University of Portsmouth

School of Computing Postgraduate Programme

MSc Degree in Forensic Information Technology

Predicting Crime Using Data Mining

By

Ginger Viviana Saltos Bernal UP707137

Supervisor: Mihaela Cocea

Project unit: PJS60P

2014

Abstract

The amount of crime data has increased dramatically, and therefore it has become a problem to analyse them in a short amount of time and make proactive decisions. For this reason, the project reviews the problem, and develops accurate models using non-used data mining techniques in the crime field such as locally weighted learning, linear regression and M5P decision tree, and the Apriori algorithm, which has been used in the field by researchers. The project executes four different experiments using each algorithm, and evaluates the performance of the models created, resulting that the M5P models have less error values, better correlation, and acceptable amount of time for the analysis than the other models.

Acknowledgements

This project would not have been made possible if not for my supervisor, Dr. Mihaela Cocea. Her guidance and invaluable contributions are the reasons for this project's success.

I would also like to thank Gary Burton from the University of Portsmouth's Institute of Cosmology and Gravitation for his support in this project.

Over the past year I have received support and encouragement from my whole family. I want to thank my mother, Fresia Bernal, and siblings for believe in me, and my wonderful grandparents Juan Bernal and Roxana Villacís for their love and support. Finally, I would like to thank my father, Carlos Saltos Villacís and grandmother Regina Pinos Piedra. They were always in my heart during this journey.

A special thanks to the government of Ecuador for giving me the opportunity to come to the United Kingdom and study the master.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Background

Crime is increasing all over the world; this is why governments invest so much money and time in the security field every year. Crime fighters around the world have been working on the analysis of data in order to develop strategies to combat crimes. However, the amount of data is growing so fast that, the analysis has become a problem, because it consumes too much time and effort making it an endless task (Malathi & Santhosh, 2011).

Data Mining is a new technology that can extract the knowledge from big amounts of data. This new technology has been used to reduce the time that crime analysts spend on finding similar characteristics between different types of crimes.

The aim of this project is to use data mining techniques to create models that could detect patterns, in order to help the police to predict crime and plan their resources.

1.2. Project Aims and Objectives

The analysis of the historical crime data is carried out to detect crime patterns and predict possible crimes so police departments can prevent them by planning their resources. Therefore, this project will analyse and preprocess historical data from a county of the United Kingdom and a city of Ecuador. It also reviews data mining techniques and methodologies that have been used to create the model and evaluate the results.

The aim of this project is to create crime prediction models for each country that will reduce the time that crime analysts spend on finding similar characteristics between different types of crime.

To achieve this aim, the following objectives have been identified:

- Review previous research related to crime data mining;
- Analyse the model process that have been used in this project; focusing on the algorithms used to create the models;
- Acquire datasets from the Police Departments of the UK;
- Acquire datasets from the Police Departments of the Ecuador;
- Design and implement models that could detect patterns and predict crimes using the WEKA software;
- Evaluate the outcomes and conclude over the results.

1.3. Project Deliverables

The models developed in WEKA are the main artifacts of this project, which will be available in the attached disc. These models and their measurements will be explained in the following chapters.

The second deliverable of this project is the report that includes the description of the methodologies and the explanation of the results of each experiment.

1.4. Constraints, Legal Issues and Ethical Considerations

Although the aims of the project are to detect patterns, and predict future crimes so that police officers can avoid them, it needs to be noted that it can also help criminals to predict where the police is making rounds, to avoid them and still commit the crimes. To avoid this, the models will only be on the attached CD.

There are several constraints on the execution of this project.

- The size of the dataset from the UK is too big to be handled in a common computer; therefore, the analysis will be executed in the supercomputer of the Institute of Cosmology and Gravitation (ICG).
- The amount of time to develop the models for both countries is short, considering the size of the datasets and, the time that takes to analyse it. To overcome this, a Gantt chart was created and included in Appendix “C”.
- To respect the citizens’ privacy, no personal data will be analysed. However, it might be interesting to include data such as gender and age, in order to analyse the trend of crimes.
- The crime data from Ecuador is not published, though the Defense Minister of this country signed a release form.

1.5. Report Structure

The report is structured in chapters that include all the aspects of the project from the background research to the conclusion of the project.

- **Chapter 1 – Introduction:** This chapter provides an overview of the project’s aims, objectives, and structure.
- **Chapter 2 – Criminal Data Mining Review:** An analysis of existing research related to data mining in the field of crime to detect patterns and predict crimes.
- **Chapter 3 – Data Mining Crime Requirements:** A brief description of the requirements needed to develop the models such as the Model Process (CRISP-DM), data pre-processing (Aggregation, dimension reduction, feature creation, variable transformation and, missing value) and algorithms (Locally weighted learning, Linear Regression, M5P and Association Rules).

- **Chapter 4 – Methodology:** A detailed description of every step followed, from the collection of the historical data, to its analysis.
- **Chapter 5 – Implementation:** This chapter shows all the experiments that were undertaken using WEKA software and their results.
- **Chapter 6 – Experiment Results:** This chapter contains a detailed description of the outcomes obtained from the experiments described on the previous chapter.
- **Chapter 7 – Evaluation and Conclusion:** It presents the conclusions drawn and the recommendations for future studies. Furthermore, it evaluates the management of the project and describes the difficulties for the execution of it.

2. Criminal Data Mining Review

Over time, the amount of data has increased dramatically; people have collected data from different sources, because they believe that more data will lead them to success. But its analysis and the extraction of the knowledge from it to make proactive decisions have become a problem.

Although in the past years computers have increased their speed, Database Management Systems (DBMS) have developed powerful and effective tools and, traditional statistics applications have helped analyse historical data, most of these tools have limitations such as:

- DBMS does not analyse the data, it allows the creation, storage and maintenance of it.
- Most of the traditional statistics applications do not analyse big amounts of data; moreover, and the data needs to be of good quality, which is not realistic for real life data (Du, 2010).

The following chapter will review some literature about data mining, focusing on the field of crime analysis and the algorithm used to develop different models.

2.1. Data Mining Definition

Fayyad, Piatetsky-Shapiro and Smyth (1996) stated that Data Mining is the analysis step in the process of Knowledge Discovery in Database (KDD). It finds patterns from large amounts of data by applying specifics algorithms. From these patterns, some useful information can be predicted to help organisations make future decisions.

Over the past years, many researchers have been using data mining technology in different areas such as shown in Table 2.1.

Table 2-1: Major Areas of Data Mining Research.

Major Areas	Authors
Marketing, sales and customer support.	Berry, & Linoff (1997) Tumpowsky (2013)
Medicine, Healthcare.	Brossette, Sprague, Hardin, Waites, Jones, & Moser (1998) Aljumah, Ahamad, & Siddiqui (2012).
Finance and Insurance.	Essays (2013). Guo (2003).
Agriculture.	Cunningham, & Holmes (1999).
Law enforcement	Skillicorn (2009). Chen, Chung, Xu, Wang, Qin, & Chau (2004).

The use of data mining in the crime field is less than other areas (Table 2-1). For this reason this report will focus on this field.

2.1.1. Challenges in Data Mining

An ideal data mining model should be accurate and efficient to detect and predict future crimes; however, regardless of the area of research, there are many challenges in Data Mining Technology. The most common are that the analysis is time consuming and the difficulty of finding patterns in big datasets (Du, 2010). In addition, big datasets often require distributed approaches (Hinman, 2013).

On the other hand, small datasets can also be challenging because they may not be a representative sample of the population, or its quality is not good enough to produce an accurate model (Hinman, 2013).

The data quality will depend on the different methods for its collection and how it is structured, making it a challenge for big or small datasets (Malathi & Santhosh, 2011). Pre-processing and processing data can take time depending on its size, errors, and missing values.

2.2. Data Mining Process

The data mining process consists of three main steps shown in Figure 2-1 and described on this section.

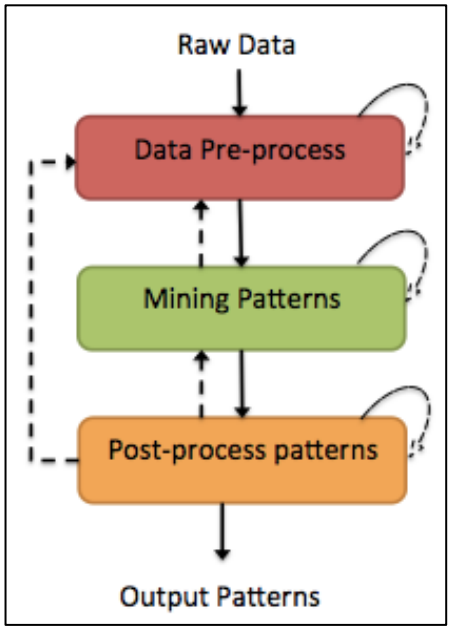


Figure 2-1: Stages of Data Mining

2.2.1. Data Pre-process

At this stage, raw data is prepared for the analysis by improving its quality and/or rearranging their attributes and records. For this reason, it usually takes between 60 and 80% of the project time (Jermyn, Dixon & Read, 1999).

While, according to Du (2010), there are seven tasks to prepare the data for the analysis: data aggregation, data sampling, dimension reduction and feature selection, feature creation, data discretization, variable transformation, and dealing with missing values, not all of these task were used by researchers, because it will depend on the datasets and the approach of each project.

Some datasets have many missing values for different reasons and the best way to handle them is by knowing why are they missing. For example Nath (2006), Dondanville, Zhang, and Lee (2007), and Yu,

Ward, Morabito, and Ding (2011) removed the records from the dataset. While Malathi and Santhosh (2011), and Hussain, Durairaj, and Farzana (2012) removed the records that were not important for the analysis and used a procedure to fill in the missing values.

For convenience or to improve the management of the data, some attributes in the dataset have to be transformed from one domain into another. For instance, Dondanville et al (2007) convert string values to categorical and Boolean. Nath (2006) extracts and transforms the data before checking for outliers and multiple abbreviations.

After this task, researchers usually aggregate the data to reduce the size of the datasets. Yu et al (2011) called this task data aggregation. Dondanville et al (2007) called it dimension reduction and feature selection, while Zubi and Mahmmud (2013) called it data reduction.

Before using any technique the researchers must have in mind that the execution of this stage will influence the outputs and performance of the data mining techniques. Therefore, a good data preparation is fundamental to achieve an excellent outcome from the data mining process (Jermyn et al, 1999).

2.2.2. Mining Patterns (Methods and Techniques)

Some research has been done in the field of crime, using different data mining techniques or a combination of them. For example, Jin, Wang, Xiao, and Pan (n.d.), Nath (2006) and, Zubi and Mahmmud (2013) use k-means to cluster the crime records and detect patterns. However, Malathi and Santhosh (2011) have showed that DBScan (Density-Based Spatial Clustering Application with Noise) technique outperforms the K-means technique results. Therefore, according to Malathi and Santhosh (2011) the K-means technique used by these authors is less effective than DBScan in terms of accuracy, speed, and efficiency.

To detect the unknown crime trend for coming years, Malathi and Santhosh (2011) used C4.5 decision tree.

Wang, Rudin, Wagner, and Sevieri (2013) use a mathematical approach, Series Finder, to analyse crime data. This algorithm is a supervised learning method that captures elements from the modus operandi (M.O.) to create a pattern and adapts itself to improve the M.O.

On the other hand, Chen et al (2004) propose a general framework called COPLINK, that combines data mining techniques with criminal and intelligence analysis. It identifies four categories that will be used (one or all of them) depending on the type of crime, because they believe that different techniques must be used on different crime investigations.

After clustering from historical data and detect patterns, future crimes can be clustered in these patterns and then, make decisions based on the results and the experiences of detectives (Nath, 2006). Another author explains, that prediction of crime can be done depending on the space, time and space and, social networks analysis. She suggests the Risk-terrain modelling (RTM) cluster, CrimeStat II software, and a Social Network Analysis (SNA) for each of the analysis (Bachner, 2013).

A comparison between several classifiers such as decision tree (J48), Support Vector Machine (SVM), Neural Network (Neural with 2 layer network), Naïve Bayes, one nearest-neighbour (1NN), and a location constrained variation, was done to forecast crime and, help with practical crime prevention solutions (Yu et al, 2011). The algorithms performed better on different experiments.

Past research has shown that some algorithms performed better than other ones depending on the datasets, and that most of the authors prefer the use of clustering and association rules. This is why this project will analyse the performance of algorithms that have not been used before such as Linear Regression (LR), Locally weighted learning (LWL), and M5P decision tree.

2.2.3. Post-Process Patterns

This stage is about analysing patterns, evaluate the extracted knowledge from the previous stage, and visualize it. Bruha and Famili (2000) explain in their paper that this stage is complementary to the previous one, because it helps refine the models. As Figure 2-1 illustrates, the analysis performed in this stage can take the project back to the first stage to rearrange the attributes and extract the knowledge again.

The models developed can be analysed through measures such as accuracy, speed of creating clusters and efficiency (Malathi & Santosh, 2011) or, plotting clusters (Nath, 2006). Other authors preferred to use crime experts, so they can examine the results, provide feedback and expert recommendations (Nath, 2006 and Wang et al, 2013).

2.3. Conclusion

The analysis of historical crime data is becoming a problem due to the amount of the data, its collection and quality, therefore it is taken too much time to analyse the data, search for patterns and make decision based on it. A number of data mining projects from different authors have been read and analysed in order to understand their approaches and learn more about this field.

Developing models that could help minimize the time on analysing crime, detecting patterns and predicting possible future crimes is the challenge that many researchers had to face and when considering all the data presented it is clear that the pre-processing of the data will affect directly the development of the models, as well as their post-processing. Therefore, it is important to decide what pre-processing techniques to use depending on the project's goals and approach.

When considering all the data presented, it is clear that the researchers' papers proposed different techniques to create a model for criminal data

mining. While some researchers use algorithms such as Series Finder (Wang et al, 2013), decision tree (Malathi & Santhosh, 2011) and COPLINK Framework (Chen et al, 2004), it seems to be a preference for some authors such as Jin et al(n.d.), Nath (2006), Zubi and Mahmmud (2013), and even Malathi and Santhosh (2011) to use clustering algorithms to group similar past crimes, and analyse the results through their visualization on a map. But, none of them have used Linear Regression (LR), locally weighted learning (LWL), and M5P decision tree algorithms, so this project will use them to develop the models and compare their efficiency.

3. Crime Data Mining Requirements

The purpose of this chapter is to briefly describe the necessary requirements to develop an accurate and efficient data mining model that will help achieve the main aim of this project, which is to reduce the time of analysis for the historical crime data.

The first section will describe the Cross-industry Standard Process (CRISP-DM) adopted to manage the project. The second section will briefly review the data mining software (WEKA) used to analyse the datasets in this project.

In the following three sections, the techniques chosen to pre-process (cleaning the data, data aggregation, dimension reduction, and feature creation, feature creation, variable transformation, and dealing with missing values), process (LR, LWL, and M5P decision tree) and post-process (Correlation Coefficient, Mean Absolute Error, and Relative Mean Squared Error) the datasets will be reviewed and discussed, including their best results to achieve the aim of the project.

3.1. Standard Process

Ozgul, Atzenbeck, Celik, and Erdem (2011) introduced and compared four methodologies for crime data mining: CRISP-DM (Cross-industry Standard Process for Data Mining), CIA intelligence, Van der Hulst, and AMPA (Actionable Mining and Predictive Analytics). They concluded that CRISP-DM and CIA Intelligence methodologies are more suitable for prediction tasks.

CRISP-DM was developed for large- and small-scale data mining projects to achieve the main objectives and produce the desire results faster and

cheaper. For this reason this model will be adapted to this project.

Daimler-Benz, SPSS, and NCR proposed the standard process CRISP-DM projects in 1996, because the use of data mining to extract knowledge from historical data, and increase business efficiency has spread to various industries.

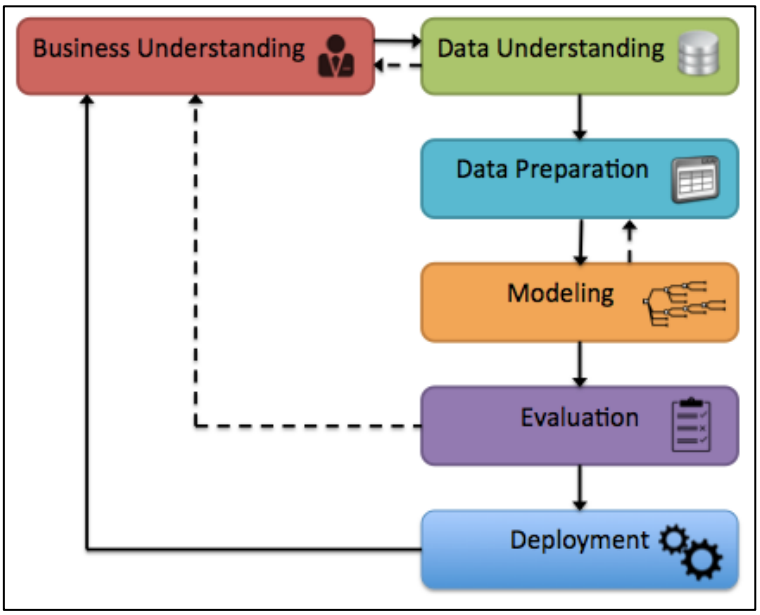


Figure 3-1: CRISP-DM Lifecycle

CRISP-DM is a reference model, which provides an overview of the lifecycle for a data mining project. It consists in 6 phases (Figure 3-1): business understanding, data understanding, data preparation, modelling, evaluation and deployment (Du, 2010).

- ***Business Understanding Phase:*** In the initial phase the background of the project is reviewed. Its aim is to understand the objectives, and requirements of the project. In this case, reduce time of data analysis, detect patterns, and predict crime.
- ***Data Understanding Phase:*** The data understanding phase is about the data's background, its aim is to identify the data source, attribute types, record descriptions and data quality (McCue, 2007).

- ***Data Preparation Phase:*** The data preparation phase includes all the techniques used from the collection of the data to the creation of the final dataset, which will be used in the modelling phase. Most of the time, this phase has to be performed multiple times before or after the creation of the model.

  Every technique used in this phase will depend on the approach of each project and will directly influence the results and performance of the model. For example, Nath (2006) explains that for missing values, outliers or multiple abbreviations such as blank or unknown that means the same, the KNN-cluster technique will group them for the same logical value. While the preparation method used by this author intents to be less invasive by not changing or deleting any records, Malathi and Santhosh (2011) used two ways to handle missing values problem; if four attributes were empty, they considered the record as irrelevant and deleted it, if not, they used a KNN-based methodology to replace the missing value with a mean value of the closest attributes.

