Article

Applied Machine Learning in Social Sciences: Neural Networks and Crime Prediction

Ricardo Francisco Reier Forradellas *, Sergio Luis Náñez Alonso [ID], Javier Jorge-Vazquez [ID] and Marcela Laura Rodriguez

Department of Economics-DEKIS Research Group, Catholic University of Ávila, 05005 Ávila, Spain; sergio.nanez@ucavila.es (S.L.N.A.); jjorge.vazquez@ucavila.es (J.J.-V.); mrodriguez.pss@gmail.com (M.L.R.)
* Correspondence: ricardo.reier@ucavila.es

Abstract: This study proposes a crime prediction model according to communes (areas or districts in which the city of Buenos Aires is divided). For this, the Python programming language is used, due to its versatility and wide availability of libraries oriented to Machine Learning. The crimes reported (period 2016–2019) that occurred in the city of Buenos Aires selected to test the model are: homicides, theft, injuries, and robberies. With this, it is possible to generate a crime prediction model according to the city area based on the SEMMA (Sample, Explore, Modify, Model, and Assess) model and after data manipulation, standardization and cleaning; clustering is performed using K-means and subsequently the neural network is generated. For prediction, it is necessary to provide the model with the information corresponding to the predictive characteristics (predict); these characteristics being according to the developed neural network model: year, month, day, time zone, commune, and type of crime.

Keywords: artificial intelligence; crime prediction; prediction model; digital administration; crime; neural networks

1. Introduction

The Autonomous City of Buenos Aires, the capital of the Argentine Republic, has a relevant crime rate (Mancini 2020); however, the same is generally true of the large metropolises throughout Latin America (Kliksberg 2002; Sain 2012; Lozano et al. 2018); they also have a relevant crime rate. This rate of crime has increased throughout the country, mainly due to the socio-economic crisis that has hit it (Cid Ferreira et al. 2017; La Ruffa 2019). In this context, the government of the City of Buenos Aires has among its main objectives to implement various preventive measures so that the crime rate will decrease, working in fields such as:

-   Increase in police officers on the streets.
-   Broad patrol schemes.
-   Increase in the fleet of police mobiles.
-   Implementation of the digital ring for the identification of vehicles entering the city.
-   Expansion of the video surveillance system.
-   Incorporation of facial recognition systems in the city's surveillance cameras.

Nowadays it is an irrefutable fact that digitalization is changing the traditional methods and balances of the current economic and social organization model (Jorge-Vázquez et al. 2019). The implementations of these measures represent a significant percentage of the city's annual budget, which is around 15%. Therefore, it is necessary to carry out an exhaustive analysis of the data on the crimes that have occurred, so that the lines of preventive action are directed towards the most affected areas (Alegre 2019). In this way, based on the open data published in the portal of the Autonomous City of Buenos Aires corresponding to the years 2016 to 2019 (Ministerio de Justicia y Seguridad. Policía de la

Ciudad 2020), in the present article, the same have been analyzed using different tools of the Machine Learning area, with the aim of developing a predictive model of crimes. The application of different Machine Learning tools that allow the analysis of the data of crimes occurred in the City of Buenos Aires between the years 2016 and 2019 inclusive will serve as a feed for the model. To do so, it is based on the available information on reported crimes, which occurred in the area of the City of Buenos Aires, and are classified into four types: Homicide, Theft (without violence), Injury and Robbery (with violence). The main result has been to generate a model of crime prediction according to communes (areas or districts into which the city territory is divided) based on neural networks. At a technical level, the Python programming language will be used, due to its versatility and wide availability of libraries oriented to Machine Learning. The use of artificial intelligence (AI) at the service of public administrations has already been presented as an opportunity (Cerrillo i Martínez 2019), and within the field of predictive AI related to crime some work has already been developed (Batarseh and Yang 2017). In this sense, on the advantages of applying AI systems, (Corvalán 2018) highlights the role of the Prometea system and its service to justice in Latin America. However, the purpose of this study, the prediction of crimes based on an applied model of AI and machine learning, has already been tested previously. We can mention, among others, the work of (Ordoñez-Eraso et al. 2019) in which they manage to detect homicide trends in Colombia, or also (Guanin et al. 2015) in their study on classification algorithms for the prediction of crimes affecting citizen security in Guayaquil. Additionally, (García-Plúa et al. 2017) who are conducting a study on the same city developing a method for automatic detection based on hot spots in the city. The use of a neuronal network to predict crimes has been analyzed in other works such as those developed by (Torres et al. 2015) focused on geographical events or the one carried out by (Esquivel et al. 2020) whose study is focused on the city of Baltimore and the use of neuronal networks as well. In the field of predicting economic crimes there are also several studies. In particular, in the area of loans, the study by (Cheng et al. 2020) stands out; on crimes associated with the theft of cell phones there are the works by (Delahoz-Dominguez et al. 2020; Rodas 2019), among others. There are also studies on the comparison of systems such as (Cinelli and Manrique 2019), at the European level. There are also studies that link these new technologies such as CBDC, whose advantage of being implemented (among others) would be the reduction of economic crimes (Náñez Alonso 2019; Náñez Alonso et al. 2020a, 2020b).

The structure of the article starts from the introduction and case studies of the use of neuronal networks for the prediction of crimes, a section of materials and methodology where the Sample, Explore, Modify, Model, and Assess (SEMMA) model is used to elaborate the predictive model of crimes based on neuronal networks, a section of results and finally the discussion of results and conclusions.

2. Materials and Methodology

There are different methodologies that are generally referenced within the Data Mining, of which the most widely adopted is SEMMA, an acronym corresponding to the five basic phases of the process (Sample, Explore, Modify, Model, Assess) (Rodríguez et al. 2003; Gómez-Palacios et al. 2017). The SAS Institute, developer of this methodology, defines it as "the process of selecting, exploring and modeling large amounts of data to discover unknown business patterns" (SAS: Analytics, Artificial Intelligence and Data Management 2020). However, this same method has been subsequently analyzed by different authors for its practical application in different sectors (Abdul Rahman et al. 2019) or (Arumugam et al. 2020). The following steps are shown in Figure 1.

Figure 1. SEMMA (Sample, Explore, Modify, Model, and Assess) Methodology. Source: Own elaboration from SAS (SAS: Analytics, Artificial Intelligence and Data Management 2020).

1. Sample: The objective of this phase is to select a representative sample of the problem under study. The SEMMA methodology establishes that for each sample considered for the analysis of the process, the level of confidence of the sample must be associated.
2. Explore: Once a sample has been determined, the available information must be explored, with a view to simplifying and subsequently optimizing the model to be created. Visualization tools or statistical techniques are usually used to help reveal relationships between variables, and then facilitate the choice of these.
3. Modify: This phase consists of the manipulation of data to achieve the appropriate format for feeding the model to be built.
4. Model: The objective of this phase is to establish a relationship between the explanatory variables and the variables under study, which will make it possible to infer the value of these with a given level of confidence. The techniques used include traditional statistical methods (such as discriminant analysis, clustering methods and regression analysis), as well as data-based techniques such as neural networks, decision trees, etc.
5. Assess: consists of the evaluation of the results through the analysis of the goodness of the model or models, contrasted with other statistical methods or with new sample populations.
6. The choice of methodology is based mainly on the simplicity and dynamism of its application, considering SEMMA as the most appropriate. The predictive model of crimes presents a low complexity, characterized mainly by (SAS: Analytics, Artificial Intelligence and Data Management 2020):

    -   It contemplates the use of public data that has a structure that requires little modification for its use.
    -   It does not require the participation or advice of specialists for the generation of an acceptable predictive model.
    -   It is possible to evaluate the results of the model with basic knowledge.
    -   Interdisciplinary work is not required for the project, therefore, neither are efforts made to coordinate and standardize work processes.

2.1. Sample

Based on the working problem, in this case, the generation of a predictive model of crimes in the City of Buenos Aires, the objective in this phase consists of obtaining a representative sample of the universe of crimes that occurred in the defined geographical area, which allows the inference of the behavior of the population. The Autonomous City of Buenos Aires (CABA) is organized in 15 communes (Ministerio de Justicia y Seguridad. Policía de la Ciudad 2020). The concept of crimes is very broad, since it can include everything from drug trafficking to money laundering. However, when it comes to restricting it to a geographical area and associating it with preventive measures that can be taken, these must be within its area of competence. In the Argentine Republic, as in most countries, there are also different areas of management and responsibility, and those that can be prevented, investigated, and prosecuted by the authorities dependent on the City of Buenos Aires are reduced to the following:

-   Homicide: Crime consisting of killing someone without the circumstances of malice, price or overkill. (Rodríguez Collao 2011)

- Theft: Crime consisting in taking for profit other people's personal belongings against the owner's will, without the circumstances that characterize the crime of theft. (Olave Albertini 2018)
- Theft: A crime committed by taking possession of another's property for profit, using violence or intimidation against people, or force on things. (Ramos Pérez 2015).

The Autonomous City of Buenos Aires, as part of its policy of transparency, and following the global trend that exists in the publication of data of facts and management made, presents a portal of open data with 380 datasets, (Ministerio de Justicia y Seguridad. Policía de la Ciudad 2020) organized under different themes. In this portal, the crimes are under the category "Security". The portal indicates that information is available from 2016 to 2019, with the same registration structure, on four types of crimes and with a level of detail at the daily and commune levels. In accordance with the preceding sections, the required data sample should include

- Classification of crimes: including homicide, theft, and robbery.
- Geographic area: corresponding to the Autonomous City of Buenos Aires, detailed by commune.
- Timing: that has a level of detail that allows conclusions to be drawn in an appropriate manner, for example, at the daily level, as opposed to annual totals that would make it difficult to properly exploit the data.

