ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY
February 2020, Vol 13(06), 696 – 711
DOI: 10.17485/ijst/2020/v13i06/149853,

# Prediction of Recidivism in Thefts and Burglaries Using Machine Learning

**Fredy Humberto Troncoso Espinosa***

Department of Industrial Engineering, Faculty of Engineering, Universidad del Bío-Bío, Concepción 4030000, Chile

**Article Type:** Article

**Article Citation:** Fredy Humberto Troncoso Espinosa. Prediction of recidivism in thefts and burglaries using machine learning. *Indian Journal of Science and Technology*. 2020; 13(06),696-711.DOI:10.17485/ijst/2020/v013i06/149853

**Received date:** January 9, 2020

**Accepted date:** January 23, 2020

***Author for correspondence:**
Fredy Humberto Troncoso Espinosa ftroncos@ubiobio.cl Department of Industrial Engineering, Faculty of Engineering, Universidad del Bío-Bío, Concepción 4030000, Chile

## Abstract

**Background/objectives:** Theft and burglary are two crimes against property that have a great social impact. Their prevention drastically lowers victimization rates and the feeling of insecurity in the population. The objective of this investigation is to obtain an index that allows the prediction of repeat offenses by criminals in these types of crimes, in order to support decision-making with respect to preventative actions. **Methodology:** In order to obtain the index, a group of machines learning was trained, with information provided by the Criminal Analysis and Investigative Focus System (CAIFS) from the Regional Public Prosecutor's Office in Biobío, Chile. The information provided was from thefts and burglaries committed between 2012 and 2017 in the city of Concepción. **Findings/application:** The results show a characterization of repeat offenders in these types of crime and a recurrence index that allows for a greater assertiveness in the prediction of recidivism than the method that is currently being used.

**Keywords:** Recidivism, Machine Learning, Criminal Analysis, Theft, Burglary.

## 1. Introduction

The Chilean Public Prosecutor's Office is an autonomous and hierarchical body whose role is to exclusively direct the investigation of constituent criminal acts and to exercise a public criminal action in the manner provided by law. In order to improve the quality of service and the results of criminal prosecution, there is a Criminal Analysis and Investigative Focus System (CAIFS) whose mission is to strengthen the criminal prosecution of crimes against property and other crimes with high social connotations.

In [1–3] order to guide the investigative work focused on reducing the number of high social connotation crimes such as crimes against property, at the beginning of each year the CAIFS of the Regional Public Prosecutor's Office in Biobío develops a ranking of the criminals with the highest probability of recidivism during that year. This ranking allows

the investigative resources to be focused on the criminals who have a greater likelihood of recidivism.

To develop this ranking, the number of criminal cases that an individual had during the last year is used. It is considered that the higher the number of criminal cases, the greater the probability of recidivism by the criminal.

Although the number of criminal cases results in being important to recidivism, it does not solely depend on this factor, but also on another set of factors associated with the personal and sociodemographic characteristics of the individual, and characteristics associated with the historical criminal conduct of the individual.

The Chilean Public Prosecutor's Office currently records information about individuals who have committed crimes, which is mainly related to the associated aspects of their historical criminal conduct. The objective of this investigation is to produce an index that measures an individual's degree of recidivism considering their criminal history so that the Regional Public Prosecutor's Office in Biobío can effectively focus on criminal prosecution work and increase the effectiveness of the intended resources for this work.

Given the complexity of the recidivism phenomenon and the difficulty of establishing a clear pattern that defines and predicts it, the recidivism index will be obtained by using data mining models. Data mining can extract patterns present in the data that are not evident [4] in order to later use them for prediction, for example, the conduct of an individual.

## 2. Machine Learning in Crime Prediction

Modern society produces a rapidly growing quantity of data, generating new problems and possibilities [5]. This has meant an important challenge for law application, where information plays a relevant role for analysts, who must carry out a precise and efficient investigation of crimes [6–7]. When studying the habitual actions of criminal, large quantities of data are generated, which makes it difficult to detect crime through traditional data analysis [8]. As a result of the aforementioned, the studio of crime turns out to be one of the most appropriate fields for the application of data mining. The knowledge that it could produce to investigate, curb and prevent crime makes it a useful tool to support police work [9–10].

The main and frequently applied data mining techniques for crime analysis are as follows: Entity extraction, Cluster Analysis, Rules of association, Classification, and Analysis of Social Networks [11–12]. In this investigation, the Classification technique will be used to obtain the recidivism index.

Within this technique, there is a wide variety of models that are also known as machines learning [13]. The machines learning learn the general hidden pattern in the data and then use it to generate a new prediction. The prediction consists of assigning a registry or observation to a previously defined category or class, such as Repeat or Not Repeat. The prediction that the machines learning delivers is a value between zero and one, known as confidence. This value will be used as a recidivism indicator, where the value close to one will be related to a high probability of recidivism and a value close to zero will be related to a low probability of recidivism.