  Both of the approaches use KNN-clustering to deal with the quality of datasets, and will depend on the selection of the k value, number of neighbours, and the distance function that KNN-cluster technique uses to group or replace the missing values.
- ***Modelling Phase:*** During this phase, different algorithms are selected such as ID3, association rules, clustering, neural networks, Support Vector Machines and decision trees. Some of these techniques need to calibrate their parameter values to create an optimal model that will solve the problem, and therefore researchers often returns to the preparation phase (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Wirth, 2000).
- ***Evaluation Phase:*** The model created in the last phases is reviewed before the deployment to validate if it reaches the objectives and requirement of the project established in the first phase “Business

Understand" (McCue, 2007). This phase can also raise some new important business issues that were not considered before (Chapman et al, 2000).

- ***Deployment Phase:*** Final phase where the model created and validated is implemented to the business practice. Measures should also be implemented in order to monitor the accuracy of the results; this may lead the model to modification (Du, 2010).

Nadali, Naghizadeh and Nosratabadi (2011), Moro, Laureano and Cortez (2011), and Suryajaya, Aryani, Devarakonda and Erwin (2014) are a few of the authors that have been used this standard to develop their projects, despite the fact that, according to Du (2010), it may need to address some issues in future editions. For example, describe the roles of data miners and domain experts on each task in the project.

3.2. Data Mining Software

Many commercial and non-commercial software tools can be used for data analysis. Those tools have different approaches, therefore it is important to choose the one that will help us to achieve the main objective of this project. This is why, the software must perform at least classification, and association analysis. A few of the tools are briefly described in the list below.

- ***R:*** It is mostly used for statistical purposes, but it can also execute data mining analysis for example, linear and nonlinear modeling, classification, and clustering. It can be downloaded under the GNU license and installed on any Operating System such as Windows, Linux and MAC (The R Foundation, n.d.).
- ***RapidMiner:*** It can also be downloaded under the GNU license. This tool was built on the WEKA data mining software, but includes better

visualization tools and a more effective Graphic Interface (Du, 2010, p.292).

- ***WEKA:*** A non-commercial tool that can be used through a graphic interface or command line. It provides many algorithms to perform pre-process, classification, cluster detection, association analysis and attribute selection tasks (Du, 2010, p. 8).
- ***Statistical Analysis System (SAS) Enterprise Miner:*** A software suit developed by SAS Institute that can perform a variety of data mining tasks including classification and association analysis. In February 2014, Gartner positioned this commercial product in the Leaders quadrant in the Magic Quadrant for Advanced Analytics Platforms (SAS, n.d.).
- ***STATISTICA:*** StatSoft developed this commercial platform originally for statistical analysis but it currently includes data analysis features. It provides a wide range of techniques for every stage of the data mining process (StatSoft, n.d.).
- ***Oracle Data Mining (ODM)****:* It is a new extension to the Oracle SQL Developer tool. It is based on a Graphic User Interface (GUI) and allows the user to work directly with data by exploring it, and create, evaluate, improve and implement models (Oracle, n.d.).

However, for this project we need a free tool that can be used through the command line as well as through the graphic interface therefore, the next section will describe more about the Waikato Environment for Knowledge Analysis (WEKA) software.

3.2.1. Waikato Environment for Knowledge Analysis Software.

The WEKA software was developed in 1993 at the University of Waikato in New Zealand. The first public release was at version 2.1 in 1996 (Hall, Frank, Holmes, Pfahringer, Reutemann & Witten, 2009). In the current time, WEKA can be portable, embed on an application or installed on any

Operating System such as Windows, Linux and Mac. This java application contains a collection of algorithms that permit the execution of data mining tasks such as pre-processing, classifying, clustering, associating, selecting attributes, and visualizing the dataset loaded in the application (Abernethy, 2010).

According to the SIGKDD Website, the WEKA team has made a big contribution to the data miners' community by freely distributing it under the terms and conditions of GNU General Public License and thus becoming one of the most used data mining applications. Data Miners can interact with it through four different environments: Explorer, Experimenter, Knowledge Flow and Simple Command Linea Interface (CLI) (Figure 3-2).

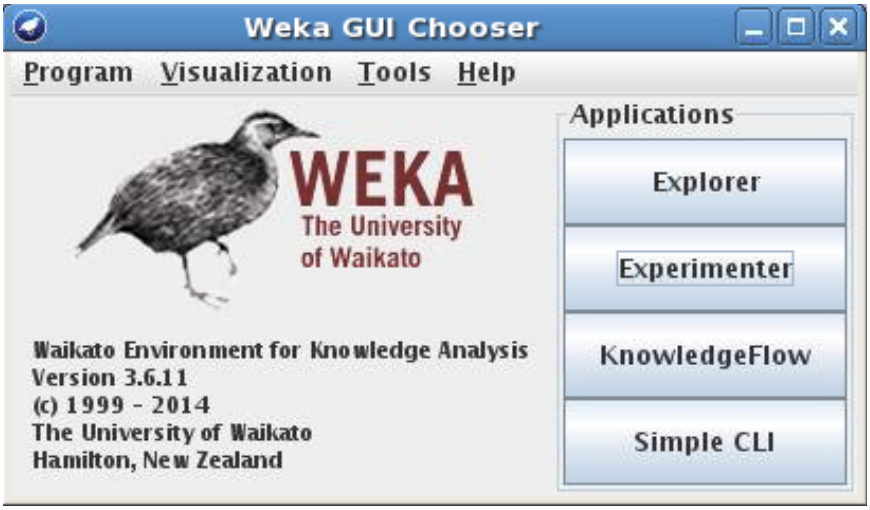


Figure 3-2: WEKA's GUI Chooser

WEKA can import the data from different types of file, such as database, Attribute Relation File Format (ARFF) and Comma-Separated Values (CSV). This project will be working with the last type of file, which can be edited using Microsoft Office or any text editor following its format.

Table 3-1: Sample of the Weather Dataset (Du, 2010, p.18).

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	FALSE	N
sunny	hot	high	TRUE	N
overcast	hot	high	FALSE	P
rain	mild	high	FALSE	P
rain	cool	normal	FALSE	P
rain	cool	normal	TRUE	N
overcast	cool	normal	TRUE	P

```
Outlook,Temperature,Humidity,Windy,Class
sunny,hot,high,FALSE,N
sunny,hot,high,TRUE,N
overcast,hot,high,FALSE,P
rain,mild,high,FALSE,P
rain,cool,normal,FALSE,P
rain,cool,normal,TRUE,N
overcast,cool,normal,TRUE,P
```

Figure 3-3: CSV format of the sample of the Weather Dataset

For instance, Figure 3-3 shows the representation in CSV format of Table 3-1, a sample of the weather Dataset (Du, 2010), where the first row defines the features or attributes of the dataset separated by commas, and the following rows represent data objects. Each data object is described by its values or records. Every dataset can have one or many classes that can be predicted. The last feature, in Figure 3-3, is the class. It will predict weather to play (P) or not (N) depending on the weather conditions (other features).

3.3. Pre-processing Techniques

As explained in Section 2.2.1, there are many methods that can be used on this stage, but for the purpose of this project only the use of the following six was necessary.

- ***Cleaning the data:*** This method deals with erroneous values and multiple variables that have the same meaning such as unk, unknown, “0”, or an empty value. Cleaning the data can be challenging and take a big amount of time depending on the size of the dataset and on how easy it is to understand it.
- ***Data Aggregation:*** This method condenses a number of data objects into a single one using some mathematics or statistics expressions, for example sum, average, and standard variation (Du, 2010). It is often known as summarization, however Kantardzic (2011) believes that even though both combine data objects, they are different in the data-warehousing context. Kantardzic (2011) states that the aggregation adds values from different business elements, and the summarization adds values within one or more data dimensions.
- ***Dimension Reduction and Feature Selection:*** The combination of less significant features, high dimensional data, noise, and irrelevant or erroneous data can make the analysis difficult and take more time to discover patterns. This problem can be solved by removing irrelevant features from the dataset (Du, 2010).

  Researchers have found that reducing the dimensionality of data improves the performance of the analysis and decrease the processing effort while maintaining reasonable accuracy (Kantardzic, 2011).
- ***Feature Creation:*** Nordman (n.d.), Ranka (2003) and Du (2010) agreed that this method creates a new feature based on the existing ones to capture the most important information efficiently.
- ***Variable Transformation:*** Although the same three authors mentioned above state that this method transforms the values of an attribute from one domain into another, Du (2010) establish a slightly difference between transforming and reducing the size of the dataset using transformation. He states that the variable transformation should be a

1:1 mapping record and therefore, should not reduce the dimension of the dataset.

- ***Dealing with missing Values:*** There are four main methods to deal with missing values.
  - To remove the data objects that contain empty values. Kantardzic (2011) and Nordman (n.d.) suggest that to use this method the missing values should not exceed the 5% of the data objects in the datasets.
  - Using of a domain expert, fill the missing values with an expectable and probable value. This method is not efficient on big dataset and can introduce noise when the value is not that obvious.
  - Fill the empty values by predicting them using statistic methods for instance Linear Regression. The dataset can be divided into a training and test sets, where the missing values are conveniently in the test set.
  - Finally, by filling the values using a constant, most frequent, average or a random value.

3.4. Data Mining Algorithms

To accomplish the main object of the project four different algorithms will be used. LWL, LR and M5P algorithms had being used to predict data in different areas of data mining, but they have not being used to predict crime data and therefore, we used them in this project.

As explained in section 2.2.2 many researchers in the crime field used the association rules because is a simple algorithm that discovers relationship between the attributes in a dataset, for this reason we tried to analyse the datasets using this approach and compare its results with the other algorithms.

3.4.1. Locally weighted Learning Algorithm.

Also called lazy learning, LWL is a classifier that creates a model based on neighboring data where higher weights are for the nearest data. For this reason, it is called locally weighted learning (Englert, n.d).

This algorithm had being used in different research to predict data. For instance, Jakkula (2007) used the algorithm along with the Multilayer perception (MLP), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) algorithms to predict health data collected from a smart home.

To create the models, LWL uses the equation below, where x is an n-dimensional vector in the continuous function $f(x)$, and "ϵ" represents the noise (Hof, 2001).

$$y = f(x) + \epsilon$$

It also uses the following basic cost function, where w_i represents the weight of the training set (x_i, y_i), x_q is the position where a value will be predicted, and β_q represents the regression coefficient (Englert, n.d.).

$$J = \frac{1}{2}\sum_{i=1}^{n} w_i(x_q)(y_i - x_i\beta_q)^2$$

This algorithm is included in the WEKA database and it can be found in the “lazy” category under the Classify tab (Figure 3-4).

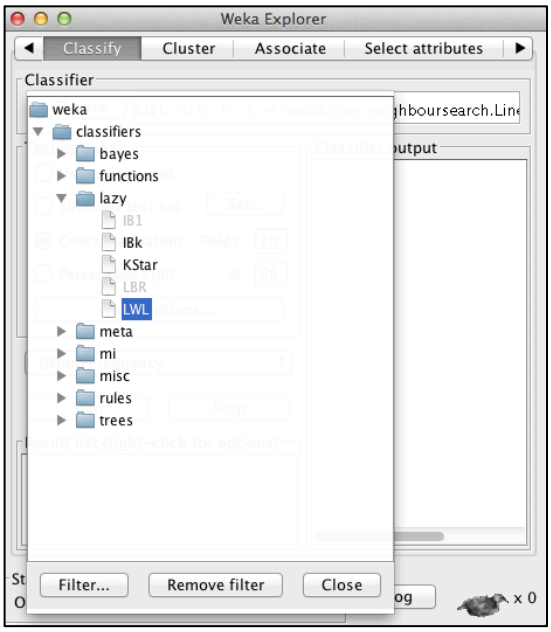


Figure 3-4: LWL algorithm in WEKA

3.4.2. Linear Regression Algorithm.

In Data mining, this algorithm is mostly used for prediction and forecasting, and it can be found in WEKA in the “functions” branch under the Classify tab (Figure 3-5).

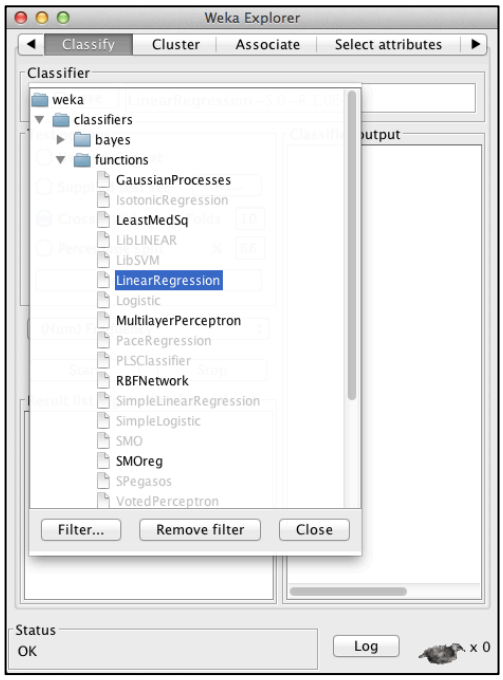


Figure 3-5: Linear Regression algorithm in WEKA

The linear regression algorithm is a statistical process that calculates weights from the independent attributes of the training data and predicts an unknown depending attribute from the training set (Witten, 2013). It can be called as simple or multivariate linear regression, depending on the amount of independent variables.

Often represented by the equation shown below (Figure 3-6), where "*x*" contains the independent attributes or explanatory variables, *"y"* is the dependent variable, "β" is a parameter vector called regression coefficient that represents the partial derivatives of *"y"* with respect to "*x*", and, "ϵ" is the error term, usually created by the noise of the data (Yale University, n.d.).

$$y = \beta x + \epsilon$$

Where

$$y = (y_1, y_2, \dots, y_n)$$

$$x = (x_1^T, x_2^T, \dots, x_n^T) = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\beta = (\beta_1, \beta_2, \dots, \beta_p)$$

$$\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$$

Figure 3-6: Linear Regression Equation.

3.4.3. M5P Decision Tree Algorithm.

It creates a model based on the M5P algorithm, which is a modified version from original M5 tree algorithm created by Quinlan (1992). The difference between the original and new version is that M5P can handle missing values and enumerated attributes (Wang & Witten, 1997).

First, while constructing a normal decision tree, it creates a linear regression model for each node of the tree. This is why the algorithm allows flexible predictions because its nodes can handle linear regression functions. Second, it prunes the tree until it reaches the root node to decrease the classification error. Finally, the smooth phase, where it predicts the value using the leaf model and while it returns back to the root node it combines that value with the predicted one obtained by each of the leaf linear model (Zhan, Gan & Hadi, 2011).

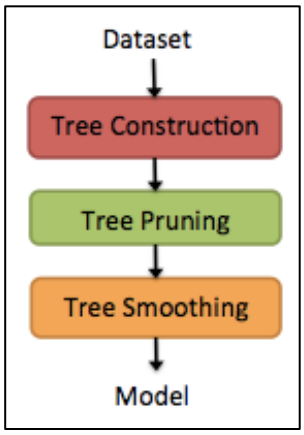


Figure 3-7: M5P algorithm process.

It can be found in WEKA under the classify tab in the “trees” category (Figure 3-8).

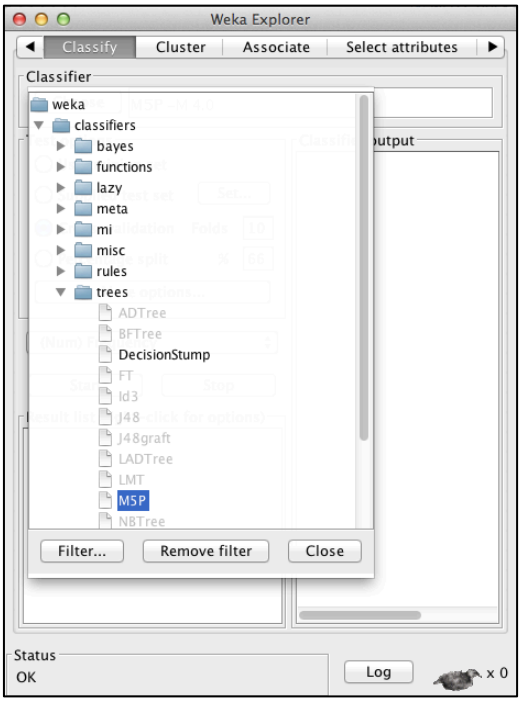


Figure 3-8: M5P algorithm in WEKA

3.4.4. Apriori Algorithm.

Proposed by Agrawal, Imielinski and Swami (1993), this algorithm is one of the most popular algorithms from the association rules, and it differs

from the classification algorithms because, it can obtain classification rules by searching frequent itemsets (Nordman, n.d.).

Depending on the type of values of the instances the association rules can be Boolean or quantitative. Figure 3-9 illustrates an example of these categories.

Boolean Association Rule,

$$Bread \implies Milk$$

Quantitative Association Rule,

$$(Age = 26, 27 \ldots 32) \implies (House = 0,1,2)$$

Figure 3-9: Examples of Boolean and Quantitative Association Rules

This algorithm searches for all frequent itemsets and generate strong association rules, which means that they meet the minimum confidence and support. The support value is the percentage of the data that satisfies the itemset and the confidence value is the percentage in which the consequent is satisfied (Técnico Lisboa, n.d.).

$$support\ (A \implies B) = \frac{P(A \cup B)}{n}$$

$$confidence\ (A \implies B) = \frac{P(A \cup B)}{P(A)} = P(A|B)$$

, where n is the number of instances

Figure 3-10: Support and Confidence equations.

First, to create the rules generate a frequent itemset of length k=1 denoted as L_1. Second, create itemsets of length 2 (k+1) denoted as L_2. Third, prune the itemsets of L_2 containing infrequent subsets of length k.

Fourth, generate the support of each subset. Fifth, remove the subsets that do not meet with the minimum support. Finally, continue with the creating new levels (L_k.) until no frequent k-itemsets can be found. The pseudo-code is written in Figure 3-11 (Wasilewska, 2014).

C_k: Candidate itemset of size k
L_k: frequent itemset of size k
L_1= {frequent items};
for (*k*=1; $L_k! = \phi$; *k*++) **do begin**
    C_{k+1}=subsets generated from L_k;
    for each transaction *t* in database **do**
        increment the count of all subsets C_{k+1} that are contained in *t*
    L_{k+1}=subsets in C_{k+1} with min_support
    end
return $\cup_k L_k$;

Figure 3-11: Pseudo-Code of the Apriori algorithm (Wasilewska, 2014).

The algorithm is available in WEKA under the Associate tab in the "associations" category (Figure 3-12).