In this way, the crime datasets published in the website of the City of Buenos Aires, corresponding to the years 2016 to 2019, will be considered. Although it is understood that the number of actual crimes is highly superior, mainly with regard to thefts and minor aspects, the information contained in the datasets corresponds to those that have been reported, constituting an appropriate sample for the work to be carried out. In principle, this source is considered sufficient for the generation of the predictive model. The data imported separately need to be consolidated in order to carry out their joint treatment. This process may present different challenges, which must be properly managed in order to obtain quality data. In the case of analysis, no problems have arisen, since it has a homogeneous structure and similar data in each dataframe, obtaining the dataset that is shown in Appendix A, Figure A1, which will form the basis of all the work.

2.2. Explore

In the exploration phase, the data from the selected information sources is analyzed, with a view to understanding its structure and the linkage it has with a view to the problem to be solved, in order to select the aspects that are most closely linked to it (Abdul Rahman et al. 2019). In order to address this phase, we will carry out the following tasks:

- Exploration of each column, whose result appears in Table A1 of Appendix A
- Analysis of quantities, which are shown in Figures 2 and 3.
- Column selection, which is shown in Table 1.

As part of the exploration of the information, it is very useful to generate graphs that show totals of records by different characteristics, such as time zone, type of crime, communes, etc. It should be noted that under this type of data analysis, unique data types, such as the case of the "id" column, should not be considered. It is frequent that, throughout the use of the methodology, it is necessary to alter the order of the phases, as in this case when it is useful to advance on the transformation of the date field, dividing it into the day, month and year columns. There are characteristics for which the data present a uniform distribution, or very close to it, such as "year", "month", and "day".

From the graph on the left, we can already see certain trends such as that relating to the month of January, in which crimes increased over the years, while in November crimes decreased over the years. Crimes increased during six months, mainly those at the beginning of the year (January, February, March, April, June, and July); while they decreased during the remaining six months, mainly during the last half of the year (May, August, September, October, November, and December).
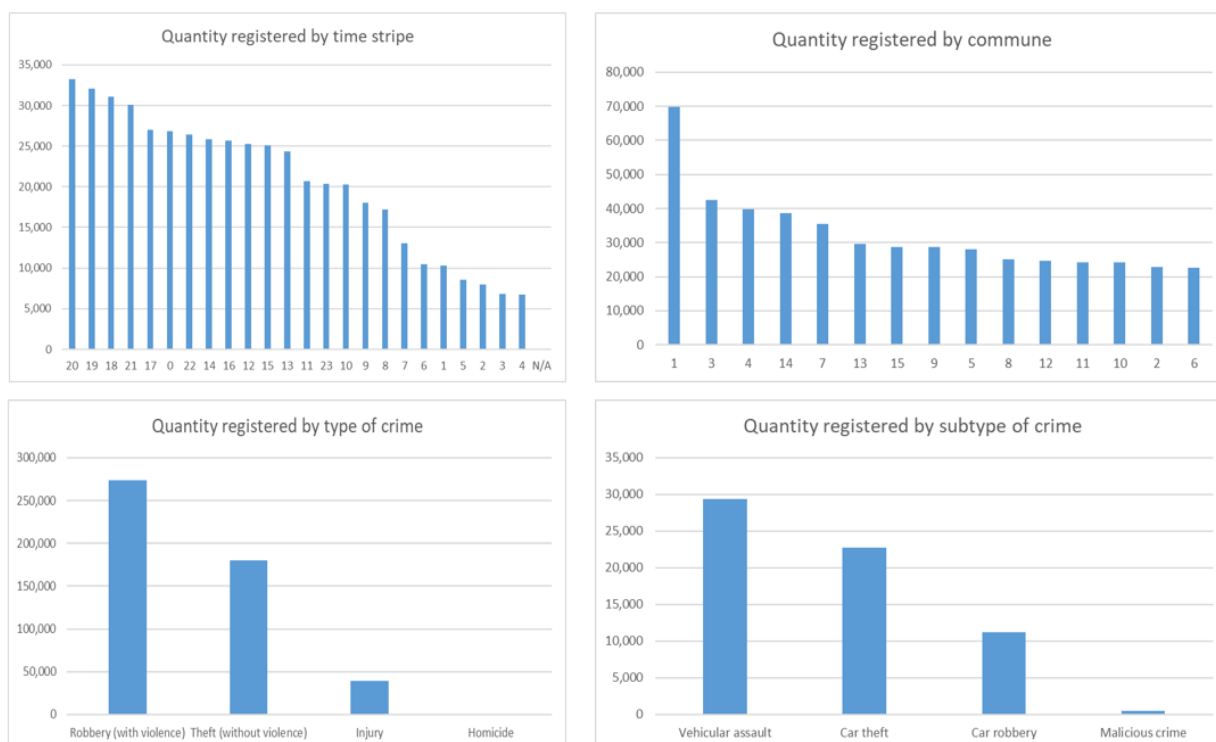
Figure 2. Non-uniform distribution of data. Source: Author's elaboration.
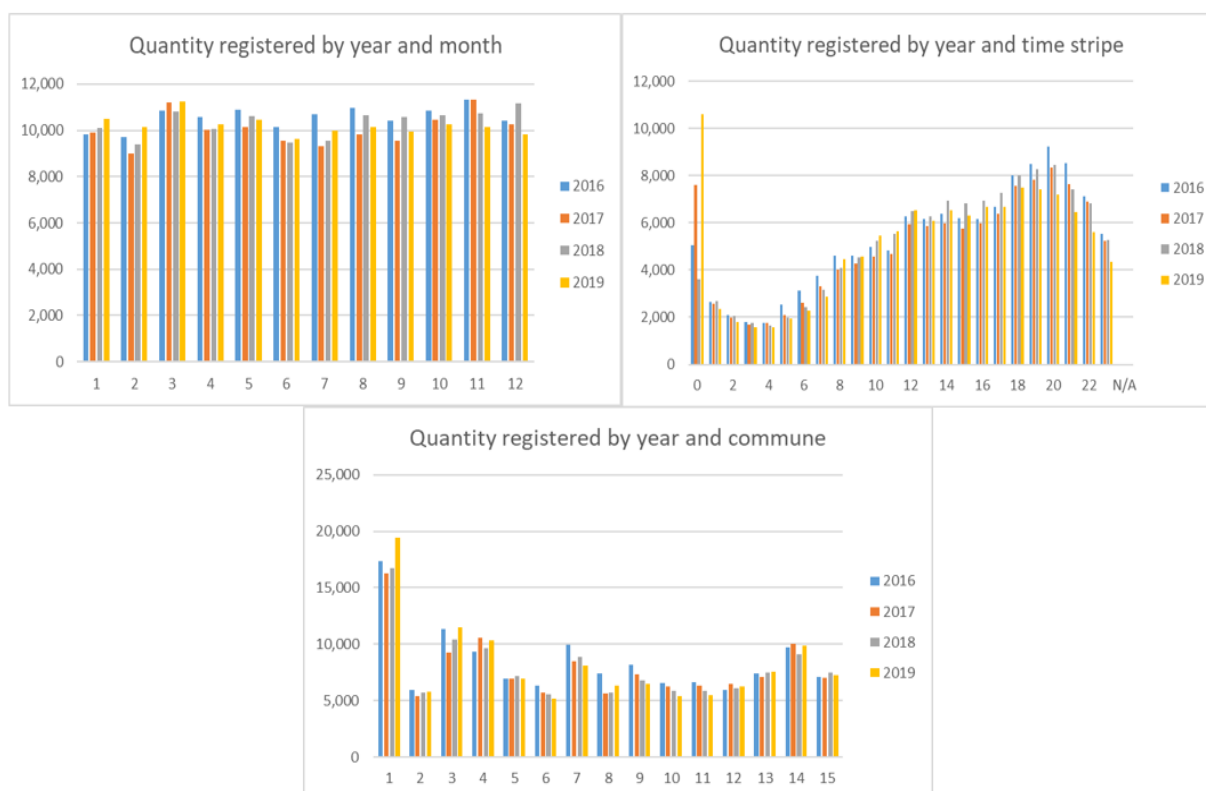

Figure 3. Amounts of crimes by year and month, by year and time zone, and by year and commune. Source: Author's elaboration.

From the central graph, a decrease is observed over time (or similar scenario to the previous year) in the time bands from 1:00 h to 7:00 h and from 20:00 h to 23:00 h. No clear

pattern of continuous growth of cases over the years is identified. In the last year there has been a decrease in crimes in the ranges of 1:00 h to 7:00 h and 13:00 h to 23:00 h. In the last year an increase is observed in the 0 hour and in the range of 8:00 h to 12:00 h.

From the right graph, it is observed that with the years a diminution of crimes is produced in three communes (6, 10 and 11). No clear pattern of continuous growth of cases over the years is identified. In the last year there has been a decrease in crimes in seven municipalities (46.67%), which are: 5, 6, 7, 9, 10, 11, and 15. In the last year there has been an increase in eight municipalities (53.33%), which are: 1, 2, 3, 4, 8, 12, 13, 14.

After analyzing the information provided by the various graphs, and by virtue of achieving the objectives of the model to be built, there are columns that will be discarded:

- Id: the fields that are unique for each record are essential for the transactional processes, however, they do not provide any information in the predictive models.
- Date: the temporary information in date format, do not provide substantial information in the predictive models, although they do provide information that can be extracted from them, such as the number of the month of the year.
- Neighborhood: although this data could be very useful, it has been defined in advance that the highest level of detail would be the commune, therefore, this column that bridles a higher level of granularity will be discarded.
- Lat and Long: both columns provide information on the geographic coordinates of the place where the reported crime occurred, therefore, it is a level of detail even higher than the one that bridles the information of the neighborhood, and since it was decided to limit the level of detail to the commune, it is appropriate to exclude it.
- Sub-type of crime: In the area of definition and application of preventive measures for the different crimes, it is very important to have the level of detail of the crime. However, since this column presents more than 85% of the null values, it is decided to exclude it.

Thus, to date, the columns or characteristics that make up the data set are:

Table 1. Structure of selected data.