Decision Tree: Corresponds to a network diagram with a tree structure in hierarchical form. It is very useful for finding structures in high dimensional spaces and in problems that mix categorical and numerical data. In this model, each node denotes an attribute on which a test is carried out. The branches derived from a node represent the categories of the attribute that show a test result. Each terminal node or leaf represents the class that a record or observation is assigned to.

Naive Bayes: Based on estimating the probability of belonging to a class, through the estimation of conditional probabilities, using the Bayes Theorem. This determines the conditional probability that a record belongs to a class, given a set of variables that characterizes it. This machine learning assumes the independence between the variables for the classification.

Neural Network: A neural network is a set of interconnected nodes, where the input nodes, known as the input layer, represent the set of variables considered to predict the class of a record. The output nodes, known as the output layer, represent the number of classes considered in the classification of each record. The input layer and the output layer are connected via a set of interconnected nodes, known as the hidden layer. The hidden layer processes the information by using a set of weightings, assigned to each connection. The Neural Network learning process consists of assigning these weightings through the training set. The neural networks have high predictive power and tolerance of out of range data, however, the interpretation of their parameters, represented by the weightings, is difficult to achieve.

When applying these methods, the use of decision trees to detect suspicious emails related to criminal activity stands out [14]. In [15] the classification, the algorithm produces an assertiveness above 95%. Decision tree, Neural Networks, and Naive Bayes have also been applied, comparing their performance in the detection of companies that issue fraudulent financial statements and when identifying factors associated with them. In Ref. [16], Yu et al. used them in the temporary space of occurrence prediction of a future residential burglary, in a city in the northeast of the United States. In Ref. [17], Bhowmik used a decision tree and Naive Bayes to identify the pattern of car insurance fraud and to predict its occurrence. In Ref. [18], Fuller et al. used machines learning based on information fusion for automatic lie detection in 371 texts of various crimes, obtaining a 73.46% accuracy rate for neural networks and 71.6% for decision tree. From the data set analysis of 2500 emails, Pandey and Ravi [19] used neural networks for the effective prediction in the detection of phishing emails. In Ref. [20], Ang and Goh explored and compared the performance of neural networks and decision tree for the prediction of crime in juvenile offenses and the identification of correlated risk factors in adolescents, obtaining accuracy rates that exceed 95%. In Ref. [21], Tollenaar and Van der Heijden implemented neural networks and other data mining techniques to produce a model that predicts three types of criminal recidivism: general recidivism, violent recidivism, and sexual recidivism. In Ref. [22], Iqbal et al. exhibited a comparison between two classification algorithms, which was used to predict the crime category for different locations in the United States. To achieve this, socioeconomic data from a census and application data from the 1990 law were used, obtaining an accuracy rate of 83.5% for decision tree and 66.4% for Naive Bayes. In Ref. [23], Wang et al. proposed a decision tree algorithm based on the Maclaurin-Priority

Value First method for the forensic study of computer crimes, achieving a performance that exceeds the previously obtained results. In Ref. [24], Rumi et al. extract the dynamic features from check-ins of foursquare users like habits or routines to predict recidivism in theft, drug offense, assault, fraud, unlawful entry, and traffic offense, using among others methods neural network.

However, if the nature of the dataset is linear or linearizable, classical statistical models like Logistic Regression or Linear Discriminant Analysis can be enough to explain the recidivism [25–27].

In the literature review, it can be seen that Decision Tree, Naive Bayes, and Neural Networks are commonly applied in contexts related to the commission of various types of crime, and for this reason, it was decided to apply them in this investigation.

## 3. Material and Methods

For the identification of the set of relevant variables and the training of the models that will obtain the recidivism index, the Knowledge Discovery in Databases (KDD) was used [28]. This methodology is composed of a sequence of stages whose main purpose is the extraction of hidden knowledge within databases [29]. This methodology is characterized as being a non-trivial discovery process of knowledge and potentially useful information, within the data contained in an information repository. It is not an automatic process; it is an iterative process that thoroughly explores large volumes of data in order to determine patterns. This methodology is comprised of five stages. The first stage is the data selection, where the data sources and type of information to be used is determined. The relevant data for the analysis are extracted from the data source(s). In order to select the information, the variables that are present in the production process must be fully known, and the variable to be predicted must be clearly established. The second stage consists of a pre-processing of the database, in order to have more trustworthy information, which brings greater value to the prediction. This stage incorporates the analysis of missing data, inconsistent data, and the analysis of out-of-range data. The third stage consists of the transformation and selection of variables. The transformation of variables includes any process that modifies the shape of the data. The variables are transformed to produce new variables, which enrich the information that will be used to train the model so that it has better predictive power. After the transformation, the process proceeds, through techniques that evaluate each variable's predictive power, so as to identify those that best predict the variable of interest [30].