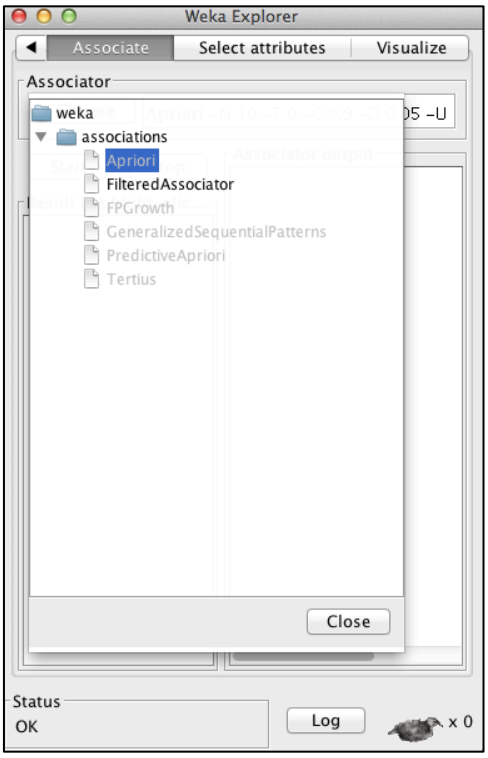


Figure 3-12: Apriori Algorithm in WEKA

3.4.5. Test Methods Models.

The models can be evaluated with different methods. Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse (2014) explain four methods, which are included in WEKA (Figure 3-13) as follow:

- ***Use training set:*** The model is trained by the input dataset, and evaluated on how well it predicts using the same input dataset (Bouckaert et al, 2014).
- ***Supplied Test Set:*** The model is trained by the input dataset, and evaluated on how well it predicts using a test-dataset (must be different from the input dataset).
- ***Cross-validation:*** The model is evaluated by cross-validation, using a number of folds previously setup.
- ***Percentage Split:*** the input dataset is divided into two sets. The amount of data for each set is established by the percentage in the %

field. The model is trained by the first dataset (bigger), and evaluated on how well it predicted using the second-dataset (smaller).

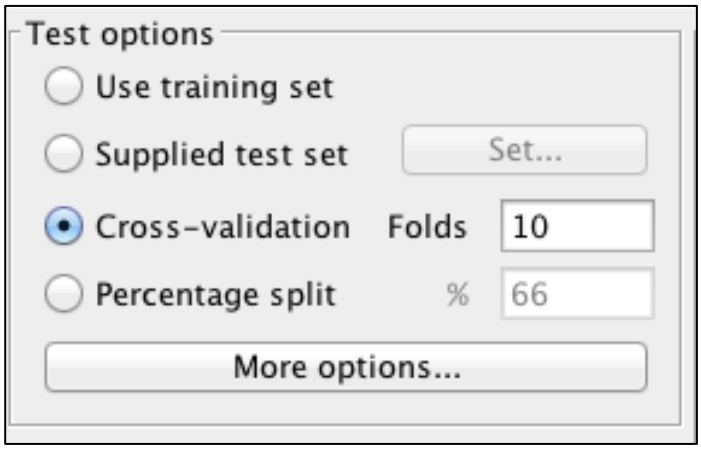


Figure 3-13: Test Options from the WEKA Software.

This project will use the most common methods: training and cross-validation methods, which are described in the following sections.

a. Training-Set Validation.

It is a method that creates a model using the training-set, and validates it by estimating the performance of its prediction. It is usually combined with the test-set validation method to validate the model using a different dataset. Figure 3-14 depicts the training-set validation, while Figure 3-15 illustrates the combination of training-test validation.

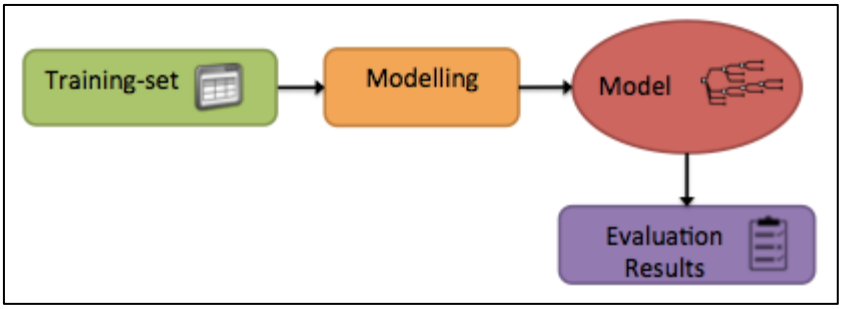


Figure 3-14: Training-set validation.

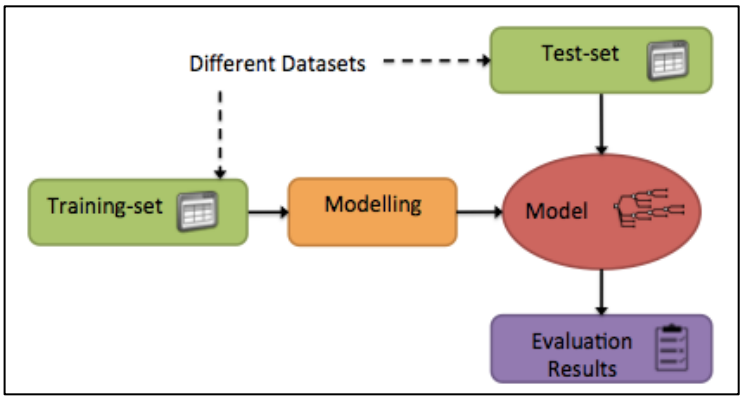


Figure 3-15: Training-Test validation.

b. 10-Fold Cross-Validation.

It is a model validation technique, similar to a combination of the training-test validation method, which also estimates the performance of a predictive model, but it divides the dataset randomly into 10 equal folds of size n/10, where n is the number of instances. The 90% of the folds correspond to the training-set, and the 10% to the test-set. Therefore, it creates a model using 9 of the folds and tests it with the last one. Finally, after repeating this process 9 more times choosing different training-set (9 folds) and test-set (1 fold), it averages the accuracies of the 10 models created (Riddle, 1998). An example of the 10-fold cross-validation is illustrated in Figure 3-16.

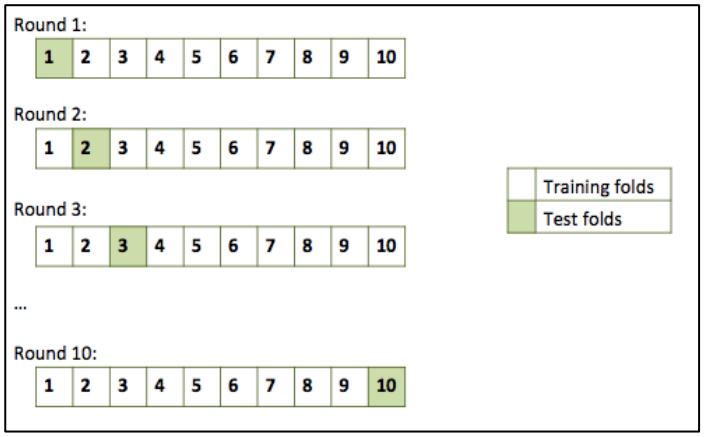


Figure 3-16: 10-fold Cross Validation process.

So, training-test validation and cross-validation methods execute conceptually similar procedures, but the cross-validation method has a more rigorous approach by averaging the accuracies of the folds created over the entire dataset.

3.5. Post-processing the patterns

Once the datasets are mined, they should be evaluated. Although this project does not have a customer in the traditional sense, its purpose is to discover interesting information related to crime prediction. Consequently, the project will be evaluated on the capacity of the models to predict crime. This evaluation varies depending on the model used and, thus, for our models, the following measures will be used for evaluation:

- ***Correlation coefficient (r):*** It measures the degree of linear association between the original and the predicted data. In WEKA, the algorithm LWL, LR and M5P provide a value between 0 and 1, where 0 indicates no linear correlation between the two types of data, while 1 indicates a perfect correlation (Eumetcal, n.d.a).

  If we plot the dataset and see them forming a straight line at an angle, we can say that they correlation is approaching to 1. For instance, Figure 3-17 shows the difference between a linear (left) and non-linear (right) correlation (Stackoverflow, 2011).

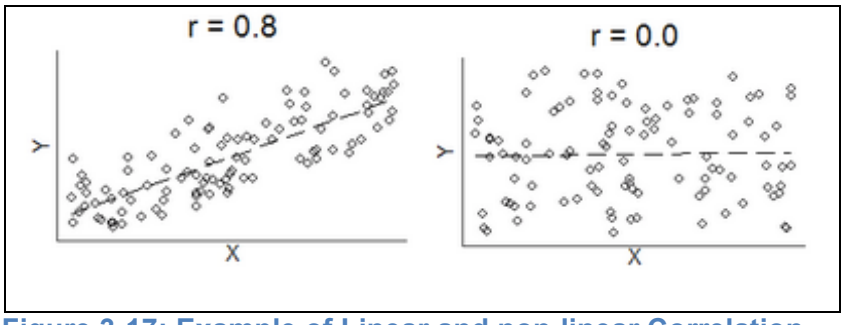


Figure 3-17: Example of Linear and non-linear Correlation

According to Mathbits (n.d.) the correlation is described as strong if the coefficient is greater than 0.8, and it is described as weak when the coefficient is less than 0.5.

- ***Mean absolute error (MAE):*** As the equation below illustrates, MAE is the average of the differences between the predicted and actual data, which means that less difference will decrease the error and therefore the model's accuracy will be better (Eumetcal, n.d.b).

$$\frac{1}{n}\sum_{i=1}^{n}|p_i - a_i|$$

where "p" is the predicted value and "a" the true value.

Figure 3-18: Mean Absolute Error Formula

- ***Root mean squared error (RMSE):*** This type of error can be analysed along with MAE, because their difference shows the variation between the individual errors. As the equation below shows, RMSE squared the errors before the average given them high weight to large errors. For this reason, RMSE will frequently be equal or higher than MAE. If their difference is bigger, the variance between the individual errors will be too (Eumetcal, n.d.b).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}|p_i - a_i|^2}$$

where "p" is the predicted value and "a" the true value.

Figure 3-19: Root Mean Squared Error Formula.

3.6. Conclusion.

The CRISP-DM standard will help us to manage the execution of the project and achieve the main objective of it. This project will be using the WEKA software for every analysis, which includes the algorithms LWL, LR, M5P and Apriori. Each experiment will be using a 10-fold cross-validation to train and test the datasets. Every model developed will be evaluated on the capacity of predicting crime. Overall, to obtain an accurate model is required that the correlation coefficient must approach 1, and the error values (MAE and RMSE) are as low as possible.

4. Methodology

This chapter describes the methodology used in this project. Following the process of CRISP-DM standard, the first section describes the collection and understanding of the datasets from the UK and Ecuador. The subsequent sections explain the data mining process conducted through the different experiments.

UK and Ecuador datasets followed similar experiments and used the same data mining algorithms in order to compare their results in the following chapters.

4.1. Understanding the Data.

The dataset used in this research were collected from the Police Departments in the UK and Ecuador.

- ***Police Department of UK:*** This data is provided by the Ministry of Justice, and published by the Police Department of UK through its website.

  As the Police Department explained in its website, crimes are directly recorded in the system by each police force. Then, it has to pass through a rigorous quality control process before being published. This quality process involves format validation, automated testing, and manual verification and approval by two separate people. Furthermore, the UK Police Department also explains the known issues, and how they are solving them, such as location accuracy, court result matching, double counting of anti-social behaviour (ASB) and crime, constantly changing data, and missing outcome data (Data.police.uk, n.d.).

The data downloaded from this website was distributed in monthly files, so it had to be integrated in one file before processing it.

- ***Police Department of Ecuador:*** This data is provided and authorized by the Ministry of Interior, and it is not published. Therefore, this Department sent it through a secure FTP server.

  The quality of this data is not as good as UK's data because each police officer has to fill a form about a crime, scan the document, and pass it to the system. Human errors can occur on each stage of this manual process. Therefore, the Ministry of Interior is working to improve the quality of the data by reviewing the location attribute with the Geographic Department and reviewing the names and personal data with the Register Department (A. Novoa, personal communication, April 26, 2014).

Both datasets have attributes in common for example, longitude, latitude, date and type of crime, and because UK's data is dated between December 2010 and February 2014, the author requested that the Ecuador's data is within the same range of dates.
A summary of the UK and Ecuador's datasets is described in the following table (Table 4-1).

Table 4-1: Datasets Summary

	UK	Ecuador
Data Objects:	609418	16219
Attributes:	12	36
Records:	7313016	583884
Missing Values:	1413564	170701
% of Missing Values:	19%	29%

The features of each datasets must be analysed to ensure their relevance in relation to the project objectives. Tables 4-2 and 4-3 describe each feature of the UK and Ecuador datasets respectively.

Table 4-2: Description of the features in the UK dataset.

Name	Type of Data	Description
Crime ID	Nominal	Id of the Crime.
Month	Nominal	Date of the crime in the format yyyy-mm.
Reported by	Nominal	The force that provided the data. For this research the author choose “Hampshire Constabulary”.
Falls within	Nominal	Same as “Reported By”. For this research “Hampshire Constabulary”
Longitude	Interval	Coordinates of the crime.
Latitude	Interval	Coordinates of the crime.
Location	Nominal	Specific or near location of the crime.
LSOA code	Nominal	Code of the Lower Layer Super Output Area (LSOA) where the crime was committed.
LSOA name	Nominal	Name of the LSOA where the crime was committed.
Crime type	Nominal	12 types of crime according to Data.police.uk (n.d.).
Last outcome category	Nominal	A reference to whichever of the outcomes associated with the crime occurred most recently.
Context	Nominal	Additional data.

Table 4-3: Description of the features in the Ecuador dataset.

Name	Type of Data	Description
Id	Nominal	Crime ID.
Provincia	Nominal	Province where the crime was committed. For this research the author choose “El Oro”.
Canton	Nominal	Canton where the crime was committed. For this research “Machala”.
Parroquia	Nominal	Parish where the crime was committed. For this research “Machala”
Zona	Nominal	Zone where the crime was committed. For this research “Zona 7”.
Distrito	Nominal	District where the crime was committed. For this research “Machala”.
Circuito	Nominal	Circuit where the crime was committed.
Subcircuito	Nominal	Sub-circuit where the crime was committed.
Latitud	Interval	Coordinates of the crime (latitude).
Longitud	Interval	Coordinates of the crime (longitude).

Fecha_infraccion	Nominal	Date of the crime in the format dd/mm/yyyy.
Hora_infraccion	Time	Time of the crime in the 24 hours format hh:mm.
Tipo_delito	Nominal	7 types of crime according to the Police Department of Ecuador.
Subtipo_delito	Nominal	61 Sub-types of crime according to the Police Department of Ecuador.
Modalidad	Nominal	Mode of the Crime. For example, assault or threat.
Agresión	Nominal	Describes whether or not the offense includes a physical or psychological aggression.
Lugar	Nominal	Place where the crime was committed. For example, bar, car, or store.
Origen_noticia	Nominal	Where did the complaint come? For example, Police document, any document, complaint.
Victima_denunciante	Nominal	Establish who made the complaint, the victim or not.
Sexo_victima	Nominal	Gender of the person who made the complaint.
Sospechoso	Nominal	Is there a suspect in the case?
Detenido	Nominal	Does the police have somebody in detention?
Alias	Nominal	The suspect Alias.
Carac. Fisicas	Nominal	Physical characteristics of the suspect.
Tez	Nominal	Type of skin of the suspect.
Estatura	Interval	Height of the suspect.
Acento	Nominal	Accent of the suspect.
Antecedentes	Nominal	Does the suspect have criminal records?
Nacionalidad	Nominal	Suspect nationality.
Vehiculo Robado	Nominal	Type of crime stolen (if it was stolen).
Marca	Nominal	Brand of the car.
Modelo	Nominal	Model of the car
Año	Interval	Year of the car.
Color	Nominal	Color of the car.
Placas	Nominal	Register number of the car.
Recuperado	Nominal	Was the car recovered?

To understand the similarities between the territorial division of the UK and Ecuador, the terms of each datasets are illustrated and compared in Figure 4-1.

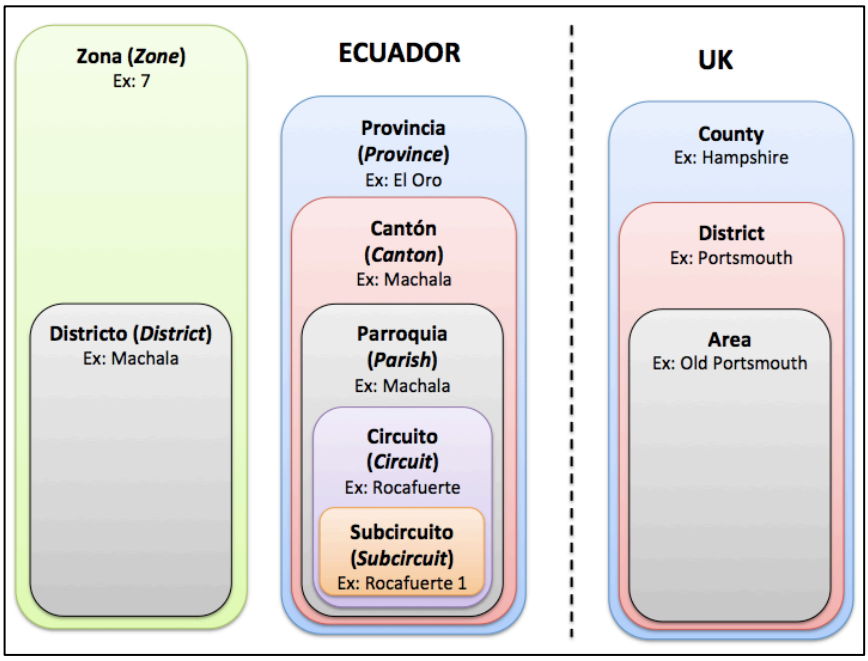


Figure 4-1: Similarities between the territorial division terms of the UK and Ecuador.

4.2. Data Preparation and Modelling.

Du (2010) explains that one of the main reasons of using the aggregation method is the time constraint along with the size of the data. For this reason, this method will be used in three of the four experiments. Despite the differences between the UK and Ecuador datasets, for each experiment similar procedures were followed. Previously to each analysis, we converted the datasets to CSV files in order to use them in WEKA.

Every experiment used 10-fold cross-validation and training-testing techniques to validate the models. Details of the procedures of each experiment are presented in the following subsections.

4.2.1. Experiment 1: General Analysis.

This experiment analyses the datasets using a new feature as a class. Figure 4-2 illustrates the procedure for the UK dataset, which initiates with the Data Understanding phase.

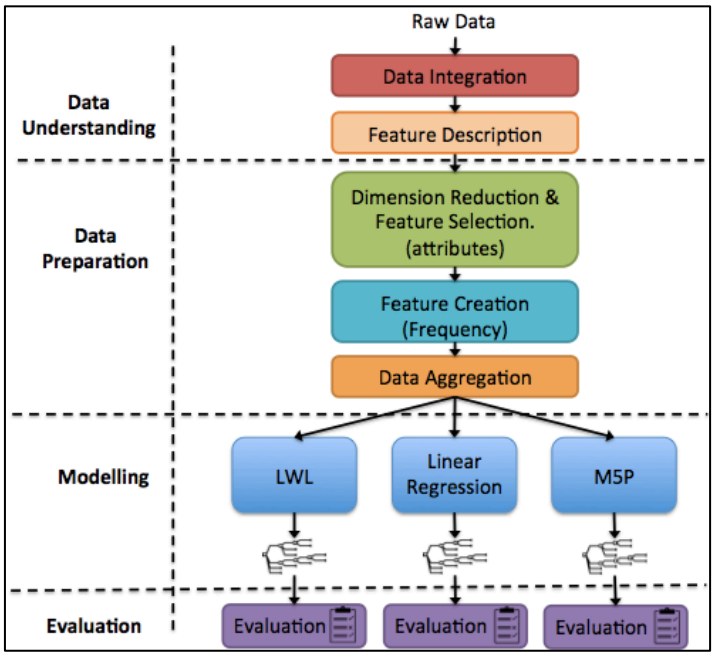


Figure 4-2: First Experiment Procedure for the UK dataset.