Predictors	Objective
Year	
Month	
Day	Quantity registered
Time_stripe	
Commune	
Type of crime	

Source: Author's elaboration.

2.3. Modify

This phase is aimed at manipulating the available data to adapt it to the form that allows obtaining adequate results from the Machine Learning models (Krishnan et al. 2016; Gudivada et al. 2017), such as:

- Tabular format: the data must be organized in table form.
- Clean data: data cleanup is a topic in itself, it can demand significant project time. Among the main aspects that should be addressed are
- Null (nan): there should be no null data, so there are techniques such as deleting those rows, completing them with a default value, imputing with statistical information (mean, median, etc.).
- Outliers: these are out-of-range, atypical data. The treatment techniques include, deleting the rows containing these data, their transformation, etc.
- Numerical values: the models cannot manage non-numerical data, so different techniques are used to transform the information, such as encoding.
- Similar value scales: this technique avoids that the larger values have a greater preponderance over the smaller ones (example: Year: 2019 vs. Crime_type=1).

Accordingly, this phase will be addressed according to the following structure: Data cleansing, Coding, Information consolidation, and Standardization.

Regarding data cleaning: The crime data have no outliers, are quite limited and therefore do not require special analysis in this regard. Only the value 'S/D' will be removed from the "time_band" field, which will allow converting the data type to integer (int). With respect to the null data, the column with a high percentage of null data was already discarded in the selection of attributes, which is the case of the subtype of crime. This leaves only the case of the 'commune' attribute, where there are 8324 rows with a null value (1.72 percent). Given that the concept of commune is highly relevant to the analysis to be undertaken, in this case it is appropriate to eliminate those rows (Krishnan et al. 2016).

With respect to coding, analyzing the data types of the columns of the data set results in the following: It can be observed (see Figure A2 in Appendix A) that the data types of 'time_stripe' and 'crime_type' are not correct, since they are indicated as 'object' when they should be of integer type (int). The field 'time_stripe' is not interpreted as a numerical data type because before it had the value 'S/D', but as the records that had this information have been eliminated, it is possible to make the modification without problems. The case of 'crime_type' is different, because it has non-numerical values, and must be transformed to incorporate it into the models. This change can be made manually or using libraries such as the case of "LabelEncoder" belonging to Sklearn.

As for the consolidation of information, in the selection of columns, characteristics that provided a higher level of granularity have been eliminated, such as the neighborhood. In this way, it is expected that the dataset has repeated rows, and depending on the model to be created, it is required to apply aggregation techniques for the consolidation of records. All this is shown in Figure A3 of Appendix A. The resulting dataset is more compact, with 123,681 rows less, equivalent to a 25% reduction. Although in most cases only one crime has been recorded, there are multiple cases that even reach 60.

Finally, with respect to standardization, once the data cleaning process has been completed, given the variation in magnitudes, provided especially by the "year" column, the standardization must be carried out (Ilyas and Chu 2019), a process by which the data is transformed in such a way that the magnitudes are now within the range of 0 and 1. In this case the standardization will be applied to the set X, while it is not believed necessary to apply it to Y. The sets are formed in the following way:

- X (independent variables): *"year"*, *"month"*, *"day"*, *"time_band"*, *"commune"*, *"crime_type"*.
- Y (dependent variable): *"quantity_registered"*.

For the standardization is used the library sklearn.preprocessing.StandardScaler, which for each column, and replaces the value of the set by z, resulting from the equation:

$$z = \frac{(x - \mu)}{\sigma}$$

where μ is the mean and σ is the standard deviation of the column analyzed.

2.4. Model

The objective of this phase is to establish a relationship between the explanatory variables and the variables under study, which will make it possible to infer their value with a given level of confidence (Nowozin and Wright 2012). The techniques used include traditional statistical methods such as clustering methods and regression analysis, as well as data-based techniques such as neural networks, decision trees, etc. (Al-Jarrah et al. 2015).

2.4.1. Clustering

K-Means is a grouping method, which aims at partitioning a set of n observations into k groups in which each observation belongs to the group whose mean value is closest. It is a method used in data mining (Morissette and Chartier 2013). This algorithm to perform the grouping requires the definition of the K parameter, for this we use the elbow method (elbow) consisting of evaluating the error that is obtained with different values of k, and at

that point where there is an abrupt fall is the value that is selected (Bholowalia and Kumar 2014; Marutho et al. 2018). It is possible to observe in Figure 4, that with K = 4 there is a decrease in the errors.
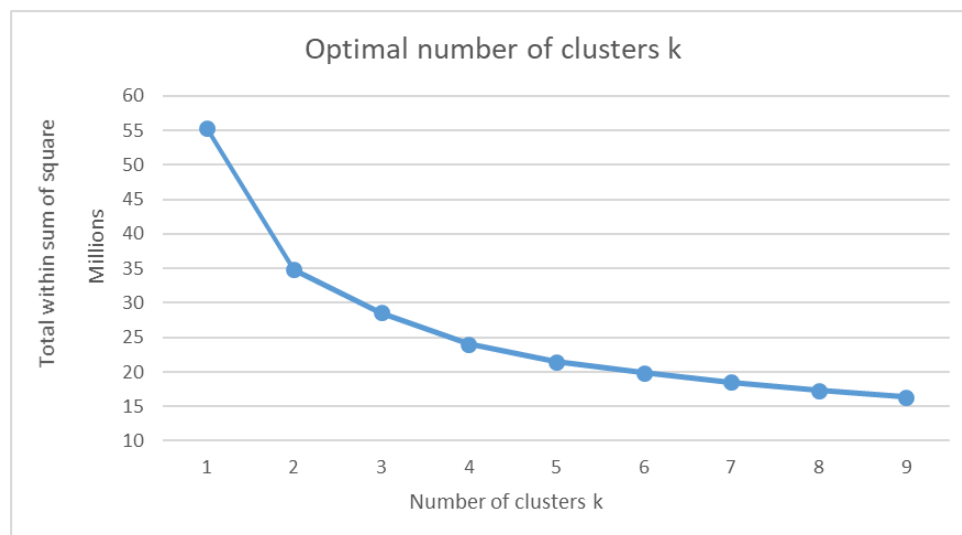

Figure 4. K-Means errors for different K values: Author's elaboration.

In the exploration phase, the application of unsupervised models is very useful since they allow the identification of relationships between the different variables of the data set, which a priori are not evident. Such is the case of K-means, which is a grouping method, which divides a set of n observations into k groups, assigning to each the group whose mean value is the closest.

Once the partition is made, the exploitation of the findings, which is generally sought in a visual way (in 2 or 3 dimensions), seeks the representation of a subset of variables in which a separation (ideally linear) between the data can be clearly identified. Thus, for example, in this case it could be observed that, the graph of the data set considering the time band and the day, there is a linear separation between four groups formed.

Although in K = 3 there is also a decrease in errors, finally K = 4 is selected. The reason is that, in other combinations of variables, it is not possible to clearly identify the 4 groups; as well as for reasons of ease and speed of processing.

The visualization of the data set for the selected K, by the predictive variables of "day" and "time_band", it is possible to visualize clearly the 4 groups that are delimited by the quadrants defined by day 16 and 12 h, as it is shown in Figure 5.

2.4.2. Neural Networks

Within the scope of neural networks, in the context of the regression problem to be solved, a multilayer perception will be employed. This type of network is formed by multiple layers, which allow you to solve problems that are not linearly separable. It can be totally or locally connected. This multilayer perception has been used in different fields such as predicting solar radiation as a function of time (Azimi et al. 2016), wind speed (Madhiarasan and Deepa 2016), landslides (Pham et al. 2017), or even rain (Esteves et al. 2018). This study will be applied to predict the commission of one of the crimes under study. In the first case each exit of a neuron of the "i" layer is an entry of all the neurons of the "i + 1" layer, while in the second case each neuron of the "i" layer is an entry of a series of neurons (region) of the "i + 1" layer.