With this, it is possible to produce simpler models that explain the problem better. The fourth stage is data mining. In this stage, the techniques that extract the relevant pattern from the data are used, in this investigation specifically, the technique of classification is used. The fifth stage of the process is the predictive performance evaluation. This consists of the evaluation of the machine learning used. For this, the results obtained from applying the test set to the trained model are used. The results summarize a matrix called Matrix and Confusion. For example, if two classes are considered, class 1 that identifies a repeat offender and class 0 that identifies a non-repeat offender, the Confusion Matrix [31] will have the shape that is shown in Table 1.

**TABLE 1.** Confusion matrix

| Classes | | Real value | |
|---|---|---|---|
| | | **1** | **0** |
| Predicted value | 1 | TP | FP |
| | 0 | FN | TN |

In the Confusion Matrix shown in Table 1, TP represents the class 1 elements that were correctly predicted by the model or true positive rate and FN represents the class 1 elements that were incorrectly predicted by the model or false positive rate. TN represents the class 0 elements that were correctly predicted by the model or true negative rate and FP represents the class 0 elements that were incorrectly predicted by the model or false positive rate.

The Confusion Matrix obtains the following performance measures [32].

Accuracy: represents the total proportion of predictions that were correctly classified.

$$Accuracy = \left(TP + TN\right) / \left(TP + FP + FN + TN\right)$$

Recall: is the percentage of observations that belong to class 1 and were correctly classified by the model.

$$Recall = TP / \left(TP + FN\right)$$

Precision: is the percentage of correctly classified elements as class 1 of the elements classified as class 1.

$$Precision = \ TP / \left(TP + FP\right)$$

Accuracy measures the general performance of the model, whilst Recall and Precision obtain a more exact measurement of how the model specifically predicts the class of interest. As mentioned previously, the prediction that the classifier delivers a value between zero and one. This value will be used as a recidivism index, where a value close to one is related to a high probability of recidivism and a value close to zero to a low probability of recidivism. The index, ordered in a descending form, produces a recidivism ranking, which allows criminal prosecution work to effectively focus on those individuals with the highest index.

# 4. KDD Process Application to Obtain the Recidivism Index

## 4.1. Dataset Selection

The data provided by the CAIFS from the Biobío Public Prosecutor's office corresponds to 12,222 burglaries and thefts, recorded between 2012 and 2017. Each record is characterized by 7 attributes as shown in Table 2.

**TABLE 2.** Attributes of the original database

| Attribute | Description |
|---|---|
| RUT (national identity N°.) | Corresponds to the unique identifier for each offender that has committed a crime. |
| Crime | Corresponds to the name of the crime committed by the offender. |
| RUC (criminal case N°.) | The unique criminal case code assigned to the offender. |
| Date of crime | The date on which the crime was committed. |
| Criminal convictions | Indicate if the person has had any convictions. |
| Gender | Sex of the accused. |
| Date of birth | Represents the date of birth of the offender. |

In Table 2, the unique identifier for each record is composed of the RUT, RUC, and Date of Crime attributes.

## 4.2. Pre-processing of Dataset

In this stage, each attribute was analyzed to identify atypical and missing data. The identification of atypical data was carried out for each attribute through using the three-sigma rule [33]. The out-of-range data were studied, some were left out and others were replaced. The replacement of atypical and missing data was performed using the Hot Deck technique, which replaces the missing value of an attribute with that of an attribute from a similar record [34]. Through the analysis of the date of birth attribute, individuals under the age of fourteen were identified, who are not within the competency of the Public Prosecutor's Office. The records with individuals under the age of fourteen were eliminated.

## 4.3. Attribute Transformation

The RUT is a unique identifier for each offender with whom it is required to obtain a recidivism index. Since the database does not show a single RUT per record, it is essential to transform the database into the required format. This transformation permitted the creation of 11 new attributes from the original five (RUT, Crime, RUC, Date of Crime, and Date of Birth). The attributes were created using the Recency Frequency Monetary (RFM) technique, which is widely used in customer segmentation [35]. In this technique, Recency represents the elapsed time since a customer's last purchase, Frequency represents their buying frequency, and Monetary represents the total amount of the purchases. In the context of this investigation, the Recency concept is represented by the elapsed time since an individual last committed a crime. The Frequency concept is represented by the average number of days between an offender's crimes and by the average number of crimes per year. The Monetary concept is represented by the offender's total amount of crimes, by the number of crimes over the last year and by the number of associated criminal cases. Other attributes of interest that are unrelated to this technique were also created. All of the created attributes are shown in Table 3.