For the project purpose, we reduced the dimension of the dataset by selecting 4 attributes: month, LSOA Code, LSOA name, and Crime Type. Then, we created a new feature called “Frequency” and aggregated the data by Crime Type to obtain the frequency of each crime type per month.

Following a similar procedure initiating with the Data Understanding phase, Figure 4-3 illustrates the procedure for the Ecuador dataset.

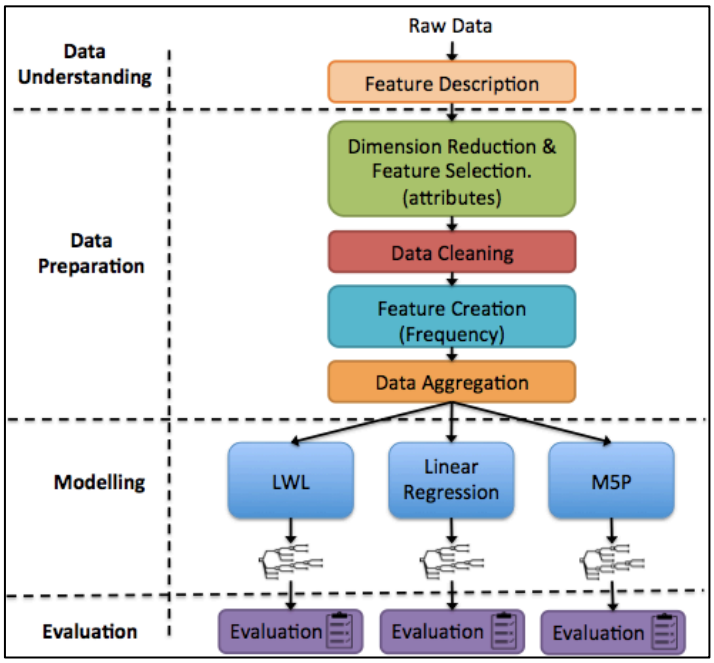


Figure 4-3: First Experiment Procedure for the Ecuador dataset.

After selecting 8 attributes: Fecha_infraccion, Circuito, Subcircuito, Tipo_delito, Subtipo_delito, Agresion, Victima_denunciante, sexo_victima, we analysed the dataset and discovered that its quality was still poor. Therefore, we had to clean it by setting the multiple variables into one; for example, the missing values were also set as “0”. Erroneous values were removed or changed, for example names in the gender (sexo_victima) feature.

Continuing with the procedure, we created the feature “Frequency” and aggregated the data using the same procedure as for the UK data.

The datasets created in this experiment were called “normal” and are detailed in Table 4-4.

Table 4-4: Summary of the Normal datasets for the First Experiment.

	UK	Ecuador
Data Objects:	196420	12902
Attributes:	5	9
Records:	982100	116118
Missing Values:	92	7881
% of Missing Values:	0%	7%

Once the data was prepared we analysed the datasets with the algorithms LWL, LR and M5P to obtained the models and evaluate their outcomes.

4.2.2. Experiment 2: Isolating the Types of Crimes

Due to the outcome of the first experiment, we decided to isolate the most frequent type of crime for each dataset and analyse it again.

Figure 4-6 and 4-7 illustrate the procedures for the UK and Ecuador datasets respectively, with an additional step. From the type of crime of each datasets (Table 4-5 and 4-6) and according to Figure 4-4 and 4-5 obtained with the visualization tool in WEKA, we selected “Antisocial Behaviour” for the type of crime in the UK dataset, and “Robbing People” (Robo a Personas) for the Sub-type of crime (Subtipo_delito) in Ecuador. These were selected because they were the most frequent types of crime in the two datasets respectively.

Table 4-5: Type of Crime values from the UK Dataset.

	Type of Crimes
1	Antisocial-behaviour
2	Burglary
3	Criminal damage and arson
4	Violent Crime
5	Other theft
6	Vehicle crime
7	Other crime
8	Drugs
9	Shoplifting
10	Violence and sexual offences

11	Public disorder and weapons
12	Public order
13	Bicycle theft
14	Robbery
15	Theft from the person
16	Possession of weapons

Table 4-6: Sub-type of crimes values from the Ecuador Dataset.

	Sub-type of Crime	**Translation**
1	Robo a personas	Robbing People
2	Intimidación/Amenaza	Intimidation / Threat
3	Robo domicilio	Robbing a house
4	Hurto	Theft
5	Varios	Other
6	Robo a motos	Motorcycle theft
7	Estafa	Scam
8	Robo a local commercial	Local commercial robbery
9	Heridas/Lesiones	Wounds / Injuries
10	Tentativa asesinato/homicidio	Murder attempted
11	Abuso de Confianza	Breach of trust
12	Robo a carros	Theft from cars
13	Robo accesorios de vehículos	Stolen vehicle accessories
14	Robo de bienes al interior del vehícúlo	Theft of goods from the vehicle
15	Falsificación	Counterfeiting
16	Otros robos	Other thefts
17	Extorsión	Extortion
18	Tenencia illegal armas/explosivos	Holding ilegal weapons / explosives
19	Violación	Violation
20	Tentativa de robo	Robbery attempted
21	Usurpación funciones/suplantación de identidad	Usurpation functions / phishing
22	Apropiación ilícita de bienes ajenos	Misappropriation of property of others.
23	Atentado al pudor	Indecent assault
24	Inves. Otras muertes	Inves. Other deaths
25	Delito aduanero	Customs crime
26	Robo a entidades privadas	Stealing from private entities
27	Asesinatos	Murders
28	Robo a fábricas/empresas	Theft from factories / companies
29	Acoso sexual	Sexual harassment
30	Usura	Usury
31	Delito ambiental	Environmental crime
32	Piratería	Piracy
33	Invasiones/Allanamiento	Invasion / Burglary
34	Plagio o secuestro personas	Kidnapping or abduction people
35	Ocultación de cosas robadas	Concealment of stolen goods
36	Desaparición personas	Missing persons

37	Rapto	Rapture
38	Robo a instituciones educativas	Stealing from educational institutions
39	Robo en carreteras	Robbing in the roads
40	Robo a entidades públicas	Stealing from public entities
41	Homicidios	Homicide
42	Rebelión y atentado	Rebellion and attacks
43	Peculado	Embesslement
44	Evasión de impuestos	Tax evasion
45	Tentativa de violación	Attempted rape
46	Estupro	Rape
47	Delito energético	Energy Offense
48	Falso testimonio/perjuicio	Perjury/Injury
49	Tentativa de plagio/secuestro	Attempted kidnapping / abduction
50	Trata persona	Person traffic
51	Prevaricato	Malfeasance
52	Asociación ilícita	Conspiracy
53	Cohecho	Bribery
54	Proxenetismo	Pimping
55	Abigeato	Rustling
56	Tráfico/tenencia de moneda falsa	Traffic / possession of false money
57	Secuestro express	Kidnapping express
58	Robo a bancos y entidades financieras	Stealing from banks and financial institutions
59	Delito contra el patrimonio	Crime against the property
60	Suicidios	Suicides
61	Tráfico de migrantes	Migrants traffic

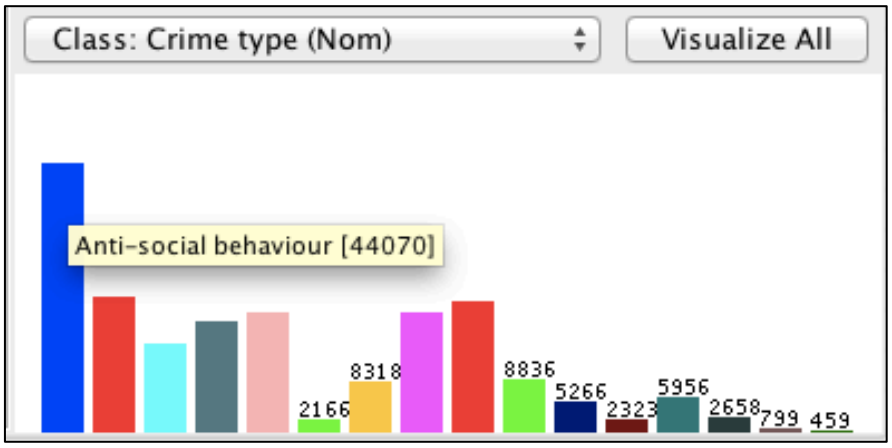


Figure 4-4: Most Frequent Type of Crime (UK dataset)

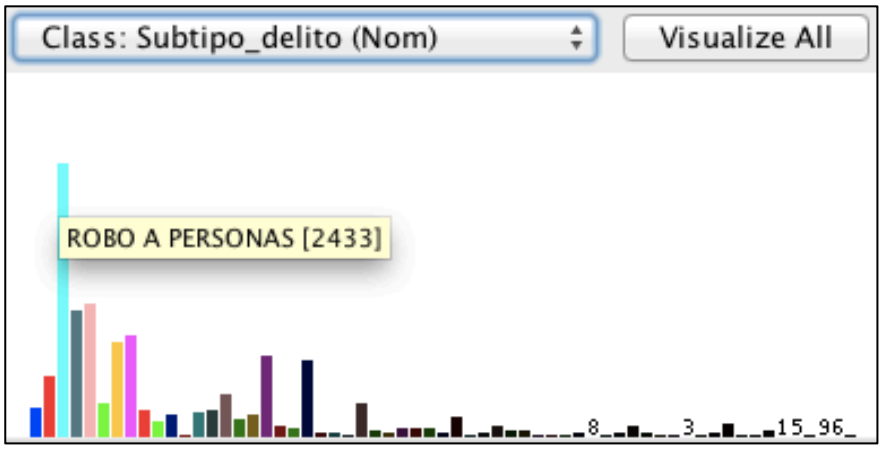


Figure 4-5: Most Frequent Type of Crime (Ecuador dataset)

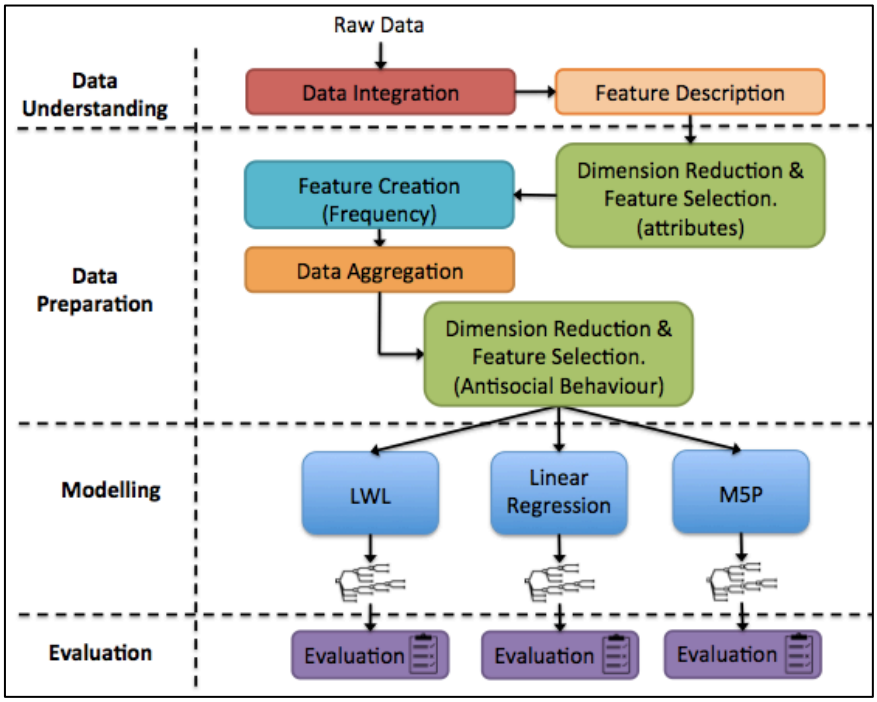


Figure 4-6: Second Experiment Procedure for the UK dataset.

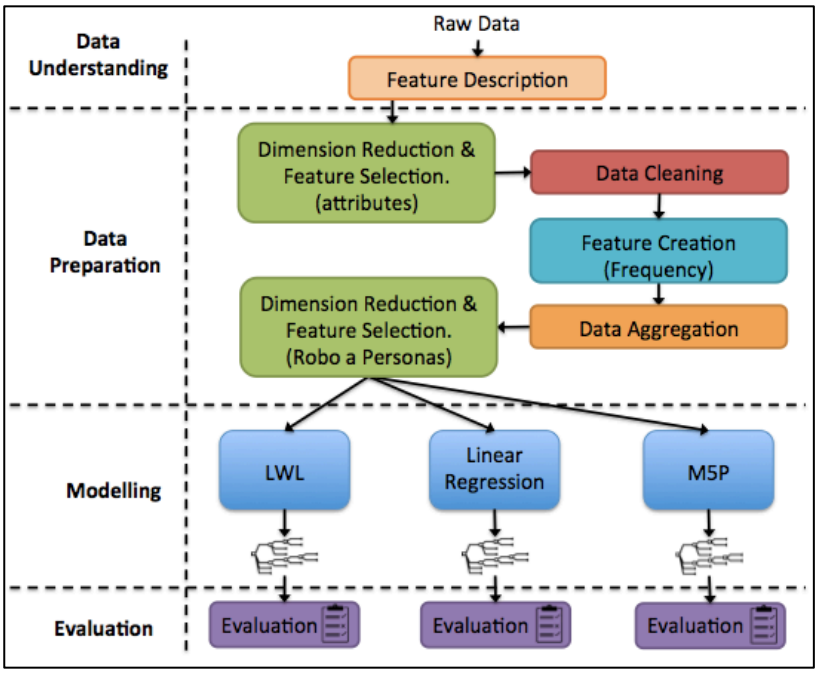


Figure 4-7: Second Experiment Procedure for the Ecuador dataset.

As Table 4-7 shows, both datasets, called "Crime Type", have fewer attributes and records after the feature selection. After the data preparation phase, we analysed the datasets with the same algorithms.

Table 4-7: Summary of the Crime Type Datasets for the Second Experiment.

	UK	Ecuador
Data Objects:	44070	2432
Attributes:	4	7
Records:	176280	17024
Missing Values:	34	1182
% of Missing Values:	0%	7%

4.2.3. Experiment 3: Variable Transformation

After the Understanding of the data phase, we selected the same 4 attributes as the first experiment, and then transformed the LSOA-Code feature into Postcodes using the database published in the Office for National Statistics (2013) Website.

To continue with the process we created the new feature “Frequency” and aggregated the data to obtain our dataset ready for the analysis.

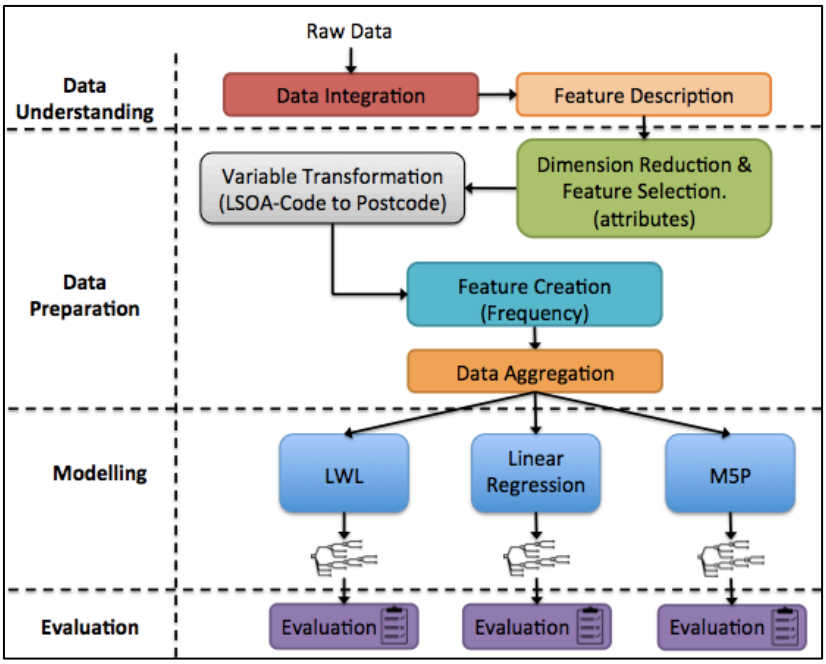


Figure 4-8: Third Experiment Procedure for the UK dataset.

Because the Ecuador dataset did not have the feature LSOA-Code, this process (Figure 4-8) was used only for the UK dataset. The one generated in this experiment were called “Transform”, and its characteristics were described in Table 4-8.

Table 4-8: Summary of the Transform Dataset for the Third Experiment.

	UK
Data Objects:	155021
Attributes:	5
Records:	775105
Missing Values:	0
% of Missing Values:	0%

4.2.4. Experiment 4: Association Rules

This experiment did not need a considerable amount of preparation. After selecting the attributes for the UK dataset, and selecting the attributes and cleaning the data for the Ecuador, we continued by analysing the datasets looking for the Association rules using the Apriori Algorithm.

Association rules do not need a numerical feature and, therefore we did not use the frequency feature. The description of datasets previous to the analysis is detailed in Table 4-9.

Table 4-9: Summary of the Datasets for the Fourth Experiment.

	UK	Ecuador
Data Objects:	609417	12902
Attributes:	5	8
Records:	3047085	103216
Missing Values:	212	7881
% of Missing Values:	0%	8%

The processes for this experiment for the UK and Ecuador datasets are graphically illustrated on Figure 4-9 and 4-10 respectively.