The layers can be classified into three types, as shown in Figure 6.
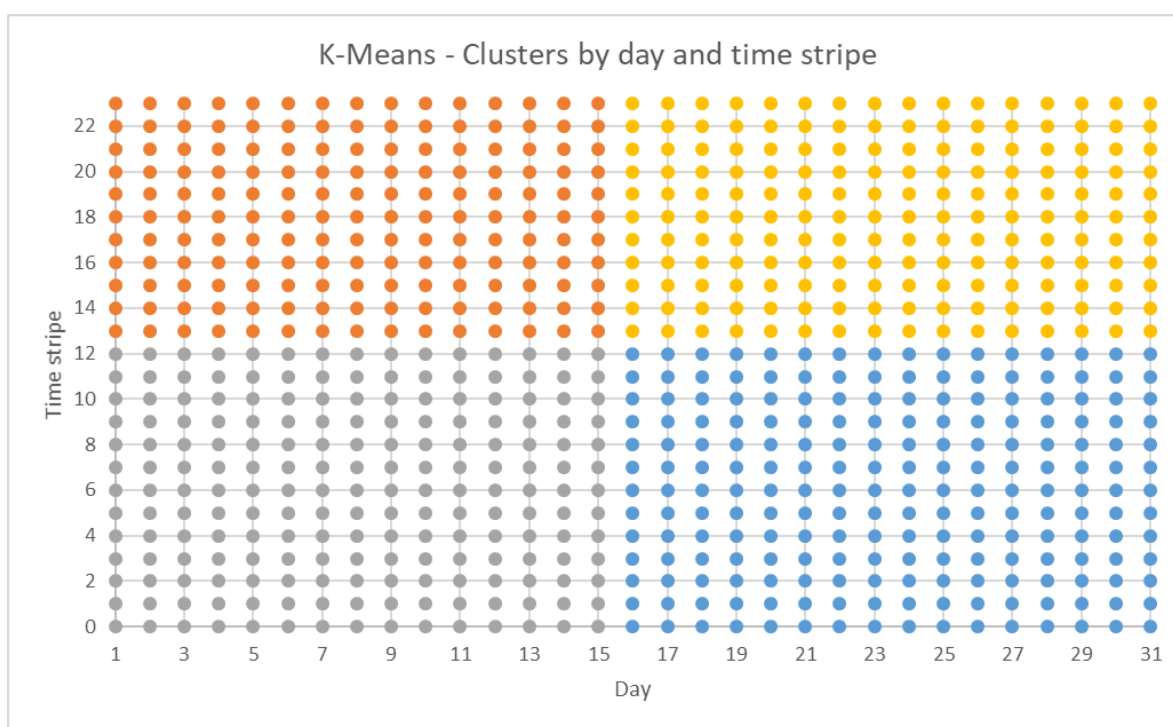
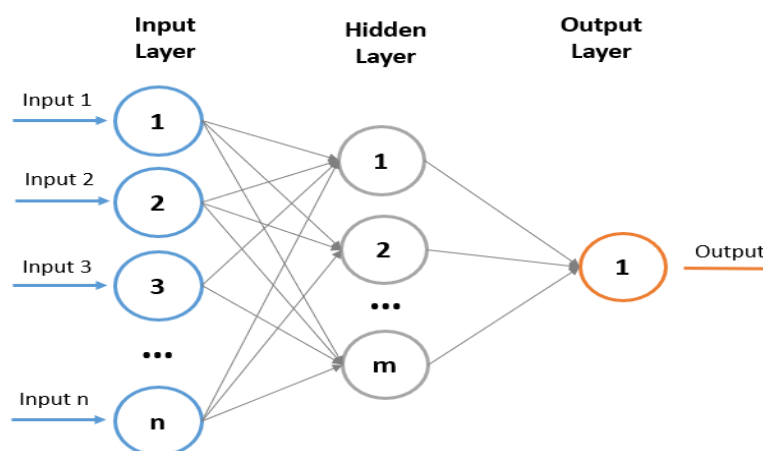**Figure 5.** Source: Author's elaboration.

Figure 6. Source: Multilayer neural network diagram. Own elaboration.

- Input Layer: Made up of those neurons that introduce the input patterns into the network. No processing takes place in these neurons.
- Hidden layers: Formed by those neurons whose inputs come from earlier layers and whose outputs pass to neurons in later layers.
- Output Layer: Neurons whose output values correspond to the outputs of the entire network.

In this case, a 3-layer model with the following characteristics will be used, which is also shown in Figure 7:

1. Input layer: 16 neurons all fully-connected with a "*ReLU*" (rectified linear unit) activation (Kutyniok 2020) that has a value of 0 for negative values, and the identity function for positive values.
2. Hidden layer: 1 hidden layer with 8 neurons and also ReLU type activation.
3. Output layer: 1 output layer with 1 neuron, and "*linear*" type activation.

```
Model: "sequential"

Layer (type)                    Output Shape              Param #
=================================================================
dense (Dense)                   (None, 16)                112

dense_1 (Dense)                 (None, 8)                 136

dense_2 (Dense)                 (None, 1)                 9

=================================================================
Total params: 257
Trainable params: 257
Non-trainable params: 0
```

Figure 7. Defined multilayer perception model. Source: Own elaboration.

The model compilation parameters are: Loss function: mean_absolute_error (MAE), Optimizer: Adam, Metric: 'mean_squared_error' (MSE), Epochs: 20 and Data separation (validation_split) = 0.2 (20% of the data) Using Keras' model.Sequential and layers.core.Dense libraries, the following results are obtained:

- Results on training data: MAE (loss): 0.3318—MSE: 0.5578
- Results on evaluation data: MAE (loss): 0.4095—MSE: 1.4609

2.5. Evaluation

In this phase, the results are evaluated by analyzing the goodness of the models, contrasting them with other statistical methods or with new sample populations. The best scenario has been provided by the multilayer perception, therefore, in this phase we will proceed to make a fine adjustment of the various parameters of that neural network. The variation in the number of layers was made starting from 2 and reaching up to 4, as shown in Figure 8.

The tests were carried out through the execution of 30 periods. Next, the results obtained on the training data (train) and the validation data (val) can be seen in Figure 9, showing the mean square error (MSE). For the training data, the lowest value of MSE was obtained by the 4-layer option, while in the validation data (val) it was obtained by the 3-layer architecture. It is considered that the most appropriate option is the 3-layer option because it has a more stable behavior against the other configurations.
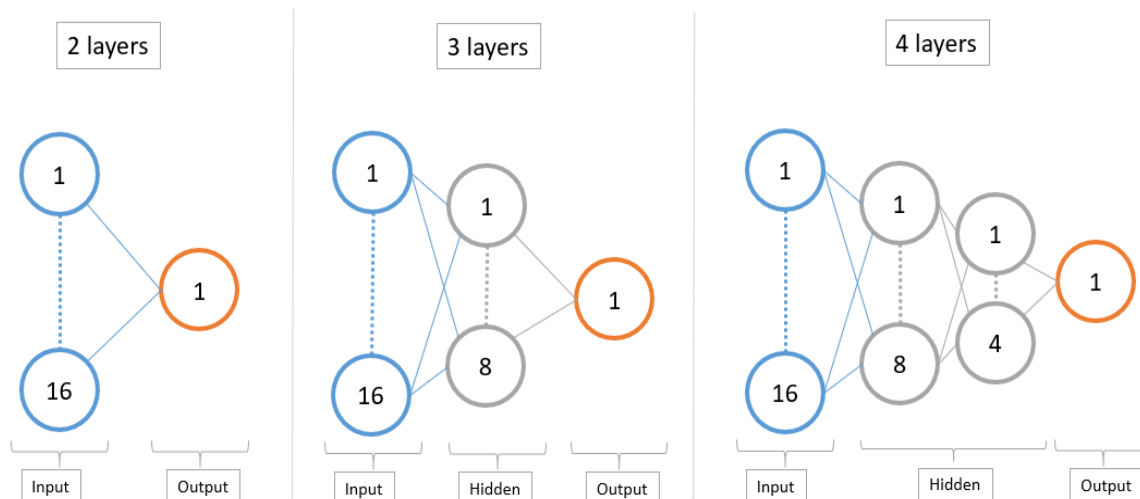

Figure 8. Source: Tests carried out on the variation of layers. Own elaboration.
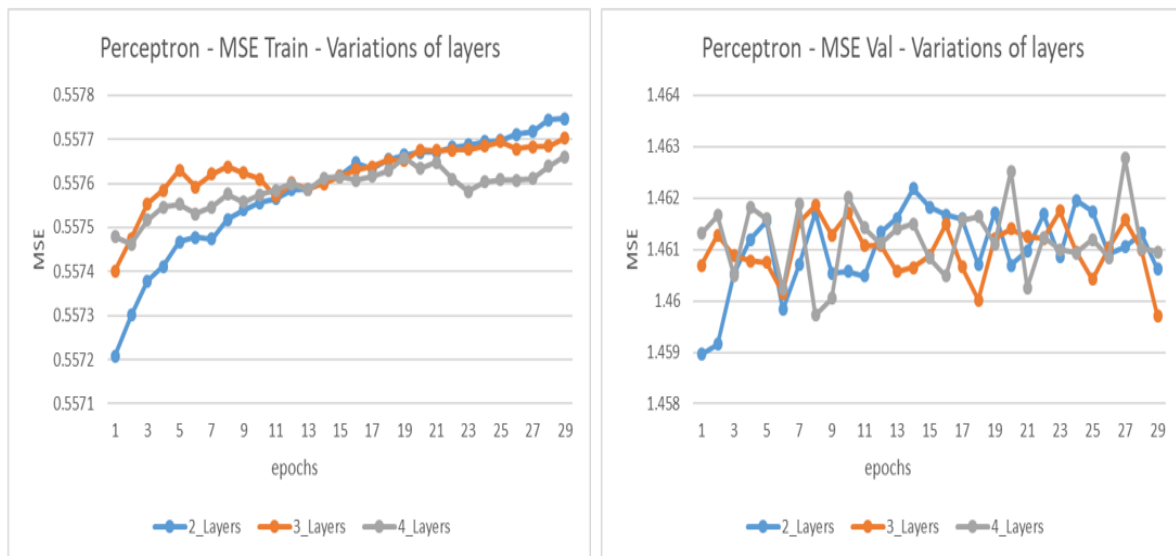
Figure 9. Comparative test results on the variation of layers Source Own elaboration.

Based on the architecture of the 3-layer perception, the alternatives considered in the variation of the amount of neurons, do so with different numbers in both the input and the hidden layers, while the output layer remains fixed in 1 neuron because it is subject to the output of the model. Five tests were carried out, with the diagrams shown in Figure A4 of Appendix A.

The tests were carried out by running 30 periods. In Figure 10, the results obtained on the training data (train) and the validation data (val) for the mean square error (MSE) can be seen. In the training data, the lowest MSE was obtained by option 32-16-1, followed by 16-8-1 and 32-8-1. The behavior for the validation data (val) the lowest error corresponds to the 32-8-1 architecture, followed by 16-8-1. It is considered as the most appropriate architecture the one with 32 neurons in the input layer, 8 in the hidden layer, and 1 in the output layer (32-8-1).

The loss function considered in the original neural network is the mean absolute error (MAE) which is a measure of the difference between two continuous variables and serves to quantify the accuracy of a predictive technique. In this neural network, the mean square error (MSE) was defined as a metric that measures the average of the squared errors, that is, the difference between the estimator and what is estimated (Tamhane 1999). Since ŷ is a predictor of n predictions, and the vector of n true values, metrics are defined as:

$$\text{MAE} = \sum_{i=1}^{n} \frac{(\hat{yi} - yi)}{n} \qquad \text{MSE} = \sum_{i=1}^{n} \frac{(\hat{yi} - yi)^2}{n}$$

Basically, the original neural network architecture was contrasted with the exchange between loss and metrics, which gave the result shown in Figure 11.