**TABLE 3.** Description of the created attributes

| | Attribute | Description |
|---|---|---|
| A | RUT (national identity N°.) | Corresponds to the unique identifier for each offender that has committed a crime. |
| B | Age at last crime | Corresponds to the age that an offender had when they committed the last crime recorded in the database. |
| C | Age during last year | Corresponds to the age during the last year of records from the database. |
| D | Days since last crime | The number of days that have elapsed since the last crime committed by an offender. |
| E | Quantity of crimes | Represents the number of crimes that an individual has committed. |
| F | Quantity of crimes in the last period | Counts all of the crimes that an offender has had during the last year in the database. |
| G | Quantity of RUC (criminal cases) | Represents the number of criminal cases in which an offender has participated. |
| H | Average number of accomplices per RUC | Represents the average number of individuals with whom an offender participates in criminal cases. |
| I | Crime accomplices | Is the number of different individuals with whom an offender is linked to in the different criminal cases in which they have participated. |
| J | Average number of days between crimes | Corresponds to the average difference in days between one crime and another. The investigation only considers repeat offenders. |
| K | Average crimes per year | Corresponds to the average number of crimes committed per year. |
| L | Quantity of RUC in the last period | Number of criminal cases recorded by an individual during the last period. |
| M | Gender | Sex of the offender. |
| N | Conviction record | Indicates if the person has had a conviction. |
| O | Recidivism | Indicates if an offender has committed a crime against property in the year following the last year of study. |

After the attributes were created, the correlation matrix shown in Table 4 was produced. The correlation identifies if there is a linear dependence between the created numerical attributes. A high correlation between two attributes implies that these two attributes will explain a phenomenon in a similar way, therefore for the training of the model, one of those attributes will be left with a correlation index greater than or equal to ±0.9. The attributes that were eliminated were as follows: Age during last year (C), Quantity of crimes in the last period (F), and Average crimes per year (K).

## 4.4. Attribute Selection

After the database transformation, the most important attributes for the prediction of recidivism are selected. The importance of each attribute is defined through its predictive power of recidivism. The attributes with the greatest predictive power are considered for the training and testing of the models. To measure the predictive power of each attribute, the Chi-squared statistic was used. This statistic is a measure of the dependence between any attribute and the attribute to be predicted. The greater the value of the statistic, the

**TABLE 4.** Correlation matrix between created attributes

| Attributes | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B | 1 | 0.994 | −0.233 | 0.101 | 0.136 | 0.226 | −0.154 | −0.109 | 0.185 | 0.101 | 0.188 |
| C | 0.994 | 1 | −0.129 | 0.078 | 0.072 | 0.196 | −0.148 | −0.114 | 0.154 | 0.078 | 0.128 |
| D | −0.233 | −0.129 | 1 | −0.235 | −0.628 | −0.325 | 0.063 | −0.042 | −0.298 | −0.235 | −0.613 |
| E | 0.101 | 0.078 | −0.235 | 1 | 0.403 | 0.856 | 0.009 | 0.166 | −0.074 | 1 | 0.434 |
| F | 0.136 | 0.072 | −0.628 | 0.403 | 1 | 0.409 | −0.010 | 0.077 | −0.004 | 0.403 | 0.926 |
| G | 0.226 | 0.196 | −0.325 | 0.856 | 0.409 | 1 | −0.127 | 0.103 | 0.081 | 0.856 | 0.523 |
| H | −0.154 | −0.148 | 0.063 | 0.009 | −0.010 | −0.127 | 1 | 0.825 | −0.101 | 0.009 | −0.06 |
| I | −0.109 | −0.114 | −0.042 | 0.166 | 0.077 | 0.103 | 0.825 | 1 | −0.042 | 0.166 | 0.06 |
| J | 0.185 | 0.154 | −0.298 | −0.074 | −0.004 | 0.081 | −0.101 | −0.042 | 1 | −0.074 | 0.063 |
| K | 0.101 | 0.078 | −0.235 | 1 | 0.403 | 0.856 | 0.009 | 0.166 | −0.074 | 1 | 0.434 |
| L | 0.188 | 0.128 | −0.613 | 0.434 | 0.926 | 0.523 | −0.06 | 0.06 | 0.063 | 0.434 | 1 |

greater the dependence between these attributes, therefore, an attribute will have greater importance over the attribute to be predicted. This method is only applicable to categorical attributes, so for its application, it was necessary to categorize each of the numerical attributes, choosing the number of categories that maximized its dependence on the attribute to be predicted. This categorization and the predictive power of each attribute are shown below in Table 5.

## 4.5. Data Mining

Three machines learning mentioned in the previous chapter were trained and tested, using the algorithms incorporated in the Scikit-Learn library of Python programming language [36]: Decision Tree Classifier, Naive Bayes, and Multilayer Perceptron. In their training, the respective parameters were adjusted to optimize their predictive performance. In order to train them, it was necessary to balance the data, so that the number of records from both classes was equal. Through this balancing, the models impartially learn the general pattern that characterizes the variable to be predicted, towards a majority class. The data balance was carried out using the (Synthetic Minority Oversampling Technique) which produces new records of the minority class within its neighborhood [37]. For the training process, validation and obtainment of parameters from the hold-out cross-validation technique for time series were used [38–39]. This technique considers the temporal division of data in a training and test set as shown in Figure 1.

Figure 1 shows that the training set formed by the records from 2012 to 2016 is divided into a training (2012–2015) and validation (2016) subset for the adjustment of parameters.