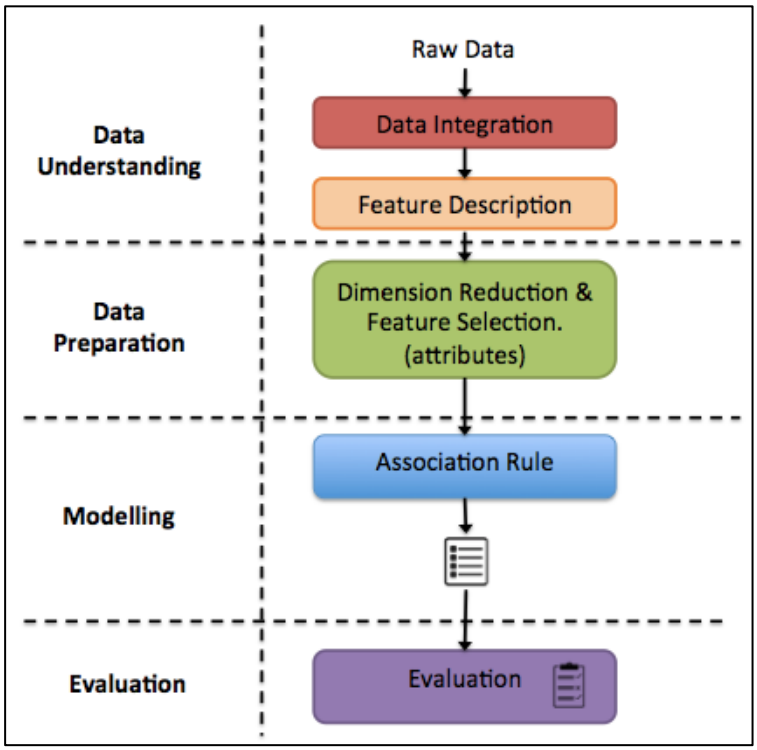


Figure 4-9: Fourth Experiment Procedure for the UK dataset.

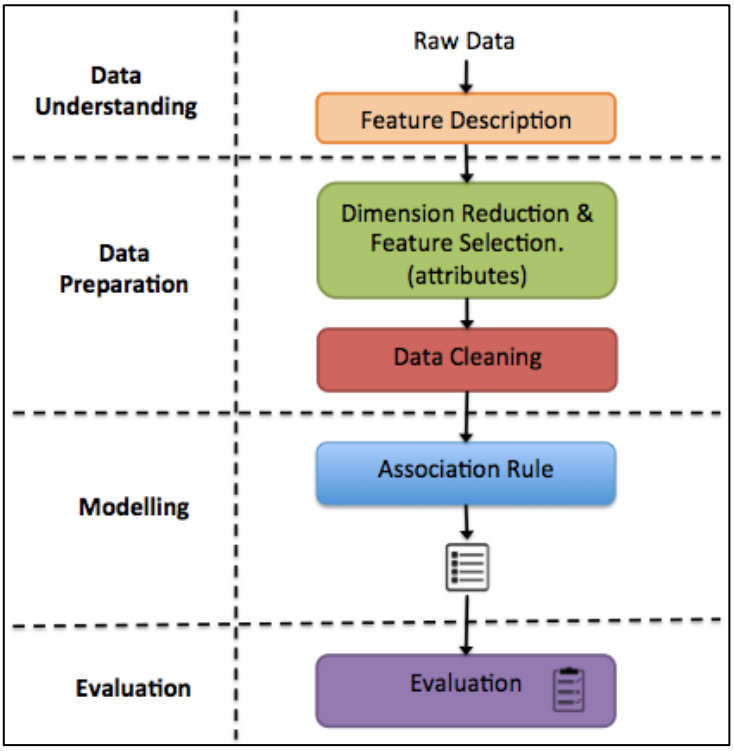


Figure 4-10: Fourth Experiment Procedure for the Ecuador dataset.

Table 4-10: Summary of the UK Dataset for the Fourth Experiment (second time).

	UK
Data Objects:	262286
Attributes:	4
Records:	1049144
Missing Values:	126
% of Missing Values:	0%

Due to the results of this experiment using the UK dataset, we had to reduce this dataset into the most frequent type of crime “Antisocial Behaviour” and analyse it again with the association rule.

5. Implementation

Using the requirements discussed in chapter three, we were able to develop the methodology explained in chapter four and implemented in this chapter. Its purpose is to review the output generated by WEKA through the command line. Because of the dimension of the datasets and the use of the supercomputer from the ICG, we had to implement the experiments through the command line of WEKA.

The first four sections will review the output and performance of every experiment described in the Methodology Chapter by using the training-set validation and 10-fold cross-validation methods, while the last section will summarize the algorithms, experiments and datasets of this project.

5.1. Initial analysis with "Data Frequency".

After collecting the datasets and preparing it for the analysis, we followed the procedure shown in Figure 4-2 for the UK dataset and Figure 4-3 for the Ecuador dataset, and start the analysis in WEKA's command line using specific scripts for each algorithm and each dataset.

- ***Linear Regression:*** The command used to execute this analysis using the Linear Regression algorithm for each dataset is written in Table 5-1, and their status and outputs are presented in Figure 5-1 and 5-2 respectively.

Table 5-1: Commands to execute experiment 1 using Linear Regression.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.functions.LinearRegression -t Data/UKData.csv -d Models/UKLR.model -o -i >> Results/UKLR.result
Ecuador	java -Xmx20480m weka.classifiers.functions.LinearRegression -t

	Data/EcuadorData.csv -d Models/EcuadorLR.model -o -i >> Results/EcuadorLR.result

```
[gsaltosb@login2 ~]$ date
Thu Sep  4 22:50:49 BST 2014
[gsaltosb@login2 ~]$ qstat -tn1u gsaltosb

headnode1.sciama.icg.port.ac.uk:
                                                                  Req'd  Req'd   Elap
Job ID               Username Queue    Jobname          SessID NDS   TSK Memory Time  S Time
-------------------- -------- -------- ---------------- ------ ----- --- ------ ----- - -----
1029411.headnode     gsaltosb cluster. LinearRegression  19265     1  12    --  1500: R 1091:   node31/
11+node31/10+node31/9+node31/8+node31/7+node31/6+node31/5+node31/4+node31/3+node31/2+node31/1+node31/0
```

Figure 5-1: The first experiment status using the UK dataset and Linear Regression algorithm (September 4th 2014).

```
Time taken to build model: 21.31 seconds
Time taken to test model on training data: 0.46 seconds

=== Error on training data ===

Correlation coefficient                  0.7391
Mean absolute error                      0.1685
Root mean squared error                  0.5358
Relative absolute error                 38.8456 %
Root relative squared error             67.3558 %
Total Number of Instances            12902



=== Cross-validation ===

Correlation coefficient                  0.7328
Mean absolute error                      0.1708
Root mean squared error                  0.5413
Relative absolute error                 39.37   %
Root relative squared error             68.0395 %
Total Number of Instances            12902
```

Figure 5-2: The first experiment output using the Ecuador dataset and Linear Regression algorithm.

- ***LWL:*** The command used to execute this analysis using the LWL algorithm for each dataset is written in Table 5-2, and their outputs are presented in Figure 5-3 and 5-4.

Table 5-2: Commands to execute experiment 1 using LWL.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.lazy.LWL -t Data/UKData.csv -d Models/UKLWL.model -x 10 -o –i >> Results/UKLWL.results
Ecuador	java -Xmx20480m weka.classifiers.lazy.LWL -t Data/EcuadorData.csv -d Models/EcuadorLWL.model -x 10 -o –i >> Results/EcuadorLWL.results

```
Time taken to build model: 0.04 seconds
Time taken to test model on training data: 19935.54 seconds

=== Error on training data ===

Correlation coefficient                  0.4722
Mean absolute error                      2.0324
Root mean squared error                  3.9724
Relative absolute error                 85.3736 %
Root relative squared error             88.1987 %
Total Number of Instances           196420


=== Cross-validation ===

Correlation coefficient                  0.4684
Mean absolute error                      2.0354
Root mean squared error                  3.9812
Relative absolute error                 85.4969 %
Root relative squared error             88.395  %
Total Number of Instances           196420
```

Figure 5-3: The first experiment output using the UK dataset and LWL algorithm.

```
Time taken to build model: 0.01 seconds
Time taken to test model on training data: 108.26 seconds

=== Error on training data ===

Correlation coefficient                  0.6199
Mean absolute error                      0.3585
Root mean squared error                  0.6373
Relative absolute error                 82.6712 %
Root relative squared error             80.1051 %
Total Number of Instances            12902


=== Cross-validation ===

Correlation coefficient                  0.5597
Mean absolute error                      0.3655
Root mean squared error                  0.6678
Relative absolute error                 84.2707 %
Root relative squared error             83.9458 %
Total Number of Instances            12902
```

Figure 5-4: The first experiment output using Ecuador dataset and LWL algorithm.

- ***M5P:*** The command used to execute this analysis using the M5P algorithm for each dataset is written in Table 5-3, and their measures are presented in Figure 5-5 and 5-6.

Table 5-3: Commands to execute experiment 1 using M5P.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.trees.M5P -t Data/UKData.csv -d Models/UKM5P.model -x 10 -o -i>> Results/UKM5P.results
Ecuador	java -Xmx20480m weka.classifiers.trees.M5P -t Data/EcuadorData.csv -d Models/EcuadorM5P.model -x 10 -o -i >> Results/EcuadorM5P.results

```
Time taken to build model: 117668.67 seconds
Time taken to test model on training data: 6.83 seconds

=== Error on training data ===

Correlation coefficient                  0.8598
Mean absolute error                      1.2639
Root mean squared error                  2.3001
Relative absolute error                 53.0905 %
Root relative squared error             51.0691 %
Total Number of Instances           196420



=== Cross-validation ===

Correlation coefficient                  0.8328
Mean absolute error                      1.3229
Root mean squared error                  2.4939
Relative absolute error                 55.5676 %
Root relative squared error             55.3714 %
Total Number of Instances           196420
```

Figure 5-5: The first experiment outcome using UK dataset and M5P algorithm.

```
Time taken to build model: 10.32 seconds
Time taken to test model on training data: 0.32 seconds

=== Error on training data ===

Correlation coefficient                  0.7442
Mean absolute error                      0.1466
Root mean squared error                  0.5313
Relative absolute error                 33.7948 %
Root relative squared error             66.7911 %
Total Number of Instances            12902



=== Cross-validation ===

Correlation coefficient                  0.7317
Mean absolute error                      0.1493
Root mean squared error                  0.5424
Relative absolute error                 34.4131 %
Root relative squared error             68.176  %
Total Number of Instances            12902
```

Figure 5-6: The first experiment outcome using Ecuador dataset and M5P algorithm.

5.2. Analysis isolating the most frequent type of crimes.

Following the process shown in Figure 4-6 and 4-7 for the UK and Ecuador dataset respectively, we analysed the datasets selecting the most frequent type of crime. To start the analysis in WEKA, we used the commands written below for each algorithm and each dataset.

- ***Linear Regression:*** The command used to execute this analysis using the Linear Regression algorithm for each dataset is written in Table 5-4, and their measures are presented in Figure 5-7 and 5-8.

Table 5-4: Commands to execute experiment 2 using LR.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.functions.LinearRegression -t Data/UKDataAntisocial.csv -d Models/UKLRAntisocial.model -o -i>> Results/UKLRAntisocial.results
Ecuador	java -Xmx20480m weka.classifiers.functions.LinearRegression -t Data/EcuadorDataRobo.csv -d Models/EcuadorLRRobo.model -o -i>> Results/EcuadorLRRobo.results

```
Time taken to build model: 294677.5 seconds
Time taken to test model on training data: 2.19 seconds

=== Error on training data ===

Correlation coefficient                  0.8559
Mean absolute error                      2.2451
Root mean squared error                  3.2811
Relative absolute error                 54.5621 %
Root relative squared error             51.7107 %
Total Number of Instances            44070


=== Cross-validation ===

Correlation coefficient                  0.8453
Mean absolute error                      2.3275
Root mean squared error                  3.3913
Relative absolute error                 56.5642 %
Root relative squared error             53.4459 %
Total Number of Instances            44070
```

Figure 5-7: The second experiment outcome using UK dataset and LR algorithm.

```
Time taken to build model: 1.15 seconds
Time taken to test model on training data: 0.12 seconds

=== Error on training data ===

Correlation coefficient                  0.7573
Mean absolute error                      0.1829
Root mean squared error                  0.5258
Relative absolute error                 40.4558 %
Root relative squared error             65.3051 %
Total Number of Instances             2432


=== Cross-validation ===

Correlation coefficient                  0.7342
Mean absolute error                      0.1894
Root mean squared error                  0.5469
Relative absolute error                 41.8827 %
Root relative squared error             67.9157 %
Total Number of Instances             2432
```

Figure 5-8: The second experiment outcome using Ecuador dataset and LR algorithm.

- ***LWL:*** The command used to execute this analysis using the LWL algorithm for each dataset is written in Table 5-5, and their measures are presented in Figure 5-9 and 5-10.

Table 5-5: Commands to execute experiment 2 using LWL.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.lazy.LWL -t Data/UKDataAntisocial.csv -d Models/UKLWLAntisocial.model -x 10 -o -i>> Results/UKLWLAntisocial.results
Ecuador	java -Xmx20480m weka.classifiers.lazy.LWL -t Data/EcuadorDataRobo.csv -d Models/EcuadorLWLRobo.model -x 10 -o -i>> Results/EcuadorLWLRobo.results

```
Time taken to build model: 0.02 seconds
Time taken to test model on training data: 486.84 seconds

=== Error on training data ===

Correlation coefficient                  0.7698
Mean absolute error                      3.2763
Root mean squared error                  4.1767
Relative absolute error                 79.6216 %
Root relative squared error             65.8247 %
Total Number of Instances            44070


=== Cross-validation ===

Correlation coefficient                  0.7481
Mean absolute error                      3.3488
Root mean squared error                  4.3171
Relative absolute error                 81.3837 %
Root relative squared error             68.0368 %
Total Number of Instances            44070
```

Figure 5-9: The second experiment outcome using UK dataset and LWL algorithm.

```
Time taken to build model: 0 seconds
Time taken to test model on training data: 2.94 seconds

=== Error on training data ===

Correlation coefficient                  0.64
Mean absolute error                      0.3783
Root mean squared error                  0.6332
Relative absolute error                 83.6953 %
Root relative squared error             78.643  %
Total Number of Instances             2432


=== Cross-validation ===

Correlation coefficient                  0.4656
Mean absolute error                      0.4
Root mean squared error                  0.7167
Relative absolute error                 88.4736 %
Root relative squared error             88.9907 %
Total Number of Instances             2432
```

Figure 5-10: The second experiment outcome using Ecuador dataset and LWL algorithm.

- ***M5P:*** The command used to execute this analysis using the M5P algorithm for each dataset is written in Table 5-6, and their outputs are presented in Figure 5-11 and 5-12.

Table 5-6: Commands to execute experiment 2 using LWL.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.trees.M5P -t Data/UKDataAntisocial.csv -d Models/UKM5PAntisocial.model -

	x 10 -o -i>> Results/UKM5PAntisocial.results
Ecuador	java -Xmx20480m weka.classifiers.trees.M5P -t Data/EcuadorDataRobo.csv -d Models/EcuadorM5PRobo.model -x 10 -o -i>> Results/EcuadorM5PRobo.results

```
Time taken to build model: 46512.61 seconds
Time taken to test model on training data: 1.52 seconds

=== Error on training data ===

Correlation coefficient                  0.8647
Mean absolute error                      2.1649
Root mean squared error                  3.1869
Relative absolute error                 52.6122 %
Root relative squared error             50.2261 %
Total Number of Instances            44070


=== Cross-validation ===

Correlation coefficient                  0.8515
Mean absolute error                      2.2601
Root mean squared error                  3.3278
Relative absolute error                 54.9243 %
Root relative squared error             52.4458 %
Total Number of Instances            44070
```

Figure 5-11: The second experiment outcome using UK dataset and M5P algorithm.

```
Time taken to build model: 1.01 seconds
Time taken to test model on training data: 0.06 seconds

=== Error on training data ===

Correlation coefficient                  0.7795
Mean absolute error                      0.1419
Root mean squared error                  0.5044
Relative absolute error                 31.3919 %
Root relative squared error             62.6444 %
Total Number of Instances             2432


=== Cross-validation ===

Correlation coefficient                  0.7354
Mean absolute error                      0.1586
Root mean squared error                  0.5462
Relative absolute error                 35.0753 %
Root relative squared error             67.8246 %
Total Number of Instances             2432
```

Figure 5-12: The second experiment outcome using Ecuador dataset and M5P algorithm.

5.3. Analysis transforming LSOA Code to Postcode.

Previous to analysis, we transform the feature from the LSOA Code to the Postcode. Then we analysed the dataset in WEKA using the commands written below for each algorithm.

- ***Linear Regression:*** The command used to execute this analysis using the Linear Regression algorithm for the UK dataset is written in Table 5-7, and its measures are presented in Figure 5-13.

Table 5-7: Commands to execute experiment 3 using Linear Regression.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.functions.LinearRegression -t Data/UKDataPostcode.csv -d Models/UKLRPostcode.model -o -i >> Results/UKLRPostcode.results

```
Time taken to build model: 389816.58 seconds
Time taken to test model on training data: 7.35 seconds

=== Error on training data ===

Correlation coefficient                  0.6301
Mean absolute error                      1.9161
Root mean squared error                  3.3126
Relative absolute error                 80.3458 %
Root relative squared error             77.6484 %
Total Number of Instances           155021


=== Cross-validation ===

Correlation coefficient                  0.5035
Mean absolute error                      1.9512
Root mean squared error                  3.8758
Relative absolute error                 81.818  %
Root relative squared error             90.8484 %
Total Number of Instances           155021
```

Figure 5-13: The third experiment outcome using Linear Regression algorithm.

- ***LWL:*** The command used to execute this analysis using the LWL algorithm for the UK dataset is written in Table 5-8, and its measures are presented in Figure 5-14.

Table 5-8: Commands to execute experiment 3 using LWL.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.lazy.LWL -t Data/UKDataPostcode.csv -d Models/UKLWLPostcode.model -x 10 -o -i>> Results/UKLWLPostcode.results

```
Time taken to build model: 0.04 seconds
Time taken to test model on training data: 15182.61 seconds

=== Error on training data ===

Correlation coefficient                  0.4964
Mean absolute error                      1.9926
Root mean squared error                  3.7113
Relative absolute error                 83.5542 %
Root relative squared error             86.9933 %
Total Number of Instances           155021


=== Cross-validation ===

Correlation coefficient                  0.4924
Mean absolute error                      1.9955
Root mean squared error                  3.72
Relative absolute error                 83.6757 %
Root relative squared error             87.1975 %
Total Number of Instances           155021
```

Figure 5-14: The third experiment outcome using LWL algorithm.

- ***M5P:*** The command used to execute this analysis using the M5P algorithm for the UK dataset is written in Table 5-9, and its measures are presented in Figure 5-15.

Table 5-9: Commands to execute experiment 3 using M5P.

Dataset	Command Line
UK	java -Xmx20480m weka.classifiers.trees.M5P -t Data/UKDataPostcode.csv -d Models/UKM5PPostcode.model -x 10 -o -i>> Results/UKM5PPostcode.results

```
Time taken to build model: 61159.3 seconds
Time taken to test model on training data: 5.05 seconds

=== Error on training data ===

Correlation coefficient                  0.8833
Mean absolute error                      1.2202
Root mean squared error                  2.0005
Relative absolute error                 51.1646 %
Root relative squared error             46.892  %
Total Number of Instances           155021


=== Cross-validation ===

Correlation coefficient                  0.8565
Mean absolute error                      1.3016
Root mean squared error                  2.2027
Relative absolute error                 54.5801 %
Root relative squared error             51.6321 %
Total Number of Instances           155021
```

Figure 5-15: The third experiment outcome using M5P algorithm.