In both the training and validation datasets, the smallest error has been thrown by the MAE option as a loss function and MSE as a metric, which is the original setting. In an artificial neural network, each neuron calculates the linear combination of weights and inputs, the result of which, in order to be the output value, the activation function is previously applied. In the model in question the prediction that the model must give is a positive integer value, as opposed to options such as the cases in which the probability of belonging to a certain class must be predicted. For this reason, among the different options of activation functions, the tests were performed between linear, which is the identity function, and rectified linear (ReLU), which is the identity function for positive values, and 0 for negative values. The tests were performed maintaining the same activation

function for the input layer and the hidden layer, and on the other hand the function of the output layer.
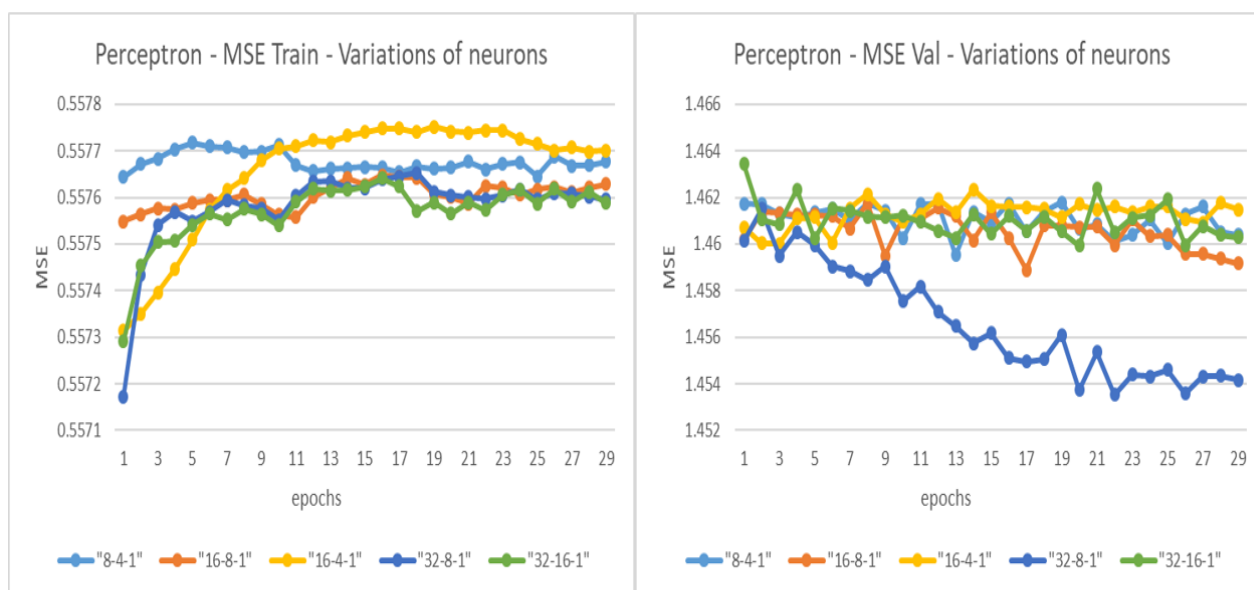

Figure 10. Source: Comparative results of tests on neuron variation. Own elaboration.
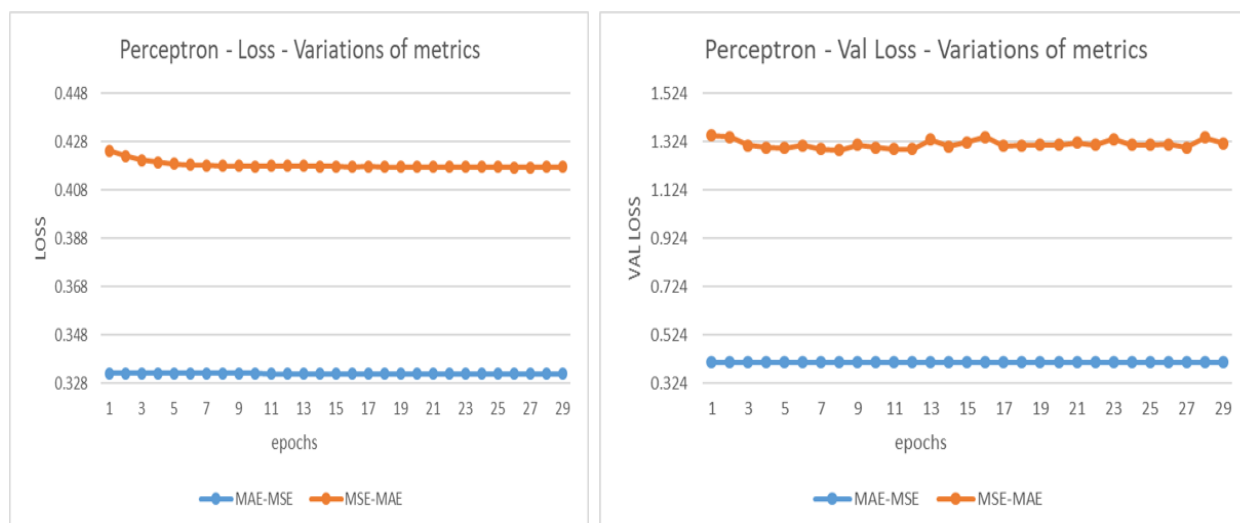

Figure 11. Source: Comparative results of tests on the variation of func. Loss. Own elaboration.

As can be seen in the results, both in the training and validation data sets the 3 options considered give similar mean square errors (MSE), however, the option of using ReLU type activation in all layers shows a more stable behavior, and therefore is considered the best alternative. This is shown in Figure 12.

Mathematical optimization is the selection of the best element (with respect to some criterion) of a set of available elements (Lan 2019). In neural networks, the optimization function is used to choose the strategy to minimize the loss function. The options considered have been:

1. SGD (Stochastic Gradient Descent) is the basic method of optimization, here the step of updating the weights (α) is fixed, it does not change during the training (Mandt et al. 2017).

2. RMSProp (Root Mean Square Propogation): in its formula it includes a parameter d, called decaying rate, which makes the effect of the previous values less influential than the new ones (Lu 2018).

3. Adam (Adaptive Moment Estimation) is similar to RMSProp but with momentum, that is, it not only considers the gradient of the current step, but also accumulates the gradients of previous steps (Kingma and Ba 2014; Tato and Roger 2018).

In both cases, i.e., with the training and validation values, the least error has been achieved by SGD. However, the optimizer that presents a more uniform behavior in both data sets is considered more appropriate, and this is only observed with Adam, which is the original network configuration. The results are shown in Figure 13.
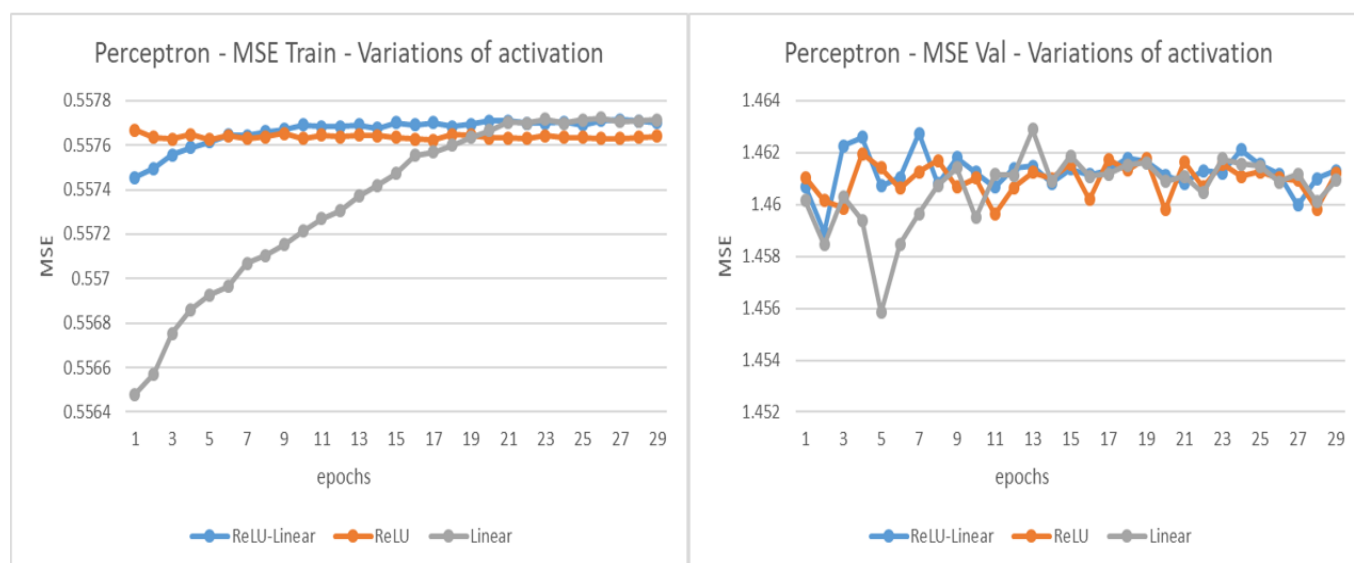