**TABLE 5.** Categories by attribute and predictive power according to the chi-squared statistic

| Attribute | Categorization | Statistical value Chi-squared |
|---|---|---|
| Quantity of RUC (criminal cases) | 5 categories: [1,2] – [3,4] – [5,6] – [7,8] – >=9 | 143.709 |
| Quantity of RUC in the last period | 4 categories: Does not have-[1,2]-[3,4]->4 | 96.073 |
| Days since last crime | 5 categories: [1, 219] – [220,438] – [439,657] – [658,876] – >877 | 82.734 |
| Average number of days between crimes | 8 categories: [1,20] – [21,40] – [41,60] – [61,90] – [91,150] – [151,300] – [301,530] – >=531 | 82.006 |
| Age at last crime | 6 categories: [14,24]-[25,34]-[35,44]-[45,54]-[55,64]->64 | 47.383 |
| Quantity of crimes | 4 categories: [1,5] – [6,10] – [11,15] – >=16 | 44.451 |
| Conviction record | 2 categories: YES – NO | 6.807 |
| Average number of accomplices per RUC | 5 categories: [0,1] – [1,2] – [2,3] – [3,4] – >4 | 3.426 |
| Average number of days between crimes | 5 categories: [0,1] – [1,2] – [2,3] – [3,4] – >4 | 1.520 |
| Gender | 2 categories: Feminine, Masculine | 0.414 |

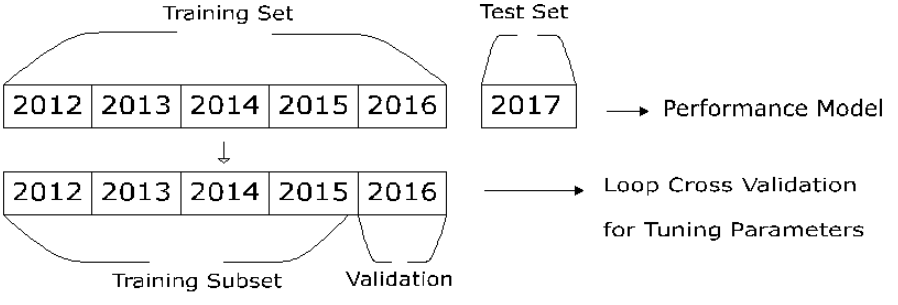


**FIGURE 1.** Hold-out cross-validation technique for time series.

This parameter adjustment is performed via an iterative process until the optimal parameters for the model are found. With the parameters adjusted, the model is trained with the training set and the recidivism or non-recidivism is predicted for 2017. Using this prediction, the predictive performance of each model is obtained. In the following chapter, the results gathered from the prediction of models are analyzed.

# 5. Analysis of Results

After testing various attribute combinations, taking the predictive powers from Table 5 as a reference, the best performance in each training model was achieved by considering the five attributes with the greatest predictive power: Quantity of RUC (Criminal Cases),

Quantity of RUC in the last period, Days since last crime, Average number of days between crimes, and Age at last crime. The best predictive power reached by each model is detailed in Table 6.

**TABLE 6.** Predictive performance of the machines learning used

| | Classification model | | |
|---|---|---|---|
| **Performance measure** | **Decision tree classifier** | **Naive Bayes** | **Multilayer perceptron** |
| Accuracy | 59% | 76% | 71% |
| Recall | 87% | 66% | 73% |
| Precision | 27% | 37% | 32% |

Table 6 it can be observed that the best overall performance is obtained by Naive Bayes with a 76% accuracy rate, followed by the neural network Multilayer Perceptron with 71% and finally the Decision Tree Classifier with an accuracy rate of 59%. In spite of these general results, we can observe that in the specific prediction of the individual recidivism class, the models demonstrate a greater Recall than Precision. This difference indicates that generally the models better predict the individuals who really are repeat offenders and who are not.

From the working point of view of the criminal analysts at the CAIFS from the Biobío Public Prosecutor's Office, there is a greater error cost when predicting whether an offender will actually re-offend (type I error) than there is when predicting if an offender will not re-offend (type II error). The cost of type 1 errors is social as preventative measures will not be taken with an individual than will actually re-offend. The cost of type II errors will be the time and resources allocated to preventative measures with an individual who will not re-offend.

In order to define the best model, the assertiveness of the recidivism ranking will be considered as a defining measure. For this, the graph called a Lift Chart is used, which shows the true positive rate variation (correctly classified repeat offenders) according to the percentage variation of the individuals within the ranking. Figure 2 shows the Gain Chart for each model.

