.0 documentation,” https://imbalanced-learn.readthedocs.io/en/stable/api.html, 10 2020, (Accessed on 10/12/2020).

[18] S. Learn, “3.1. cross-validation: evaluating estimator performance — scikit-learn 0.23.2 documentation,” https://scikit-learn.org/stable/modules/cross_validation.html, 10 2020, (Accessed on 10/12/2020).

Open Science Index, Computer and Systems Engineering Vol:15, No:2, 2021 publications.waset.org/10011881.pdf