Figure 12. Comparative test results on the variation of func. Activation. Source: Own elaboration.
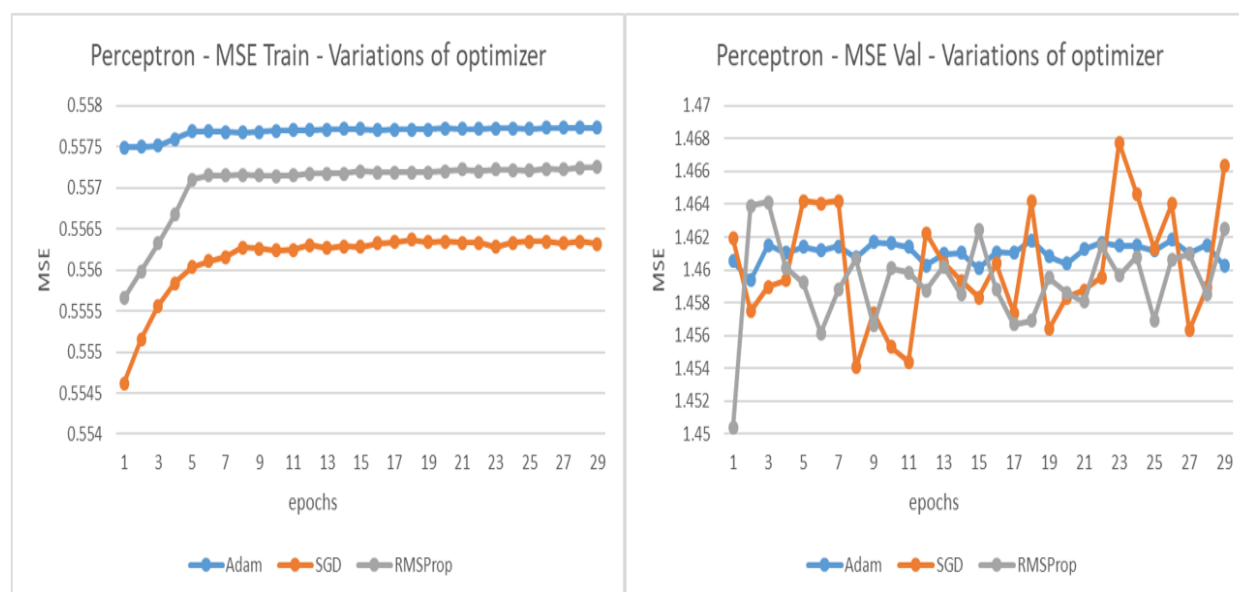

Figure 13. Source: Comparative results of tests on the optimizer variation. Own elaboration.

2.6. Selección y uso del Modelo Predictivo

Depending on the different tests carried out, it is identified as the best artificial neural network model for the predictive model of crimes, which has the following characteristics,

shown in Figures 14 and 15. To use the predictive model, the following information is required: Year, Day, Month, Time Zone, Commune, and Type of Crime. This information needs to be scaled up in order to be entered as input into the model, and then use the predict option to obtain the number of recorded crimes expected.
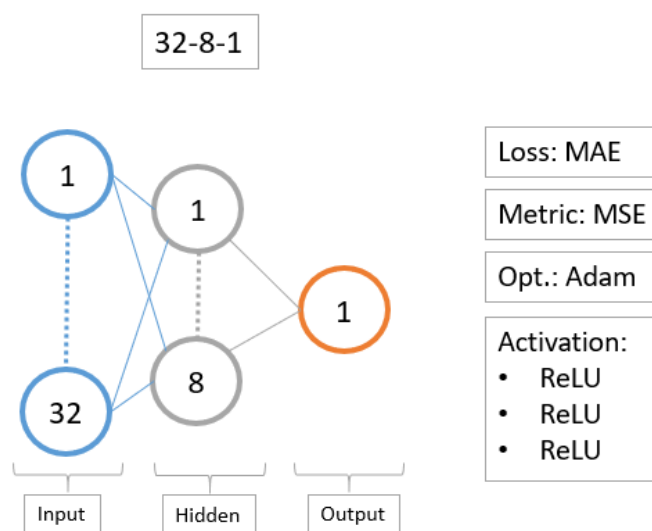

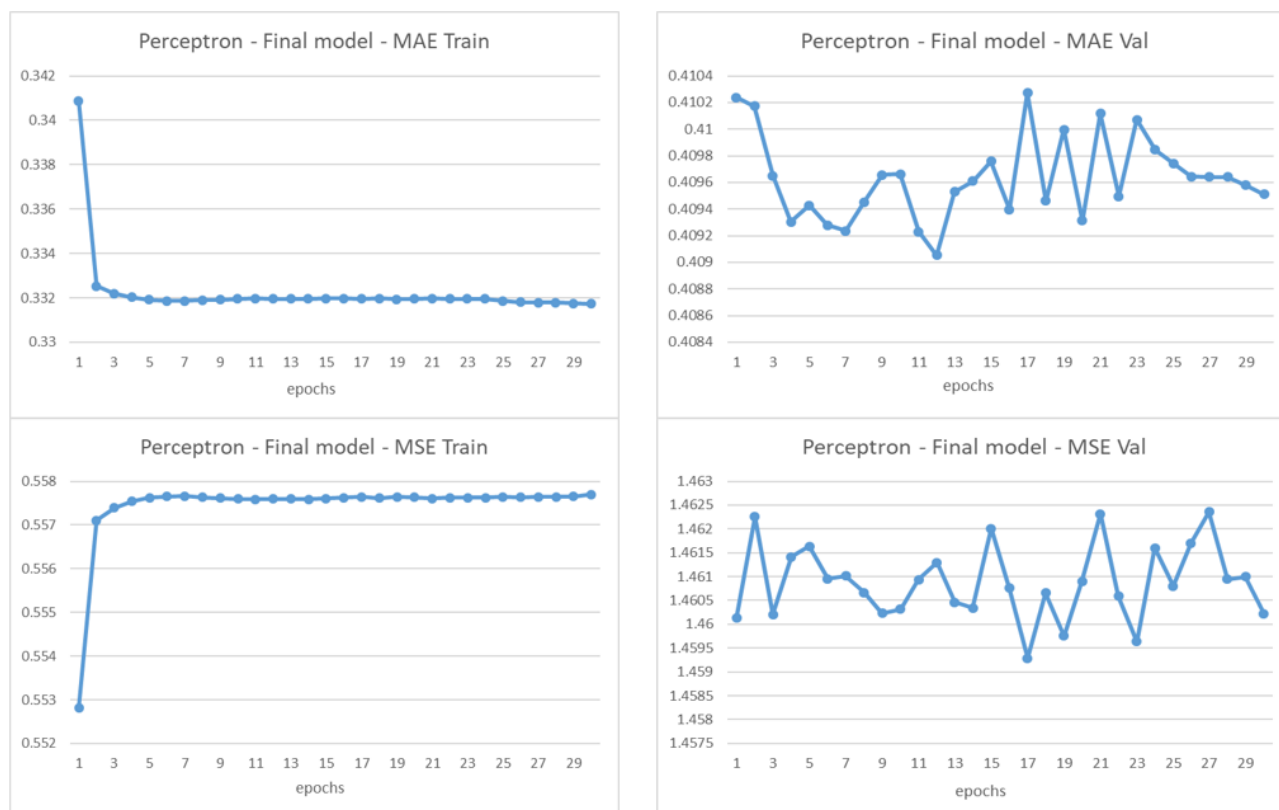**Figure 14.** Selected network architecture. Source: Own elaboration.

Figure 15. Source: Metrics resulting from the final trained model, Own elaboration.

3. Results

Using the crime data published by the Government of the Autonomous City of Buenos Aires, a predictive regression model of the number of crimes expected for a given date, time, commune and type of crime has been defined and trained. The available data had

information from 2016 to 2019, for the types of reported crimes that are: theft, robbery, injuries, and homicide. In the information exploration phase, it was possible to identify a close relationship between the information of the day and the time zone, being able to visualize 4 groupings, corresponding to 4 quadrants, delimited by the 16th day and the 12th hour. For the construction of the model, a multi-layer artificial neuronal network was selected, with a structure called perception, on which multiple tests were made that allowed identifying the appropriate hyperparameters to achieve better results. Therefore, the multilayer perception has the architecture shown in Figure 16.
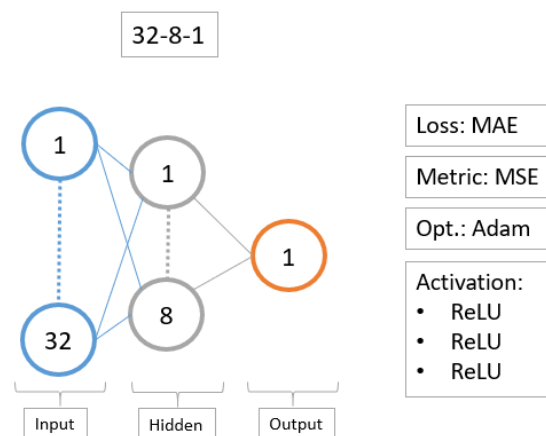

Figure 16. Selected network architecture. Source: Own elaboration.

The results obtained are:

- MAE: 0.3317 in training data and 0.4095 in validation dat.
- MSE: 0.5577 in the training data and 1.4602 in the validation data

This determines that for the prediction it is required to use the information corresponding to the predictive characteristics (predict), being these: year, month, day, time zone, commune and type of crime, to then scale them to feed the model. The model for the prediction of crimes in the city of Buenos Aires, created and implemented through a neural network called multilayer perception, allows for obtaining future information, not only with respect to possible crimes, but with the appropriate level of detail for the definition of preventive and detection measures in the different communes in which the city is organized. The information provided by the model is also useful in the decision-making process for the distribution of the budget allocated to Security.

From the initial data set, after the cleaning, transformation and grouping processes, a total of 361,184 records were obtained. This data set was divided into two groups: the training group, with 80%, equivalent to 288,947 records, and the validation group, with 20%, equivalent to 72,237 records.

Given that this is a regression problem, in contrast to those of classification, in the evaluation of the predictive quality of the model the concept of "accuracy" or quantity of correct vs. incorrect predictions is not used, but rather other types of metrics are used, such as the average of the sum of the absolute value of the differences between the estimated value and the correct value, which is called MAE (mean absolute error), and in this case the final model obtained has yielded a value of 0.4095, so it is expected to obtain an average error of approximately less than 0.5 cases, for example if in reality were recorded 3 cases, the prediction would yield values between 2.6 and 3.6, which, in terms of the possible modes of use of the model is appropriate.