In the graph shown in Figure 2, a better performance implies a curve that is closer to the point (0.1), which is associated with the greater number of individuals classified in the highest positions of the ranking. It can be observed that the best performance is achieved by the Multi-Layer Perceptron, followed by Naive Bayes and finally Decision Tree Classifier. All of the models show a superior performance than the method used by the CAIFS from the Biobío Public Prosecutor's Office, which consists of prioritizing offenders based on the number of criminal cases over the previous year. This method only works well with individuals with a high number of criminal cases, where the curve is very close to the origin. This can be seen in greater detail in Figure 3, which shows the Decision Tree produced through processing the results provided by the Decision Tree Classifier algorithm

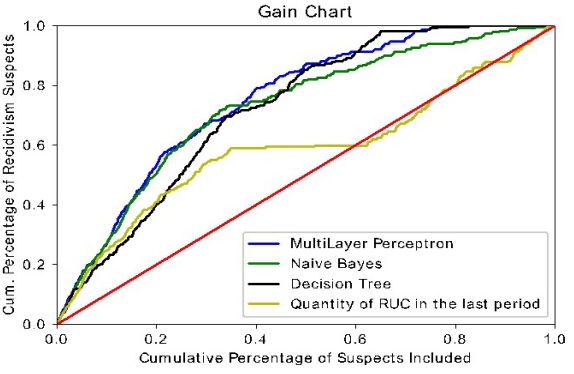


**FIGURE 2.** Machines Learning Performance and the current method used by the Public Prosecutor's Office.

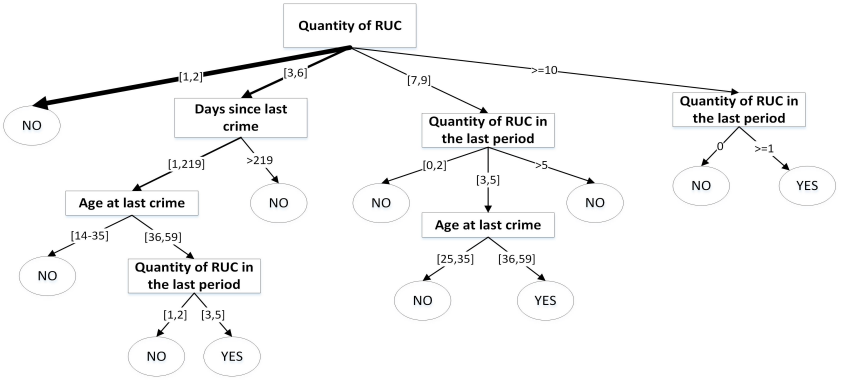


**FIGURE 3.** Decision tree created to characterize the recidivism in thefts and burglaries.

The decision tree in Figure 3 shows that in general, it is possible to characterize any repeat offender by using four attributes: Total number of cases, Number of cases in the last period, Days since last crime, and Age at last crime. When the offenders record two or more criminal cases associated with crimes against property they do not re-offend. When the number of cases recorded by an offender is high, with at least one case recorded in the last period, the individual will re-offend in the next period. These two branches validate the criteria currently being used by the CAIFS from the Biobío Public Prosecutor's Office, which can be simple defined as: the greater (lesser) the number of recorded cases, the greater (lesser) the probability of recidivism will be, above all for a low number of cases.

The situation becomes more complicated for an intermediate number of crimes, which is shown in the central branches of the tree. These branches show the following general rules for recidivism:

- If an offender has between three and six criminal cases, less than 220 days have passed since the last crime, his/her age is between 36 and 59, and of the criminal cases between three and five were committed during the last year, the individual will re-offend.

- If an offender has between seven and nine recorded criminal cases, of which between three and five were during the last year and furthermore the age of the individual is between 36 and 56, he/she will re-offend in the following period.

Although the decision tree obtains general rules that make an individual have a high probability of being a repeat offender, it is necessary to apply the neural network as a model to generate the recidivism index that decides the actions to be taken with an offender.

## 6. Discussion

The presence of unequal or unbalanced classes is a problem that regularly occurs in some issues when application techniques are applied [40] such as: Medical diagnosis (90% without the disease, 10% with the disease); e-commerce (99% non-purchase, 1% purchase); Cyber security (more than 99.99% of connections are not attacks). This problem is also present in the theme of this investigation, where 84% of the total records belong to the next period repeat offender class and 16% to the next period non-repeat offender class. The use of some aids in order to improve the results, such as the minority class multiplication technique, was used in this investigation. However, other techniques allow for the incorporation of type I and II error costs [41]. In this investigation, there was no accurate knowledge of these costs, only of the general difference between them. In a study carried out by De la Fuente and Mejías [1], a set of factors that define crime were determined, including age and gender. In this investigation, gender did not have any great relevance, but it was found that age does play a relevant role in recidivism.

The best predictive performance achieved by the models was 76% where the remaining 24% is attributable to other important factors as explained by De la Fuente and Mejías [1] that consider educational level, poverty and employment, which are not recorded by the Public Prosecutor's Office. Another important factor in the understanding of recidivism is the nature and distribution of available opportunities that criminals take in order to commit their crimes.