5.4. Analysis using Apriori algorithm.

Continuing with the process (Figure 4-9 and 4-10), after preparing the data, we analysed each dataset in WEKA using the commands written in Table 5-10 for the UK and Ecuador datasets.

Table 5-10: Commands to execute experiment 4 using Apriori algorithm.

Dataset	Command Line
UK	java -Xmx20480m weka.associations.Apriori -N 100 -T 1 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -t Data/UKDataAsociationRule.csv >>Results/UKAR.results
Ecuador	java -Xmx20480m weka.associations.Apriori -N 100 -T 1 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -t Data/EcuadorDataAsociationRule.csv >>Results/EcuadorAR.results

The commands above define the extraction of the 100 best association rules from each dataset with a minimum confidence of 0.9 and a minimum support of 0.1. The measures for the UK dataset are presented in Figure 5-16, and part of the measures for the Ecuador dataset are presented in Figure 5-17.

```
No large itemsets and rules found!

=== Evaluation ===

Elapsed time: 1500.044s
```

Figure 5-16: The fourth experiment outcome using the UK dataset.

```
Apriori
=======

Minimum support: 0.1 (1290 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 28

Size of set of large itemsets L(3): 9

Best rules found:

  1. Subcircuito=ROCAFUERTE 1 3188 ==> Circuito=ROCAFUERTE 3188    conf:(1)
  2. Circuito=ROCAFUERTE 3188 ==> Subcircuito=ROCAFUERTE 1 3188    conf:(1)
  3. Subtipo_delito=ROBO A PERSONAS 2433 ==> Tipo_delito=CONTRA LA PROPIEDAD 2433    conf:(1)
  4. Subcircuito=ROCAFUERTE 1 Tipo_delito=CONTRA LA PROPIEDAD 2013 ==> Circuito=ROCAFUERTE 2013    conf:(1)
  5. Circuito=ROCAFUERTE Tipo_delito=CONTRA LA PROPIEDAD 2013 ==> Subcircuito=ROCAFUERTE 1 2013    conf:(1)
  6. Subcircuito=ROCAFUERTE 1 Sexo_victima=masculino 1722 ==> Circuito=ROCAFUERTE 1722    conf:(1)
  7. Circuito=ROCAFUERTE Sexo_victima=masculino 1722 ==> Subcircuito=ROCAFUERTE 1 1722    conf:(1)
  8. Subcircuito=ROCAFUERTE 1 Victima_denunciante=denunciante 1513 ==> Circuito=ROCAFUERTE 1513    conf:(1)
  9. Circuito=ROCAFUERTE Victima_denunciante=denunciante 1513 ==> Subcircuito=ROCAFUERTE 1 1513    conf:(1)
 10. Subtipo_delito=ROBO A PERSONAS Sexo_victima=masculino 1447 ==> Tipo_delito=CONTRA LA PROPIEDAD 1447    conf:(1)
 11. Subcircuito=ROCAFUERTE 1 Victima_denunciante=victima 1405 ==> Circuito=ROCAFUERTE 1405    conf:(1)
 12. Circuito=ROCAFUERTE Victima_denunciante=victima 1405 ==> Subcircuito=ROCAFUERTE 1 1405    conf:(1)
 13. Subtipo_delito=ROBO A PERSONAS Victima_denunciante=victima 1384 ==> Tipo_delito=CONTRA LA PROPIEDAD 1384    conf:(1)

=== Evaluation ===

Elapsed time: 0.41s
```

Figure 5-17: Part of the fourth experiment outcome using the Ecuador dataset.

As explained at the end of section 4-2-4, we narrow down the UK dataset by selecting the type of crime “Antisocial Behaviour”, and analysed it with the Apriori algorithm again. But, we obtained the same output as with the previous analysis (Figure 5-18).

```
No large itemsets and rules found!

=== Evaluation ===

Elapsed time: 558.915s
```

Figure 5-18: The fourth experiment outcome using the UK dataset selecting the Antisocial Behaviour Crime.

5.5. Summary

As explained in previous sections, this chapter described the implementation of the procedures using the requirements described in prior chapters.

Four different algorithms were used to analyse the UK and Ecuador datasets, but they have been prepared differently for each experiment (Table 5-11).

Table 5-11: Summary of Algorithms and Datasets Implementation.

Country	Algorithm	Dataset	Experiment # : Dataset
UK	LR	Normal	1: Frequency Attribute.
		Crime Type	2: Most frequent Type of Crime.
		Transform	3: Converted LSOA-Code to Postcode.
	LWL	Normal	1: Frequency Attribute.
		Crime Type	2: Most frequent Type of Crime.
		Transform	3: Converted LSOA-Code to Postcode.
	M5P	Normal	1: Frequency Attribute.
		Crime Type	2: Most frequent Type of Crime.
		Transform	3: Converted LSOA-Code to Postcode.
	Apriori	Normal	4a: General Dataset
		Crime Type	4b: Most frequent Type of Crime
Ecuador	LR	Normal	1: Frequency Attribute.
		Crime Type	2: Most frequent Type of Crime.
	LWL	Normal	1: Frequency Attribute.
		Crime Type	2: Most frequent Type of Crime.
	M5P	Normal	1: Frequency Attribute.
		Crime Type	2: Most frequent Type of Crime.
	Apriori	Normal	4a: General Dataset

6. Experiment Results

The following chapter will evaluate and discuss the results shown in the previous chapter.

The results will be evaluate in three sections; first, the correlation coefficient that describes how well the model performed, the mean absolute error, and the root mean squared error that describe the error between the predicted data and the actual data.

The second section will analyse and explained the association rules using the apriori algorithm. Finally, the last section will analyse how the time consumed for each analysis changes depending on the amount of data.

6.1. Measures of the models created using Linear Regression, LWL and M5P algorithms.

Three different experiments were carried out on the process of obtaining the most accurate model between three algorithms. Figure 6-1 illustrates the correlation coefficient for the models obtained through the three experiments using the UK datasets (normal, crime type and transform) generated for each experiment, and the LWL, LR and M5P algorithms. In like manner, Figure 6-2 illustrates the correlation coefficients for the models obtained through two experiments using the Ecuador dataset and the same algorithms.

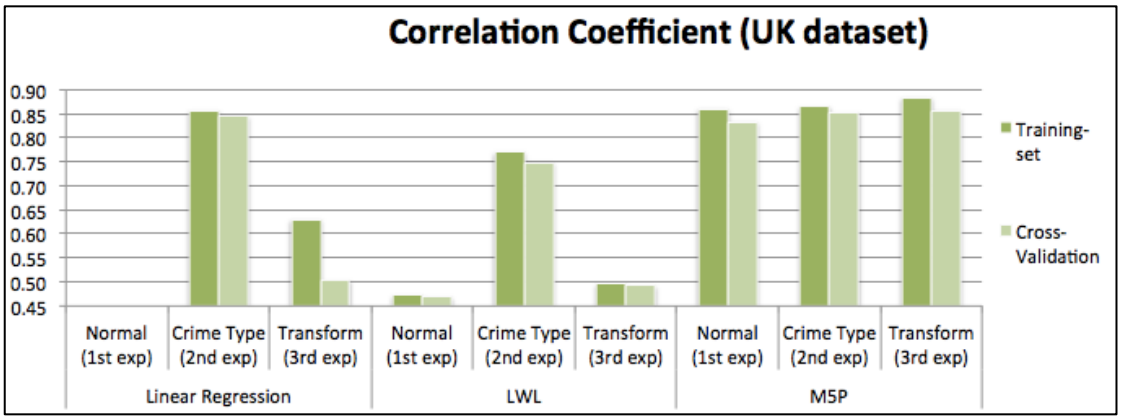


Figure 6-1: Correlation Coefficient for the UK dataset.

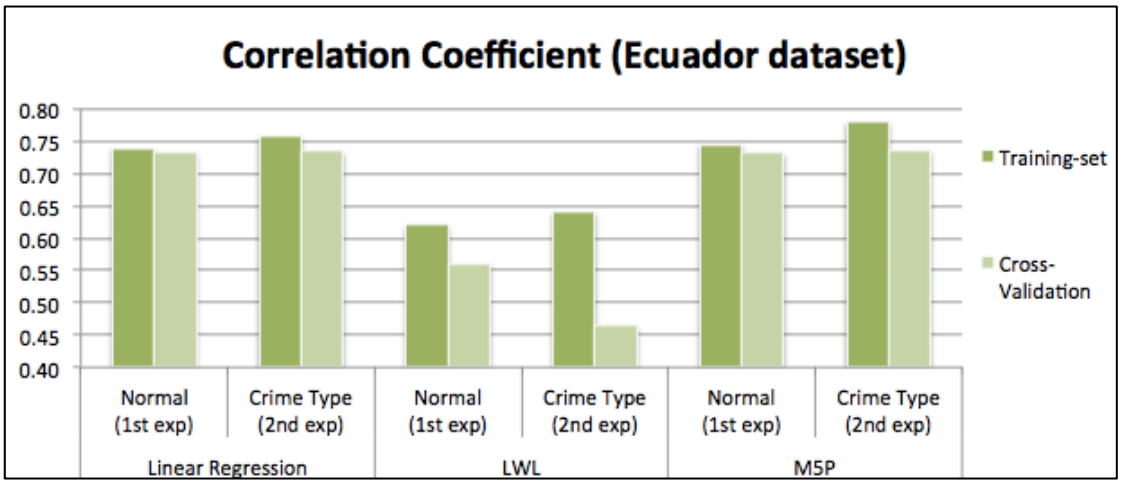


Figure 6-2: Correlation Coefficient for the Ecuador dataset.

As explained in Section 3.5, if the coefficient is approaching 1 then, we can say that the model has a strong correlation. Having this in mind, we can analyse the Figures above, and find that for the UK dataset (Figure 6-1) the models with the strongest correlation on the three experiments were developed by the M5P algorithm, and the one developed by the linear regression on the second experiment. Similarly, for the Ecuador dataset (Figure 6-2) the strongest correlation can be seen in the models developed by the LR, and M5P algorithms on every experiment.

On the other hand, note that the coefficients for the LR using the Ecuador dataset are similar on both experiments, while for the UK dataset; the

value becomes stronger when we selected the most frequent crime type (second experiment).

Overall, both Figures illustrate that the second experiment has better coefficients than the other ones using any of the algorithms.

In addition to the correlation, it is also necessary to compare the error values in order to analyse the accuracy of each model. Figures 6-3 and 6-4 show the MAE and Figures 6-5 and 6-6 show the RMSE for each experiment using the UK and Ecuador dataset respectively.

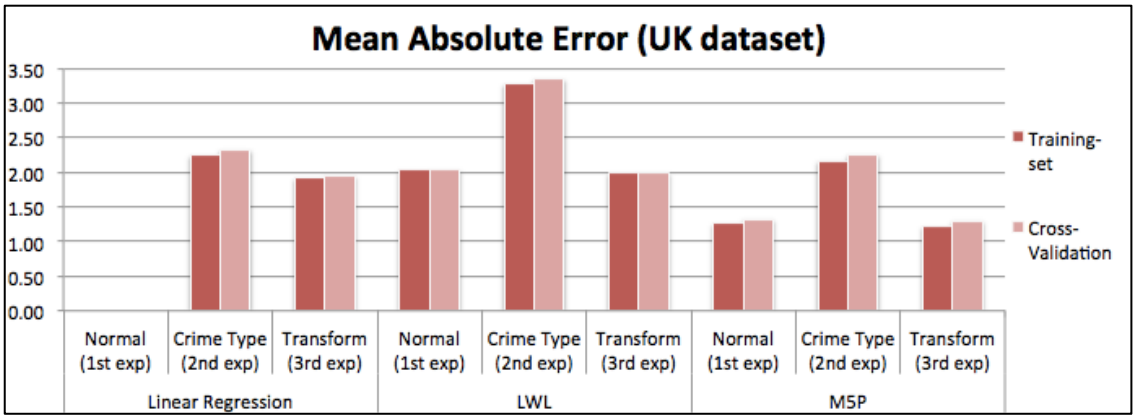


Figure 6-3: Mean Absolute Error for the UK dataset.

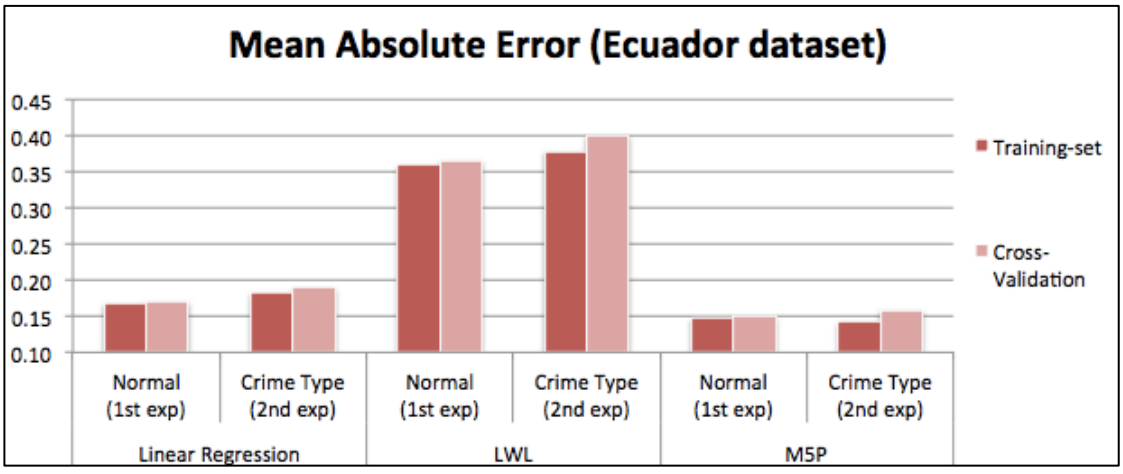


Figure 6-4: Mean Absolute Error for the Ecuador dataset.

Despite the fact that the MAE values for the UK dataset are higher than the values for the Ecuador dataset, we cannot say that the UK models are

worst than the Ecuador ones, because the outcomes on every algorithm will vary depending on the differences of each dataset. But we can clearly identify lower values on the LR and M5P algorithms for the Ecuador dataset and lower values on the first and third experiment using the M5P algorithm on the UK dataset. In like manner, we can analyse the following Figures, where the lower values are depicted on the same algorithms and experiments for the UK and Ecuador datasets as the MAE.

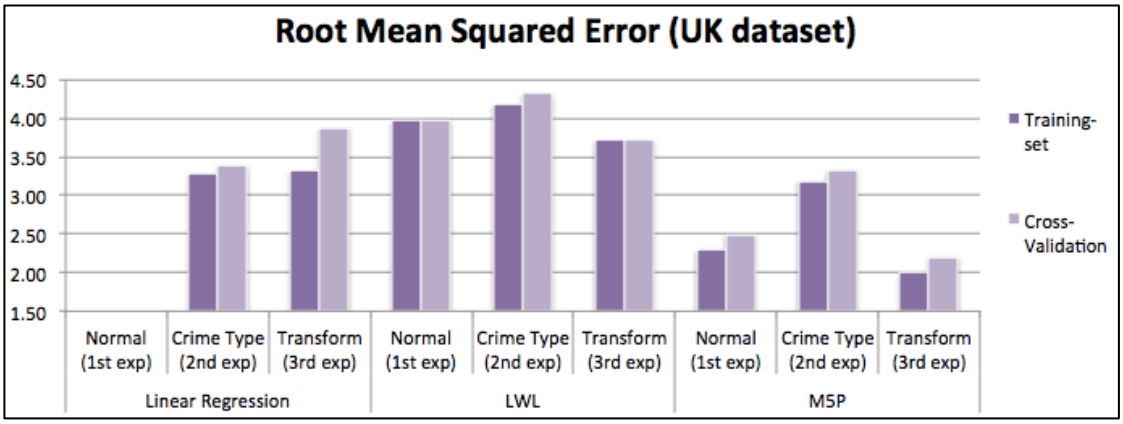


Figure 6-5: Root Mean Squared Error for the UK dataset.

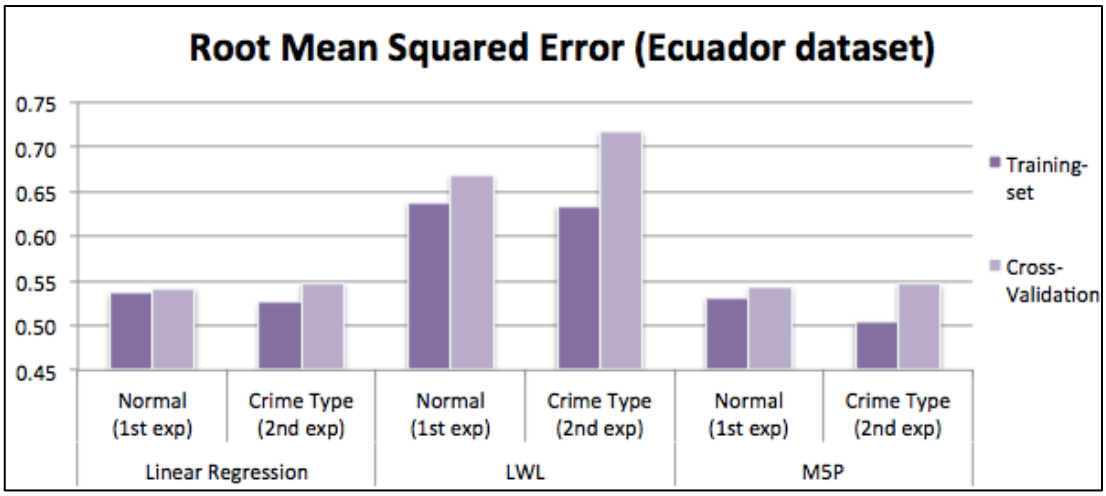


Figure 6-6: Root Mean Squared Error for the Ecuador dataset.

Analogous to the MAE Figures in the RMSE ones, we can identify lower values on the first and third experiment using the M5P algorithm on the UK

dataset (Figure 6-5), and lower values on the LR and M5P algorithms for the Ecuador dataset (Figure 6-6).

Overall, the M5P algorithm obtained correlation coefficient and error values (MAE, RMSE) on both datasets that achieve the main purpose of this project, which is to develop accurate models.

Figures 6-1, 6-3 and 6-5 did not shown the results on the first experiment using the UK dataset and the LR algorithm, because it is still analysing the data and so far (September 4th 2014) it has been 1090 hours of analysis.

Also, all Figures illustrate two different validation methods, which are training-set and 10-fold cross-validation. As explained in section 3.4.5, both of the methods assess the performance of the models.

Even though, all the Figures reflect that the training-set method gives better results (correlation coefficient, MAE and RMSE) than the Cross-validation method, according to Bouckaert et al (2014) the cross-validation gives more realistic measures of the expected accuracy on unseen data..