4. Conclusions

The model of crime prediction in the city of Buenos Aires created, implemented through a neural network called multilayer perception, allows to obtain future information,

not only regarding possible crimes, but with the appropriate level of detail for the definition of preventive and detection measures in the different communes where the city is organized. Therefore, the information provided by the model can also be useful in the decision-making process for the distribution of the budget allocated to Security, more so in a context of crisis where government resources are scarce; it becomes even more necessary to optimize public spending. For the use of the model, it should be considered that the information regarding crimes is published by the government of the Autonomous City of Buenos Aires, within three months after the end of the year. The model created has been trained with the data corresponding to the years 2016 to 2019. Therefore, it is expected that the information for the year 2020 will be available in March 2021, and it would be convenient to update the model with this information. However, it will be a situation, given that due to the issues of the pandemic linked to the COVID-19 virus, restrictive mobility measures have been taken, among which home confinement predominates between the months of March and October. As a result of this situation, it is to be expected that a very different scenario will arise from the usual one and described by the data for the period 2016 to 2019. In this way, the incorporation of the information from 2020 will have to be properly tested and evaluated, considering different scenarios that could vary from making transformations, to even discarding them.

Appendix A

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 493364 entries, 0 to 493363
Data columns (total 10 columns):
 #  Column              Non-Null Count   Dtype
---  ------              --------------   --------
 0  id                  493364 non-null  int64
 1  date                493364 non-null  object
 2  time_stripe         493364 non-null  object
 3  crime_type          493364 non-null  object
 4  crime_subtype       63776 non-null   object
 5  registered_quantity 493364 non-null  float64
 6  commune             484905 non-null  float64
 7  neighborhood        484905 non-null  object
 8  lat                 484905 non-null  float64
 9  long                484905 non-null  float64
dtypes: float64(4), int64(1), object(5)
memory usage: 37.6+ MB
```

Figure A1. Initial data frame. Source: Own elaboration.

Table A1. Exploration of each column.

Name	Type of Information	Examples
id	Unique registration identification number.	51,585 83,256 486,683
	Integer data type. Has no values	
	null.	
date	Record date in "YY-mm-dd" format. It has no null values.	1 January 2016 31 December 2017
time_band	Time in which the occurrence of the event was reported, ranging from 0 to 23. It has no null values	0 13 22
crime_type	Type of crime reported. It corresponds to the 4 values contained in the dataset. It has no null values.	Stole Theft
crime_subtype	Subtype of crime reported.	Sinister Road Automotive Theft Automotive Theft
registered_quantity	There are 4 values, and 85.84% is indicated as null (nan).	1 2 3 4
commune	Amount of crimes included in the registry. They are values that go from 1 to 4. It does not have null values.	1 2 14
neighborhood	Commune in which the reported crime occurred. They are integer values that go from 1 to 15. 1.71% of the data are null.	Palermo Belgrano Chacarita
lat	Neighborhood (within the Commune) in which the crime reported occurred. 1.71% of the data are null.	−34.64872987 −34.64519741 −34.692347
long	Latitude. It is part of the geographical coordinates of the place where the crime reported occurred. 1.71% of the data are null.	−58.46853592 −58.40109886 −58.492738

Source: Own elaboration based on data extracted from https://data.buenosaires.gob.ar/dataset/delitos.

```
df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 484905 entries, 0 to 122483
Data columns (total 7 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   --------
 0   time_stripe          484905 non-null  object
 1   crime_type           484905 non-null  object
 2   registered_quantity  484905 non-null  float64
 3   commune              484905 non-null  float64
 4   year                 484905 non-null  int64
 5   month                484905 non-null  int64
 6   day                  484905 non-null  int64
dtypes: float64(2), int64(3), object(2)
memory usage: 29.6+ MB
```

Figure A2. Result of the execution of df.info(). Source: Own elaboration.

```
df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 484865 entries, 0 to 493363
Data columns (total 7 columns):
 #  Column              Non-Null Count     Dtype
--- ------              --------------     --------
 0  time_stripe         484865 non-null  int32
 1  crime_type          484865 non-null  int32
 2  registered_quantity 484865 non-null  int32
 3  commune             484865 non-null  int32
 4  year                484865 non-null  int32
 5  month               484865 non-null  int32
 6  day                 484865 non-null  int32
dtypes: int32(7)
memory usage: 16.6+ MB
```

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 361184 entries, 0 to 361183
Data columns (total 7 columns):
 #  Column              Non-Null Count     Dtype
--- ------              --------------     --------
 0  year                361184 non-null  int64
 1  month               361184 non-null  int64
 2  day                 361184 non-null  int64
 3  time_stripe         361184 non-null  int64
 4  commune             361184 non-null  int64
 5  crime_type          361184 non-null  object
 6  registered_quantity 361184 non-null  int32
dtypes: int32(1), int64(5), object(1)
memory usage: 17.9+ MB
```

Figure A3. Data frame before and after the consolidation process Source Own elaboration.
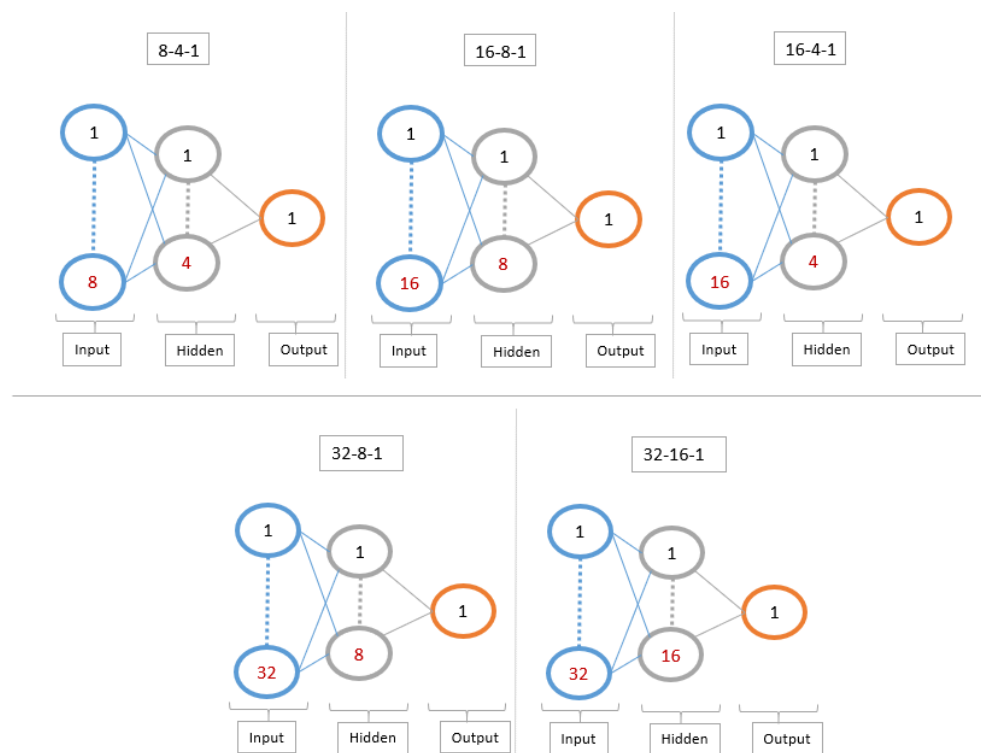

Figure A4. Tests performed on the variation of neurons. Source: Own elaboration.

References

Abdul Rahman, Fauziah, Rahimah Kassim, Zirawani Baharum, Helmi Adly Mohd Noor, and Norhaidah Abu Haris. 2019. Data Cleaning in Knowledge Discovery Database-Data Mining (KDD-DM). *International Journal of Engineering and Advanced Technology* 8: 2196–99. [CrossRef]

Alegre, Romina. 2019. Policías comunales y policías locales. Un estudio de la seguridad en una ciudad del interior de la provincia de Buenos Aires. In *Actores e instituciones de la seguridad en la provincia de Buenos Aires (2010–2018)*. Buenos Aires: Clacso, pp. 115–45. Available online: https://www.jstor.org/stable/j.ctvt6rmwh.7 (accessed on 17 October 2020).

Al-Jarrah, Omar, Paul Yoo, Sami Muhaidat, George Karagiannidis, and Kamal Taha. 2015. Efficient Machine Learning for Big Data: A Review. *Big Data Research* 2: 87–93. [CrossRef]

Arumugam, Paramasivan, A. Poompavai, Gamesa Manimannan, and Suruliandi Poongothai. 2020. Comparative Study of Data Mining Classifiers with Different Features and Different Databases Domain. *International Journal of Scientific and Innovative Mathematical Research* 8: 1–11. [CrossRef]

Azimi, Rasool, Mohadeseh Ghayekhloo, and Mahmoud Ghofrani. 2016. A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. *Energy Conversion and Management* 118: 331–44. [CrossRef]

Batarseh, Feras A., and Ruixin Yang. 2017. *Federal Data Science: Transforming Government and Agricultural Policy Using Artificial Intelligence*. Cambridge: Elsevier Academic Press.

Bholowalia, Pumina, and Arvind Kumar. 2014. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications* 105. [CrossRef]

Cerrillo i Martínez, Agustí. 2019. Retos y oportunidades del uso de la inteligencia artificial en las administraciones públicas. *Oikonomics* 6. [CrossRef]

Cheng, Dawei, Zibin Niu, and Liqing Zhang. 2020. Delinquent Events Prediction in Temporal Networked-Guarantee Loans. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13. [CrossRef]

Cid Ferreira, Lucía, Matías Lorenzo Pisarello, and Roxana Laks. 2017. Observaciones sobre el delito en relación con los contextos económicos en la Argentina contemporánea. In *I Jornadas de estudios sociales sobre delito, violencia y policía 20 y 21 de abril de 2017 La Plata y Quilmes, Argentina. La seguridad en cuestión*. La Plata: Universidad Nacional de La Plata, Facultad de Humanidades y Ciencias de la Educación.