Repeat offenders are expected to be found within the first places in the generated ranking. However, there are individuals who are in the first places and have not actually reoffended. This error is not necessarily attributable to the neural network performance because there are two other probable causes: firstly, certain individuals could have been convicted and this is not recorded in the database, and secondly, that certain individuals have reoffended and that this fact has not been recorded. Many of the cases that the CAIFS from the Biobío Public Prosecutor's Office records do not have an individual associated with them.

## 7. Conclusion

The recidivism index proposed in this investigation is based on a set of general information managed by the CAIFS from the Biobío Public Prosecutor's Office regarding crimes. Based on this information, a model that retrieves the pattern that characterizes repeat offenders could be trained and used as a recidivism predictor from one period to another.

The best overall performance machine learning is Naive Bayes. However, the usefulness of this index is through the recidivism ranking and in said ranking the model that concentrates the repeat offenders higher in the rankings is the neural network Multi-Layer Perceptron.

The five attributes that maximize the predictive performance of the trained models were as follows: Quantity of RUC (Criminal Cases), Quantity of RUC in the last period, Days since last crime, Average number of days between crimes, and Age at last crime. These attributes coincide with those considered by the decision tree, only differing in the Average number of days between crimes. The definition of these attributes is relevant since they are easily understandable for analysts from the Public Prosecutor's Office and produces a characterization based on easily accessible information.

The RFM technique is especially useful attribute generation and each one of these concepts is present within the most important attributes. The Recency concept is present in the Days since last crime attribute, the Frequency concept is present in the Average number of days between crimes attribute, and the Monetary concept is present in the Quantity of RUC (Criminal Cases) and Quantity of RUC in the last period attributes.

The Age at last crime stands out since the literature indicates that age is a determining factor in crime.

These two branches validate the criteria that are currently used by the Public Prosecutor's Office, which can be simply defined as, the greater (lesser) the number of recorded cases, the greater (lesser) the recidivism probability will be.

From the decision tree, it can be seen that for individuals with a high number of criminal cases or a low number of criminal cases, the fact is that: the greater (lesser) number of recorded cases, the greater (lesser) the recidivism probability will be. However, when the values of the cases associated with individuals are at intermediate levels, there are other factors present and the recidivism pattern is not evident, so therefore the use of machine learning resulted in being the appropriate way of extracting the pattern and obtaining the recidivism index.

The generation of a recidivism ranking is an effective tool given that it shows high assertiveness when identifying repeat offenders. This high assertiveness improves the allocation of public resources related to the implementation of preventive recidivism measures by the Public Prosecutor's Office.

In order to improve the neural network performance in future work, considering two elements included in the discussions is proposed: try different techniques for balancing the data and consider other attributes that could be available, such as educational level.

## Acknowledgement

The author acknowledges the support given by the Macro Facultad de Ingeniería of Universidad del Bío-Bío through i+T node during the execution of this project. Also acknowledges to the Criminal Analysis Unit of the Public Prosecutor's Office of Región del Biobío-Chile by the dataset provided under an Internship Agreement.

# References


1. De la Fuente H, Mejías C. Análisis econométrico de los determinantes de la criminalidad en Chile. *Política Criminal*. 2011; 6(11), 192–208. http://dx.doi.org/10.4067/S0718-33992011000100007
2. Archwamety T, Katsiyannis A. Factors related to recidivism among delinquent females at a state correctional facility. *Journal of Child and Family Studies*. 1998; 7(1), 59–67. https://doi.org/10.1023/A:1022960013342
3. Katsiyannis A, Archwamety T. Factors related to recidivism among delinquent youths in a state correctional facility. *Journal of Child and Family Studies*. 1997; 6(1), 43–55. https://doi.org/10.1023/A:1025068623167
4. Han J, Kamber K, Pei J. Data mining: concepts and techniques. 3rd edn. The Morgan Kaufmann. 2011. https://www.academia.edu/download/43034828/Data_Mining_Concepts_And_Techniques_3rd_Edition.pdf.
5. Cocx T, Kosters W. A distance measure for determining similarity between criminal investigations. In: Industrial conference on data mining. Springer, Berlin, Heidelberg. 2006; 511–525. https://doi.org/10.1007/11790853_40
6. Chen H, Chung W, Xu J, Wang G, Qin Y, Chau M. Crime data mining: a general framework and some examples. *Computer*. 2004; 37(4), 50–56. https://doi.org/10.1109/MC.2004.1297301
7. De BruinJ, Cocx T, Kosters W, Laros J, Kok J. Data mining approaches to criminal career analysis. In: Sixth international conference on data mining (ICDM'06). 2006; 171–177. https://doi.org/10.1109/ICDM.2006.47
8. Thongtae P, Srisuk S. An analysis of data mining applications in crime domain. In: 2008 IEEE 8th international conference on computer and information technology workshops. 2008; 122–126. https://doi.org/10.1109/CIT.2008.Workshops.80
9. Keyvanpour M, Javideh M, Ebrahimi M. Detecting and investigating crime by means of data mining: a general crime matching framework. *Procedia Computer Science*. 2011; 3, 872–880. https://doi.org/10.1016/j.procs.2010.12.143
10. Nath S. Crime pattern detection using data mining. In: 2006 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology workshops. 2006; 41–44. https://doi.org/10.1109/WI-IATW.2006.55
11. Hassani H, Huang X, Silva E, Ghodsi M. A review of data mining applications in crime. statistical analysis and data mining: *The ASA Data Science Journal*. 2016; 9(3), 139–154. https://doi.org/10.1002/sam.11312
12. Troncoso F, WeberR. A novel approach to detect associations in criminal networks. *Decision Support Systems*. 2020; 128, 113–159. https://doi.org/10.1016/j.dss.2019.113159
13. Kotsiantis S, Zaharakis I, Pintelas P. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*. 2006; 26(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3
14. Appavu S, Pandian M, Rajaram R. Association rule mining for suspicious email detection: a data mining approach. In: 2007 IEEE intelligence and security informatics. 2007; 316–323. https://doi.org/10.1109/ISI.2007.379491
15. Kirkos E, Spathis C, Manolopoulos Y. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*. 2007; 32(4), 995–1003. https://doi.org/10.1016/j.eswa.2006.02.016
16. Yu C, Ward M, Morabito M, Ding W. Crime forecasting using data mining techniques. In: 2011 IEEE 11th international conference on data mining workshops. 2011; 779–786. https://doi.org/10.1109/ICDMW.2011.56