6.2. Analysis of the Linear Regression models created on every experiment.

Due to the size of the models developed on each experiment using the LR algorithm, Figures 6-7, 6-8, 6-9, and 6-10 depict a sample of each one of them, but the original ones will be attached in a CD to this report.

```
=== Classifier model ===


Linear Regression Model

FREQUENCE2 =

      0.9087 * Fecha_infraccion=2010-01,2010-08,2010-05,2010-04,2010-07,2010-11,2010-03,2010-02,2010-10,2010-09,2010-06 +
      0.531  * Fecha_infraccion=2010-08,2010-05,2010-04,2010-07,2010-11,2010-03,2010-02,2010-10,2010-09,2010-06 +
      0.0547 * Fecha_infraccion=2010-05,2010-04,2010-07,2010-11,2010-03,2010-02,2010-10,2010-09,2010-06 +
      0.1178 * Fecha_infraccion=2010-07,2010-11,2010-03,2010-02,2010-10,2010-09,2010-06 +
      0.0603 * Fecha_infraccion=2010-10,2010-09,2010-06 +
      0.2272 * Fecha_infraccion=2010-06 +
     -0.1544 * Circuito=MACHALA OCCIDENTAL,19 DE NOVIEMBRE,RAYITO DE LUZ,CRISTO DEL CONSUELO,SIMON BOLIVAR,ROCAFUERTE,LA CU
      0.0694 * Circuito=RAYITO DE LUZ,CRISTO DEL CONSUELO,SIMON BOLIVAR,ROCAFUERTE,LA CUATRO MIL,EL CAMBIO,PAEZ,27 DE FEBRE
      0.1153 * Circuito=SIMON BOLIVAR,ROCAFUERTE,LA CUATRO MIL,EL CAMBIO,PAEZ,27 DE FEBRERO,LAS KATYAS,PARQUE LINEAL,JAMBEL
      0.1057 * Circuito=ROCAFUERTE,LA CUATRO MIL,EL CAMBIO,PAEZ,27 DE FEBRERO,LAS KATYAS,PARQUE LINEAL,JAMBELI CENTRO,GRAL
     -0.0738 * Circuito=LA CUATRO MIL,EL CAMBIO,PAEZ,27 DE FEBRERO,LAS KATYAS,PARQUE LINEAL,JAMBELI CENTRO,GRAL SERRANO,9 D
      0.0912 * Circuito=EL CAMBIO,PAEZ,27 DE FEBRERO,LAS KATYAS,PARQUE LINEAL,JAMBELI CENTRO,GRAL SERRANO,9 DE MAYO,NUEVO P
     -0.21   * Circuito=PAEZ,27 DE FEBRERO,LAS KATYAS,PARQUE LINEAL,JAMBELI CENTRO,GRAL SERRANO,9 DE MAYO,NUEVO PILO +
     -0.6553 * Circuito=27 DE FEBRERO,LAS KATYAS,PARQUE LINEAL,JAMBELI CENTRO,GRAL SERRANO,9 DE MAYO,NUEVO PILO +
     -0.2091 * Circuito=LAS KATYAS,PARQUE LINEAL,JAMBELI CENTRO,GRAL SERRANO,9 DE MAYO,NUEVO PILO +
      1.3256 * Circuito=PARQUE LINEAL,JAMBELI CENTRO,GRAL SERRANO,9 DE MAYO,NUEVO PILO +
     -0.6497 * Circuito=JAMBELI CENTRO,GRAL SERRANO,9 DE MAYO,NUEVO PILO +
      0.4856 * Circuito=GRAL SERRANO,9 DE MAYO,NUEVO PILO +
     -1.2797 * Circuito=9 DE MAYO,NUEVO PILO +
      0.5126 * Circuito=NUEVO PILO +
      0.8461 * Subcircuito=LAS KATYAS 1,LA CUATRO MIL 2,RAYITO DE LUZ 1,MACHALA OCCIDENTAL 1,19 DE NOVIEMBRE 1,CRISTO DEL C
     -0.9094 * Subcircuito=LA CUATRO MIL 2,RAYITO DE LUZ 1,MACHALA OCCIDENTAL 1,19 DE NOVIEMBRE 1,CRISTO DEL CONSUELO 2,SIM
      0.1964 * Subcircuito=RAYITO DE LUZ 1,MACHALA OCCIDENTAL 1,19 DE NOVIEMBRE 1,CRISTO DEL CONSUELO 2,SIMON BOLIVAR 1,ROC
      0.0351 * Subcircuito=MACHALA OCCIDENTAL 1,19 DE NOVIEMBRE 1,CRISTO DEL CONSUELO 2,SIMON BOLIVAR 1,ROCAFUERTE 1,JAMBEL
     -0.0967 * Subcircuito=CRISTO DEL CONSUELO 2,SIMON BOLIVAR 1,ROCAFUERTE 1,JAMBELI CENTRO 1,CRISTO DEL CONSUELO 1,EL CAM
     -0.0633 * Subcircuito=SIMON BOLIVAR 1,ROCAFUERTE 1,JAMBELI CENTRO 1,CRISTO DEL CONSUELO 1,EL CAMBIO 1,PAEZ 1,LAS KATYA
     -0.1052 * Subcircuito=ROCAFUERTE 1,JAMBELI CENTRO 1,CRISTO DEL CONSUELO 1,EL CAMBIO 1,PAEZ 1,LAS KATYAS 3,27 DE FEBRER
      0.3123 * Subcircuito=JAMBELI CENTRO 1,CRISTO DEL CONSUELO 1,EL CAMBIO 1,PAEZ 1,LAS KATYAS 3,27 DE FEBRERO 1,LA CUATRO
     -0.2512 * Subcircuito=CRISTO DEL CONSUELO 1,EL CAMBIO 1,PAEZ 1,LAS KATYAS 3,27 DE FEBRERO 1,LA CUATRO MIL 1,PARQUE LIN
     -0.1365 * Subcircuito=EL CAMBIO 1,PAEZ 1,LAS KATYAS 3,27 DE FEBRERO 1,LA CUATRO MIL 1,PARQUE LINEAL 1,RAYITO DE LUZ 2,
      0.2906 * Subcircuito=PAEZ 1,LAS KATYAS 3,27 DE FEBRERO 1,LA CUATRO MIL 1,PARQUE LINEAL 1,RAYITO DE LUZ 2,9 DE MAYO 2,
```

Figure 6-7: Sample of the Linear Regression Model for the Ecuador Dataset in the First Experiment.

```
=== Classifier model ===


Linear Regression Model

Frequence =

      0.2876 * Month=2014-01,2013-12,2014-03,2012-12,2011-12,2011-01,2013-11,2010-12,2012-01,2013-02,2012-02,2012-11,2013
      0.7129 * Month=2013-12,2014-03,2012-12,2011-12,2011-01,2013-11,2010-12,2012-01,2013-02,2012-02,2012-11,2013-03,2011
      0.3107 * Month=2012-12,2011-12,2011-01,2013-11,2010-12,2012-01,2013-02,2012-02,2012-11,2013-03,2011-02,2012-04,2013
      0.1254 * Month=2010-12,2012-01,2013-02,2012-02,2012-11,2013-03,2011-02,2012-04,2013-04,2011-11,2013-10,2013-01,2011
      0.251  * Month=2012-02,2012-11,2013-03,2011-02,2012-04,2013-04,2011-11,2013-10,2013-01,2011-03,2013-09,2012-10,2013
      0.1711 * Month=2012-04,2013-04,2011-11,2013-10,2013-01,2011-03,2013-09,2012-10,2013-05,2012-03,2011-09,2011-06,2012
      0.1221 * Month=2011-11,2013-10,2013-01,2011-03,2013-09,2012-10,2013-05,2012-03,2011-09,2011-06,2012-06,2012-09,2011
      0.1979 * Month=2013-01,2011-03,2013-09,2012-10,2013-05,2012-03,2011-09,2011-06,2012-06,2012-09,2011-10,2012-05,2013
      0.1522 * Month=2011-03,2013-09,2012-10,2013-05,2012-03,2011-09,2011-06,2012-06,2012-09,2011-10,2012-05,2013-06,2011
      0.0852 * Month=2013-05,2012-03,2011-09,2011-06,2012-06,2012-09,2011-10,2012-05,2013-06,2011-05,2013-08,2012-07,2011
      0.2235 * Month=2012-03,2011-09,2011-06,2012-06,2012-09,2011-10,2012-05,2013-06,2011-05,2013-08,2012-07,2011-08,2013
      0.2629 * Month=2012-06,2012-09,2011-10,2012-05,2013-06,2011-05,2013-08,2012-07,2011-08,2013-07,2011-07,2011-04,2012
      0.1001 * Month=2011-10,2012-05,2013-06,2011-05,2013-08,2012-07,2011-08,2013-07,2011-07,2011-04,2012-08 +
      0.3408 * Month=2011-05,2013-08,2012-07,2011-08,2013-07,2011-07,2011-04,2012-08 +
      0.2777 * Month=2013-08,2012-07,2011-08,2013-07,2011-07,2011-04,2012-08 +
      0.1401 * Month=2012-07,2011-08,2013-07,2011-07,2011-04,2012-08 +
      0.1804 * Month=2013-07,2011-07,2011-04,2012-08 +
      0.3054 * Month=2012-08 +
      1.0284 * LSOA Code=E01016302,E01031591,E01030935,E01030902,E01030896,E01030893,E01016194,E01016236,E01020345,E01020
     -1.4938 * LSOA Code=E01030902,E01030896,E01030893,E01016194,E01016236,E01020345,E01020349,E01030805,E01030942,E01031
     -0.9757 * LSOA Code=E01032697,E01031434,E01016199,E01030930,E01016294,E01031535,E01030903,E01031534,E01031995,E01016
      1.1011 * LSOA Code=E01031434,E01016199,E01030930,E01016294,E01031535,E01030903,E01031534,E01031995,E01016689,E01030
     -0.8757 * LSOA Code=E01016294,E01031535,E01030903,E01031534,E01031995,E01016689,E01030767,E01030774,E01030890,E01030
     -1.288  * LSOA Code=E01031995,E01016689,E01030767,E01030774,E01030890,E01030770,E01031507,E01028663,E01020404,E01031
      1.0816 * LSOA Code=E01030767,E01030774,E01030890,E01030770,E01031507,E01028663,E01020404,E01031860,E01016301,E01016
     -1.1719 * LSOA Code=E01028663,E01020404,E01031860,E01016301,E01016366,E01030929,E01031508,E01020411,E01020358,E01016
      1.1914 * LSOA Code=E01031860,E01016301,E01016366,E01030929,E01031508,E01020411,E01020358,E01016643,E01030432,E01032
```

Figure 6-8: Sample of the Linear Regression Model for the UK Dataset in the Second Experiment.

```
=== Classifier model ===


Linear Regression Model

FREQUENCE =

      0.0594 * Fecha_infraccion=2013-04,2010-01,2010-04,2010-02,2010-05,2010-08,2010-07,2010-09,2010-03,2010-11,2010-10,
      0.8272 * Fecha_infraccion=2010-01,2010-04,2010-02,2010-05,2010-08,2010-07,2010-09,2010-03,2010-11,2010-10,2010-06
      0.46   * Fecha_infraccion=2010-04,2010-02,2010-05,2010-08,2010-07,2010-09,2010-03,2010-11,2010-10,2010-06 +
      0.1192 * Fecha_infraccion=2010-02,2010-05,2010-08,2010-07,2010-09,2010-03,2010-11,2010-10,2010-06 +
      0.1254 * Fecha_infraccion=2010-07,2010-09,2010-03,2010-11,2010-10,2010-06 +
      0.0896 * Fecha_infraccion=2010-03,2010-11,2010-10,2010-06 +
      0.085  * Fecha_infraccion=2010-11,2010-10,2010-06 +
      0.2955 * Fecha_infraccion=2010-06 +
      0.0436 * Circuito=19 DE NOVIEMBRE,EL CAMBIO,PUERTO BOLIVAR,SIMON BOLIVAR,LAS KATYAS,ROCAFUERTE,PAEZ,MACHALA OCCIDE
      0.0813 * Circuito=ROCAFUERTE,PAEZ,MACHALA OCCIDENTAL,JAMBELI CENTRO,27 DE FEBRERO,LA CUATRO MIL,GRAL SERRANO,PARQU
      0.061  * Circuito=27 DE FEBRERO,LA CUATRO MIL,GRAL SERRANO,PARQUE LINEAL,NUEVO PILO,9 DE MAYO +
     -0.045  * Subcircuito=PUERTO BOLIVAR 1,LAS KATYAS 1,19 DE NOVIEMBRE 1,EL CAMBIO 1,SIMON BOLIVAR 1,RAYITO DE LUZ 2,C
     -0.0628 * Subcircuito=LAS KATYAS 1,19 DE NOVIEMBRE 1,EL CAMBIO 1,SIMON BOLIVAR 1,RAYITO DE LUZ 2,CRISTO DEL CONSUEL
      0.0999 * Subcircuito=19 DE NOVIEMBRE 1,EL CAMBIO 1,SIMON BOLIVAR 1,RAYITO DE LUZ 2,CRISTO DEL CONSUELO 1,JAMBELI C
     -0.0414 * Subcircuito=EL CAMBIO 1,SIMON BOLIVAR 1,RAYITO DE LUZ 2,CRISTO DEL CONSUELO 1,JAMBELI CENTRO 1,ROCAFUERTE
      0.0703 * Subcircuito=RAYITO DE LUZ 2,CRISTO DEL CONSUELO 1,JAMBELI CENTRO 1,ROCAFUERTE 1,PAEZ 1,MACHALA OCCIDENTAL
     -0.2036 * Subcircuito=CRISTO DEL CONSUELO 1,JAMBELI CENTRO 1,ROCAFUERTE 1,PAEZ 1,MACHALA OCCIDENTAL 1,LA CUATRO MIL
      0.0956 * Subcircuito=JAMBELI CENTRO 1,ROCAFUERTE 1,PAEZ 1,MACHALA OCCIDENTAL 1,LA CUATRO MIL 2,27 DE FEBRERO 1,GRA
      0.0404 * Subcircuito=ROCAFUERTE 1,PAEZ 1,MACHALA OCCIDENTAL 1,LA CUATRO MIL 2,27 DE FEBRERO 1,GRAL SERRANO 1,JAMBE
     -0.0585 * Subcircuito=LA CUATRO MIL 2,27 DE FEBRERO 1,GRAL SERRANO 1,JAMBELI CENTRO 2,LAS KATYAS 2,PARQUE LINEAL 1,
     -0.0934 * Subcircuito=JAMBELI CENTRO 2,LAS KATYAS 2,PARQUE LINEAL 1,LA CUATRO MIL 1,9 DE MAYO 2,NUEVO PILO 1,9 DE M
      0.0828 * Subcircuito=LAS KATYAS 2,PARQUE LINEAL 1,LA CUATRO MIL 1,9 DE MAYO 2,NUEVO PILO 1,9 DE MAYO 1,PUERTO BOLI
      0.3734 * Subcircuito=9 DE MAYO 1,PUERTO BOLIVAR 2 +
     -0.3075 * Subcircuito=PUERTO BOLIVAR 2 +
      0.0372 * Victima_denunciante=victima +
      0.9303
```

Figure 6-9: Linear Regression Model for the Ecuador Dataset in the Second Experiment.

```
=== Classifier model ===


Linear Regression Model

Frequence per month =

      0.1834 * Month=2013-12,2014-03,2013-11,2012-12,2013-10,2013-02,2013-09,2012-11,2012-01,2013-04,2013-05,2012-02,20
      0.1364 * Month=2013-11,2012-12,2013-10,2013-02,2013-09,2012-11,2012-01,2013-04,2013-05,2012-02,2013-03,2011-12,20
     -0.1002 * Month=2012-12,2013-10,2013-02,2013-09,2012-11,2012-01,2013-04,2013-05,2012-02,2013-03,2011-12,2012-04,20
      0.2052 * Month=2013-10,2013-02,2013-09,2012-11,2012-01,2013-04,2013-05,2012-02,2013-03,2011-12,2012-04,2013-01,20
     -0.1328 * Month=2013-02,2013-09,2012-11,2012-01,2013-04,2013-05,2012-02,2013-03,2011-12,2012-04,2013-01,2012-10,20
      0.2048 * Month=2013-09,2012-11,2012-01,2013-04,2013-05,2012-02,2013-03,2011-12,2012-04,2013-01,2012-10,2011-11,20
     -0.1227 * Month=2012-11,2012-01,2013-04,2013-05,2012-02,2013-03,2011-12,2012-04,2013-01,2012-10,2011-11,2013-06,20
      0.1136 * Month=2012-01,2013-04,2013-05,2012-02,2013-03,2011-12,2012-04,2013-01,2012-10,2011-11,2013-06,2012-03,20
     -0.0669 * Month=2013-04,2013-05,2012-02,2013-03,2011-12,2012-04,2013-01,2012-10,2011-11,2013-06,2012-03,2012-09,20
      0.081  * Month=2013-05,2012-02,2013-03,2011-12,2012-04,2013-01,2012-10,2011-11,2013-06,2012-03,2012-09,2012-06,20
     -0.0904 * Month=2013-03,2011-12,2012-04,2013-01,2012-10,2011-11,2013-06,2012-03,2012-09,2012-06,2013-08,2012-05,20
      0.0697 * Month=2011-12,2012-04,2013-01,2012-10,2011-11,2013-06,2012-03,2012-09,2012-06,2013-08,2012-05,2013-07,20
      0.0633 * Month=2012-04,2013-01,2012-10,2011-11,2013-06,2012-03,2012-09,2012-06,2013-08,2012-05,2013-07,2011-09,20
     -0.0702 * Month=2013-01,2012-10,2011-11,2013-06,2012-03,2012-09,2012-06,2013-08,2012-05,2013-07,2011-09,2011-10,20
      0.132  * Month=2012-10,2011-11,2013-06,2012-03,2012-09,2012-06,2013-08,2012-05,2013-07,2011-09,2011-10,2012-07,20
      0.0986 * Month=2011-11,2013-06,2012-03,2012-09,2012-06,2013-08,2012-05,2013-07,2011-09,2011-10,2012-07,2012-08,20
      0.0832 * Month=2012-06,2013-08,2012-05,2013-07,2011-09,2011-10,2012-07,2012-08,2010-12,2011-01,2011-02,2011-06,20
      0.066  * Month=2012-05,2013-07,2011-09,2011-10,2012-07,2012-08,2010-12,2011-01,2011-02,2011-06,2011-03,2011-05,20
      0.084  * Month=2013-07,2011-09,2011-10,2012-07,2012-08,2010-12,2011-01,2011-02,2011-06,2011-03,2011-05,2011-08,20
     -0.0929 * Month=2012-07,2012-08,2010-12,2011-01,2011-02,2011-06,2011-03,2011-05,2011-08,2011-04,2011-07 +
      0.137  * Month=2012-08,2010-12,2011-01,2011-02,2011-06,2011-03,2011-05,2011-08,2011-04,2011-07 +
      0.1013 * Month=2010-12,2011-01,2011-02,2011-06,2011-03,2011-05,2011-08,2011-04,2011-07 +
```

Figure 6-10: Sample of the Linear Regression Model for the UK Dataset in the Third Experiment.