Cinelli, Virginia, and Alberto Manrique. 2019. El uso de programas de análisis predictivo en la inteligencia policial: Una comparativa europea. *Revista de Estudios En Seguridad Internacional* 5: 1–20. [CrossRef]

Corvalán, Juan Gustavo. 2018. Inteligencia artificial: Retos, desafíos y oportunidades—Prometea: La primera inteligencia artificial de Latinoamérica al servicio de la Justicia. *Revista de Investigações Constitucionais* 5: 295. [CrossRef]

Delahoz-Dominguez, Enrique José, Tomás José Fontalvo-Herrera, and Adel Alfonso Mendoza-Mendoza. 2020. Definición de Perfiles Geográficos de hurto de automóviles. Caso Aplicado en Cartagena. *Justicia* 25: 99–108. [CrossRef]

Esquivel, Nicolás, Orietta Nicolis, Billy Peralta, and Jorge Mateu. 2020. Spatio-temporal prediction of Baltimore crime events using CLSTM neural networks. *IEEE Access* 8: 209101–12. [CrossRef]

Esteves, Joao Trevizoli, Glauco de Souza Rolim, and Antonio Sergio Ferraudo. 2018. Rainfall prediction methodology with binary multilayer perceptron neural networks. *Climate Dynamics* 52: 2319–31. [CrossRef]

García-Plúa, Juan, Wellington Villota-Oyarvide, and Julio Litardo-Unda. 2017. Detección automática del nivel de crimen basado en el análisis de puntos calientes en la ciudad de Guayaqui. *Dominio de Las Ciencias* 3: 367–79.

Gómez-Palacios, Toledo, Giovanni Hernández, and Navarro. 2017. A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Advances in Science, Technology and Engineering Systems Journal* 2: 598–604. [CrossRef]

Guanin, Fajardo, Jorge Humberto, Carlos González Guanin, Damian Santacruz Fernández, and Byron Oviedo. 2015. Utilización de algoritmos de clasificación para la predicción de los delitos que afectan la seguridad ciudadana en Guayaquil. pp. 1–5. Available online: https://www.researchgate.net/profile/Jorge_Guanin_Fajardo/publication/308631782_Utilizacion_de_algoritmos_de_clasificacion_para_la_prediccion_de_los_delitos_que_afectan_la_seguridad_ciudadana_en_Guayaquil/links/57e9381a08aed0a291302f6d.pdf (accessed on 15 October 2020).

Gudivada, Venkat, Amy Apon, and Junhua Ding. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10: 1–20.

Ilyas, Ihab, and Xu Chu. 2019. *Data Cleaning*. New York: Morgan &Claypool.

Jorge-Vázquez, Javier, María Peana Chivite-Cebolla, and Francisco Salinas-Ramos. 2019. La transformación digital en el sector cooperativo agroalimentario español: Situación y perspectivas. *CIRIEC-España. Revista de Economía pública, social y cooperativa* 79: 5–34. [CrossRef]

Kingma, Diederik, and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv* arXiv:1412.6980.

Kliksberg, Bernardo. 2002. El crecimiento de la criminalidad en América Latina: Un tema urgente. *Multiciencias* 2: 85–91. Available online: https://www.redalyc.org/articulo.oa?id=904/90420201 (accessed on 4 October 2020).

Krishnan, Sanjay, Michael Franklin, Ken Goldberg, Jiannan Wang, and Eugene Wu. 2016. Activeclean: An interactive data cleaning framework for modern machine learning. Paper presented at the 2016 International Conference on Management of Data, San Francisco, CA, USA, June 26–July 1; pp. 2117–20.

Kutyniok, Gitta. 2020. Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function". *Annals of Statistics* 48: 1902–5. [CrossRef]

La Ruffa, Silvia Paola. 2019. Informe sobre Seguridad Pública en la Ciudad de Buenos Aires (2016–2018): Primeras aproximaciones. Cartografías Del Sur. *Revista de Ciencias, Artes y Tecnología* 10: 148–68. [CrossRef]

Lan, Guanghui. 2019. Lectures on Optimization Methods for Machine Learning. Retrieved from Lectures on Machine Learning. Available online: https://cpb-us-w2.wpmucdn.com/sites.gatech.edu/dist/f/330/files/2019/08/LectureOPTML.pdf (accessed on 11 October 2020).

Lozano, Cortés, René Leticia, Luis Fernando Cabrera Castellanos, and Maribel Lozano Cortés. 2018. Causas sociales y económicas de los delitos en México. *Boletín Criminológico* 24. [CrossRef]

Lu, Yanmin. 2018. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. Retrieved from IOP Conference Series: Materials Science and Engineering-Papers. Available online: https://www.researchgate.net/publication/324265077_Analysis_of_the_Mean_Absolute_Error_MAE_and_the_Root_Mean_Square_Error_RMSE_in_Assessing_Rounding_Model/link/5ac81d9ca6fdcc8bfc8143f8/download (accessed on 2 October 2020).

Madhiarasan, M., and S. N. Deepa. 2016. Comparative analysis on hidden neurons estimation in multi layer perceptron neural networks for wind speed forecasting. *Artificial Intelligence Review* 48: 449–71. [CrossRef]

Mancini, Inés. 2020. Las luchas contra el delito y sus efectos en la Argentina. Sobre la expansión del encarcelamiento y las dificultades de seguir a nuestros presos. *Revista CS*, 139–58. [CrossRef]

Mandt, Stephan, Hoffman Matthew, and David Blei. 2017. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research* 18: 4873–907.

Marutho, Dhendra, Sunarna Hendra Handaka, and Mujolno Wijaya. 2018. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. Paper presented at 2018 International Seminar on Application for Technology of Information and Communication, Semarang, Indonesia, September 21–22; pp. 533–38.

Ministerio de Justicia y Seguridad. Policía de la Ciudad. 2020. Buenos Aires Data—Delitos. Buenos Aires Data Website. Available online: https://data.buenosaires.gob.ar/dataset/delitos (accessed on 18 November 2020).

Morissette, Laurence, and Sylvain Chartier. 2013. The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology* 9: 15–24. [CrossRef]

Náñez Alonso, Sergio Luis. 2019. Activities and Operations with Cryptocurrencies and Their Taxation Implications: The Spanish Case. *Laws* 8: 16. [CrossRef]

Náñez Alonso, Sergio Luis, Miguel Ángel Echarte Fernández, David Sanz Bas, and Jaroslaw Kaczmarek. 2020a. Reasons Fostering or Discouraging the Implementation of Central Bank-Backed Digital Currency: A Review. *Economies* 8: 41. [CrossRef]

Náñez Alonso, Sergio Luis, Javier Jorge-Vázquez, and Ricardo Francisco Reier Forradellas. 2020b. Detection of Financial Inclusion Vulnerable Rural Areas through an Access to Cash Index: Solutions Based on the Pharmacy Network and a CBDC. Evidence Based on Ávila (Spain). *Sustainability* 12: 7480. [CrossRef]

Nowozin, Sebastian, and Stephen Wright. 2012. *Optimization for Machine Learning*. Cambridge: MIT Press.

Olave Albertini, Alejandra. 2018. El delito de hurto como tipo de delito de resultado. *Política Criminal* 13: 175–207. [CrossRef]

Ordoñez-Eraso, Hugo, Cesar Jesús Pardo-Calvache, and Carlos Cobos-Lozada. 2019. Detección de tendencias de homicidios en Colombia usando Machine Learning. *Revista Facultad de Ingeniería* 29: e11740. [CrossRef]

Pham, Binh Thai, Dieu Tien Bui, Indra Prakash, and Mrugen Dholakia. 2017. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA* 149: 52–63. [CrossRef]

Ramos Pérez, César. 2015. Comentario al nuevo artículo 443 bis del código penal, que sanciona el robo con fuerza de cajeros automáticos. *Revista de Estudios de La Justicia* 19: 127–56. [CrossRef]

Rodas, Alejandro. 2019. Análisis espacial y temporal del hurto de celulares, Pereira, Risaralda, año 2018. *Revista Logos Ciencia and Tecnología* 11: 167–75. [CrossRef]

Rodríguez Collao, Luis. 2011. Robo con homicidio. *Revista de Estudios de La Justicia* 11: 131–51. [CrossRef]

Rodríguez, Montequín, María Teresa, Valeriano Álvarez Cabal, José Manuel Fernández, and Adolfo González Valdés. 2003. Metodologías para la realización de proyectos de Data Mining. Obtenido de. Available online: https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf (accessed on 25 October 2020).

Sain, Marcelo. 2012. La criminalidad común y la criminalidad organizada en la Provincia de Buenos Aires. *Cuadernos de Seguridad* 15: 39–77.

SAS: Analytics, Artificial Intelligence and Data Management. 2020. Available online: https://www.sas.com/es_es/home.html (accessed on 19 November 2020).

Tamhane, Ajit. 1999. *Statistics and Data Analysis: From Elementary to Intermediate*. Upper Saddle River: Prentice Hall, Retrieved from Basic Concepts of Inference. Available online: https://ocw.aprende.org/courses/sloan-school-of-management/15-075j-statistical-thinking-and-data-analysis-fall-2011/lecture-notes/MIT15_075JF11_chpt06b.pdf (accessed on 10 November 2020).

Tato, Ange, and Nkambou Roger. 2018. Improving Adam Optimizer. Available online: https://openreview.net/pdf?id=HJfpZq1DM (accessed on 22 October 2020).

Torres, Francisco, Leticia Neira, and Luis Torres-Treviño. 2015. Análisis Criminal Inteligente Usando Eventos Geográficos y Temporales con la Red Neuronal de Máxima Sensibilidad. *Daena: International Journal of Good Conscience* 10: 92–100.