17. Bhowmik R. Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*. 2011; 2(4), 156–162. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.3602&rep=rep1&type=pdf
18. Fuller C, Biros D, Delen D. An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications*. 2011. 38(7), 8392–8398. https://doi.org/10.1016/j.eswa.2011.01.032
19. Pandey M, Ravi V. Detecting phishing e-mails using text and data mining. In: 2012 IEEE international conference on computational intelligence and computing research. 2012; 1–6. https://doi.org/10.1109/ICCIC.2012.6510259
20. Ang R, Goh D. Predicting juvenile offending: a comparison of data mining methods. *International Journal of Offender Therapy and Comparative Criminology*. 2013; 57(2), 191–207. https://doi.org/10.1177/0306624X11431132
21. Tollenaar N, Van der Heijden P. Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society*. 2013; 176(2), 565–584. https://doi.org/10.1111/j.1467-985X.2012.01056.x
22. Iqbal R, Murad M, Mustapha A, Panahy P, Khanahmadliravi N. An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*. 2013; 6(3), 4219–4225. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.921.1753&rep=rep1&type=pdf
23. Wang Y, Peng X, Bian J. Computer crime forensics based on improved decision tree algorithm. *Journal of Networks*. 2014; 9(4), 1005–1011. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.600.5169&rep=rep1&type=pdf#page=203.
24. Rumi S, K Deng K, Salim F. Crime event prediction with dynamic features. *EPJ Data Science*. 2018; 7(1), 43. https://doi.org/10.1140/epjds/s13688-018-0171-7
25. Dressel J, Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*. 2018; 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580
26. Ozkan T. Predicting recidivism through machine learning. Doctoral dissertation. 2017. http://hdl.handle.net/10735.1/5405
27. Tollenaar N, van der Heijden P. Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PloS One*. 2019; 14(3), e0213245. https://doi.org/10.1371/journal.pone.0213245
28. Fayyad U, Piatetsky G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. Association for the Advancement of Artificial Intelligence. 1996; 82–88. https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf
29. Brachman R, Anand T. The process of knowledge discovery in databases. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence. 1996; 37–57. https://dl.acm.org/citation.cfm?id=257944
30. Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003; 3: 1157–1182. http://www.jmlr.org/papers/v3/guyon03a.html
31. Hay A. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*. 1988; 9(8), 1395–1398. https://doi.org/10.1080/01431168808954945
32. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 2009; 45(4),427-437. https://doi.org/10.1016/j.ipm.2009.03.002.
33. Lakshminarayan K, Harp S, Goldman R, Samad T. Imputation of missing data using machine learning techniques. KDD. 1996; 140–145. https://www.aaai.org/Papers/KDD/1996/KDD96-023.pdf

34. Andridge R, LittleR. A review of hot deck imputation for survey non-response. *International Statistical Review*. 2010; 78(1), 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x.
35. Hosseini S, Maleki A, Gholamian M. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*. 2010; 37(7), 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Vanderplas J. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. 2011; 12(Oct), 2825–2830. https://hal.inria.fr/hal-00650905
37. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002; 16, 321–357. https://doi.org/10.1613/jair.953
38. Tashman L. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*. 2000; 16(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0
39. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006; 7(1), 91. https://doi.org/10.1186/1471-2105-7-91
40. Chawla N, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*. 2004; 6(1), 1–6. https://doi.org/10.1145/1007730.1007733
41. Weiss G. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*. 2004; 6(1), 7–19. https://doi.org/10.1145/1007730.1007734.