To validate the models and explain how they work, an example is done using the model (Figure 6-9) for the Ecuador dataset in the second experiment. Following the model and using a random instance (548th instance) of the Ecuador Dataset (Table 6-1), the predicted value is generated as Figure 6-11 shown.

Table 6-1: Instance 548th from the Ecuador Dataset of the Second Experiment.

Attributes	548th Instance
Fecha_infraccion	2011-02
Circuito	Rocafuerte
Subcircuito	Rocafuerte 1
Agresion	Fisica
Victima_denunciante	Victima
Sexo_victima	Masculino
Frequency	1

For example, the first line of the regression model "*0.0594 * Fecha_infraccion=2013-04,2010-01,2010-04,2010-02,2010-05,2010-08,2010-07,2010-09,2010-03,2010-11,2010-10,2010-06 +",* means that if the crime was committed within those dates, it places a 1 if not a 0. Therefore, for the 548th row of instances in the Ecuador dataset, this model predicted the output of *1.0458* on a crime with an actual frequency of *1* (Figure 6-11) validating the accuracy of the model.

```
Data (548) = 2011-02, Rocafuerte, Rocafuerte 1, Fisica, Victima, Masculino, 1

Linear Regression Model

FREQUENCY =
     0.0594 * 0 +
     0.8272 * 0 +
     0.46   * 0 +
     0.1192 * 0 +
     0.1254 * 0 +
     0.0896 * 0 +
     0.085  * 0 +
     0.2955 * 0 +
     0.0436 * 1 +
     0.0813 * 1 +
     0.061  * 0 +
    -0.045  * 1 +
    -0.0628 * 1 +
     0.0999 * 1 +
    -0.0414 * 1 +
     0.0703 * 1 +
    -0.2036 * 1 +
     0.0956 * 1 +
     0.0404 * 1 +
    -0.0585 * 0 +
    -0.0934 * 0 +
     0.0828 * 0 +
     0.3734 * 0 +
    -0.3075 * 0 +
     0.0372 * 1 +
     0.9303
Expected value (Frequency) = 1
Regression Model Output = 1.0458
```

Figure 6-11: The use of the 548th instance with the linear regression model obtained in the second experiment

The same procedure can be follow with the rest of the linear regression models and datasets to validate the predicted vs. the expected value.

6.3. Analysis of the Pruned Model Trees created on every experiment.

Based on the UK and Ecuador dataset, the M5 algorithm generates pruned model trees for every experiment (Figure 6-12, 6-13, and 6-14). Each model tree are formed by several linear models in their final leaves, for example Figure 6-12 and 6-14 depict the pruned model trees for the first and second experiments using the Ecuador dataset and they are formed by two and four Linear Models respectively. Similarly the pruned model trees for the first, second (Figure 6-13), and third experiments using the UK dataset are formed by 82, 6 and 164 Linear Models respectively.

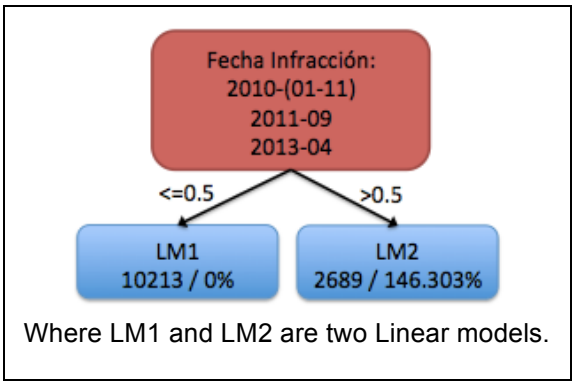
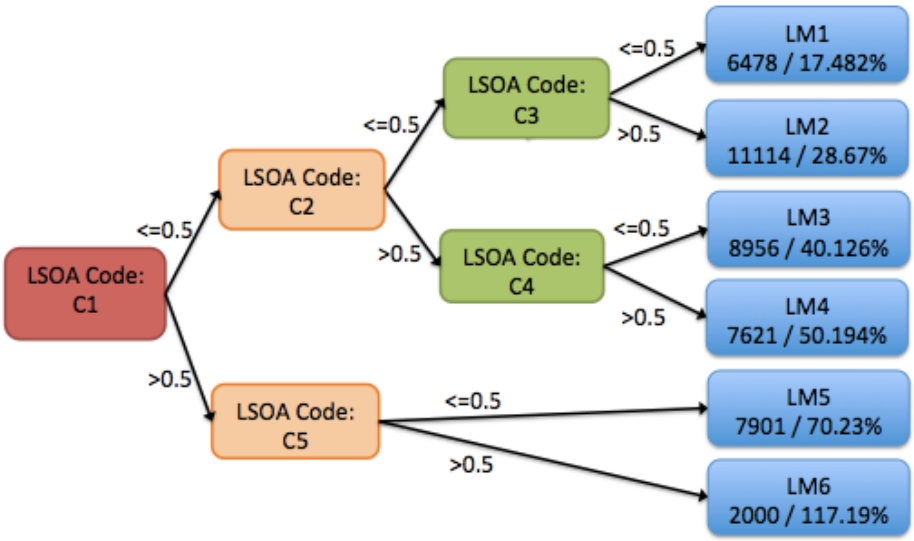
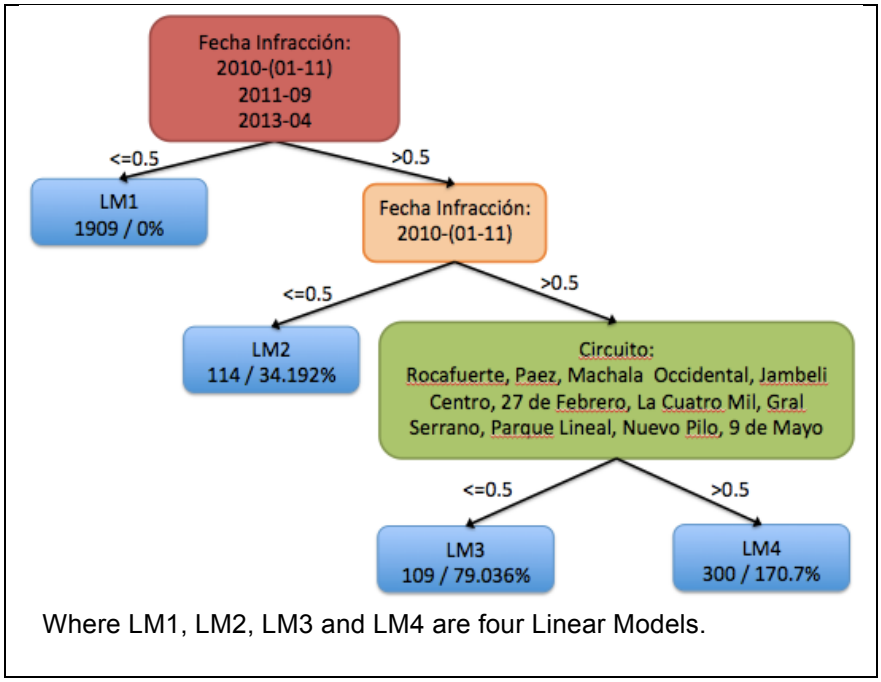


Figure 6-12: The first experiment pruned model tree for the Ecuador Dataset.



Where C1, C2, C3, C4 and C5 are different ranges of LSOA Code, and LM1, LM2, LM3, LM4, LM5 and LM6 are six Linear Models.

Figure 6-13: The second experiment pruned model tree for the UK dataset



Where LM1, LM2, LM3 and LM4 are four Linear Models.

Figure 6-14: The second experiment pruned model tree for the Ecuador Dataset.

Similarly as section 6.2, to validate the models and explain how they work, an example is done using the model (Figure 6-13) for the UK dataset in the second experiment. Following the tree model and using a random instance (instance 32nd) of the UK Dataset (Table 6-2), the predicted value is generated as Figure 6-16 shown.

Table 6-2: Instance 33rd from the UK Dataset of the Second Experiment.

Attributes	32nd Instance
Month	2010-12
LSOA Code	E01017045
LSOA name	Portsmouth 012C
Frequency	2

For example, the twenty-first line of the linear model number 6 (LM6) "*+ 0.0001 * Month=2011-10,2012-05,2013-06,2011-05,2013-08,2012-07,2011-08,2013-07,2011-07,2011-04,2012-08",* means that if the crime was committed within those dates, it places a 1 if not a 0. Therefore, for the 32nd row of instances in the UK dataset, this model predicted the output of *2.3897* on a crime with an actual frequency of *2* (Figure 6-11) validating the accuracy of the model.

Using the value of the LSOA Code of the 32nd instance, it follows down the prune model tree (Figure 6-15) comparing its value with each leave, which is not similar (<=0.5) to any of the LSOA Code of C1. Continuing it is not similar (<=0.5) to C2, but similar (>0.5) to C3, therefore for this instance the output is generated using the Linear Model two (Figure 6-16).

Data (32) = 2010-02, E01017045, Portsmouth 012C, 2

M5 pruned model tree:
(using smoothed linear models)

LSOA Code C1 = E01017317, E01017108, E01017325, E01017262, E01017322, E01022732, E01017158, E01017079, E01023118, E01017142, E01023042, E01022824, E01017348, E01017181, E01017075, E01017217, E01022800, E01023255… <= 0.5 :
| LSOA Code C2 = E01022889, E01022740, E01022757, E01022685, E01022943, E01017215, E01023258, E01023182, E01022926, E01017016, E01022947, E01017359, E01023178, E01022924 <= 0.5 :
| | LSOA Code C3 = E01022534, E01017369, E01022747, E01022792, E01022565, E01032857, E01023069, E01022590, E01017145, E01023274, E01017232, E01017303, E01017045… <= 0.5 : LM1 (6478/17.482%)
| | LSOA Code C3 = E01022534, E01017369, E01022747, E01022792, E01022565, E01032857, E01023069, E01022590, E01017145, E01023274, E01017232, E01017303, E01017045… > 0.5 : LM2 (11114/28.675%)
| LSOA Code C2 = E01022889, E01022740, E01022757, E01022685, E01022943, E01017215, E01023258, E01023182, E01022926, E01017016, E01022947, E01017359, E01023178, E01022924… > 0.5 :
| | LSOA Code C4 = E01022746, E01022657, E01017204, E01022795, E01023201, E01022952, E01017241, E01022891, E01023150, E01022722, E01022678, E01017175, E01022738 … <= 0.5 : LM3 (8956/40.126%)
| | LSOA Code C4 = E01022746, E01022657, E01017204, E01022795, E01023201, E01022952, E01017241, E01022891, E01023150, E01022722, E01022678, E01017175, E01022738… > 0.5 : LM4 (7621/50.194%)
LSOA Code C1 = E01017317, E01017108, E01017325, E01017262, E01017322, E01022732, E01017158, E01017079, E01023118, E01017142, E01023042, E01022824, E01017348, E01017181, E01017075, E01017217, E01022800, E01023255… > 0.5 :
| LSOA Code C5 = E01017099, E01017168, E01017345, E01032748, E01022900, E01017027, E01017036, E01017035, E01017077, E01023153, E01022862, E01023264, E01023077… <= 0.5 : LM5 (7901/70.23%)
| LSOA Code C5 = E01017099, E01017168, E01017345, E01032748, E01022900, E01017027, E01017036, E01017035, E01017077, E01023153, E01022862, E01023264, E01023077… > 0.5 : LM6 (2000/117.19%)

Figure 6-15: The use of the pruned decision tree on the 33rd instance.

Data (32) = 2010-02, E01017045, Portsmouth 012C, 2

LM num: 2
Frequency =
0.0001 * 0 + 0.3498 * 0 + (…) * 0 + 0.0012 * 1 + 0.0005 * 1 - 0.0004 * 1 - 0.0008 * 1 + 0.001 * 1 - 0.0003 * 1 + 0.0005 * 1 - 0.0003 * 1 + 0.0005 * 1 - 0.0002 * 1 + 0.0006 * 1 - 0.0002 * 1 + 0.0001 * 1 - 0.0006 * 1 + 0.0006 * 1 + 0.0002 * 1 - 0.0005 * 1 + 0.0005 * 1 - 0.0006 * 1 + 0.0005 * 1 + 0.0001 * 1 - 0.0001 * + 0.0001 * 1 - 0.0003 * 1 + 0.0002 * 1 + 0.0001 * 1 + 0.0002 * 1 - 0.0004 * 1 + 0.0002 * 1 - 0.0002 * 1 + 0.0005 * 1 - 0.0002 * 1 + 0.0002 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0002 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0003 * 1 + 0.0005 * 1 - 0.0002 * 1 + 0.0002 * 1 + 0.0004 * 1 - 0.0005 * 1 + 0.0003 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0002 * 1 + 0.0002 * 1 - 0.0001 * 1 - 0.0002 * 1 + 0.0001 * 1 + 0.0002 * 1 - 0.0003 * 1 + 0.0004 * 1 - 0.0001 * 1 + 0.0002 * 1 - 0.0002 * 1 + 0.0001 * 1 + 0.0003 * 1 - 0.0002 * 1 + 0.0001 * 1 + 0.0002 * 1 - 0.0001 * 1 + 0.0001 * 1 + 0.0001 * 1 + 0.0001 * 1 - 0.0004 * 1 + 0.0001 * 1 + 0.0004 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0005 * 1 - 0.0003 * 1 + 0.0001 * 1 - 0.0004 * 1 + 0.0005 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 - 0.0002 * 1 + 0.0003 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.1566 * 1 - 0.0004 * 1 + 0.0003 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 + 0.0003 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0002 * 1 + 0.1435 * 1 + 0.0002 * 1- 0.0002 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 + 0.0002 * 1 + 0.1214 * 1 - 0.0002 * 1 + 0.0001 * 1 + 0.0001 * 1 - 0.0002 * 1 - 0.0001 * 1 + 0.0003 * 1 - 0.0001 * 1 + 0.0002 * 1 - 0.0001 * 1 + 0.0002 * 1 + 0.1132 * 1 - 0.0001 * 1 + 0.0002 * 1 - 0.0001 * 1 + 0.1192 * 1 + 0.0001 * 1 - 0.0002 * 1 + 0.0002 * 1 - 0.0001 * 1 + 0.0002 * 1 + 0 * 1 - 0.0001 * 1 + 0.1701 * 1 + 0 * 1 - 0.0001 * 1 + 0.0001 * 1 + 0.0001 * 1 - 0.0001 * 1 + 0.0001 * 1 + (…) * 0 + 1.5607

Expected value (Frequency) = 2
Regression Model Output = 2.3897

Figure 6-16: The use of the 32nd instance with the Linear Model 2 (LM2) of the decision tree model obtained in the second experiment.

The same procedure can be follow with the rest of the tree models and datasets to validate the predicted vs. the expected value.

It is observed the importance of the LSOA Code from the model trees of the UK datasets, for being the root node and it splits down with the same

attribute. For the Ecuador dataset, the important attribute is the date when the crime was or will be committed.

6.4. Results of the analysis using the Apriori Algorithm.

Even though, we selected the most frequent type of crime to narrow down the UK dataset, the Apriori algorithm still could not find any association rule that meet the minimum confidence of 0.9 and a minimum support of 0.1. On the contrary, the Ecuador dataset met 13 rules using the same algorithm and setting the same values for the minimum confidence and support.

Analysing Figure 5-17, for the Ecuador association rules we can discharge the rules 1, 3-8, 10, 11 and 13 because they go from a lower level of a category to an upper lower, for instance rule one goes from the subcircuit to a circuit. On the other hand, rules 2, 5, 7, 9 and 12 give us the idea that most of the crime committed in the Circuit “Rocafuerte” focus on the Subcircuit “Rocafuerte 1”. Figure 6-17 depicts the relationship between the features using the association rules obtained by the Apriori algorithm.

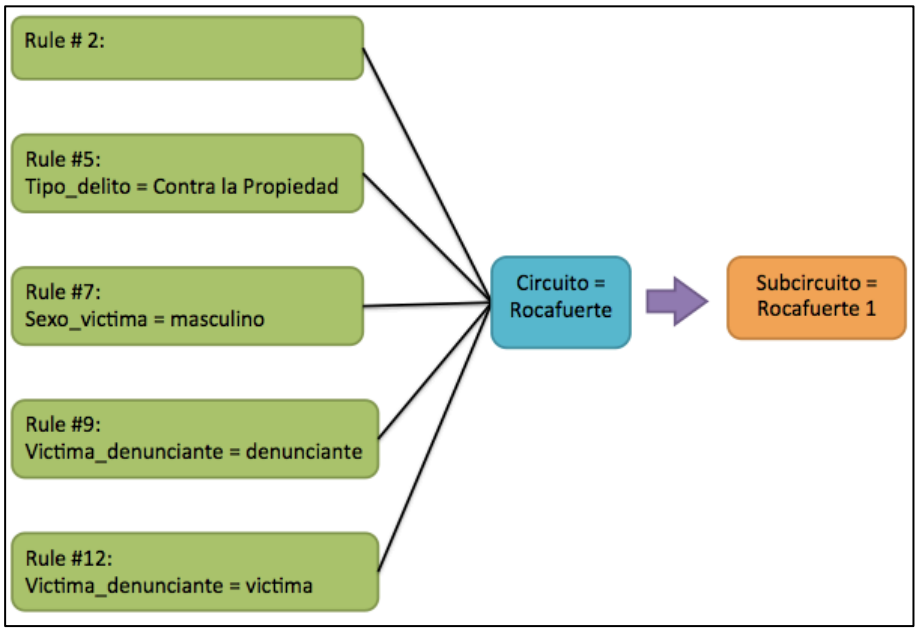


Figure 6-17: Summary of the association rules 2, 5, 7, 9 and 12 for the Ecuador dataset

6.5. Time vs. Data Objects.

Despite the fact that there were not results in the first experiment using the UK datasets and the LR algorithm due to the amount of time of analysis, its performance is not good enough to achieve the main objective of the project, which is to reduce the time of crime data analysis.

Figures 6-18 and 6-19 deals the number of instances of each dataset (blue color) and the time spent (red color) of each experiment. In both datasets the algorithm that spent more time to analyse and create the models is the LR. Even decreasing the number of instances on the second and third experiment for the UK dataset, the time spent on the analysis using the LR is still higher than the time spent on the other algorithms.

The time spent on analysing both datasets using the LWL algorithm did not vary a lot between the experiments, despite the fact that the number of instances changes between them.

On the other hand, as we saw in Figures 6-1 and 6-2 the M5P algorithm obtained good correlation coefficients, but it took less time to analyse the

Ecuador dataset than to analyse the UK dataset. This result does not mean that M5P works better on fewer instances, but gives us the outcome to analyse with this algorithm the UK data using fewer instances.

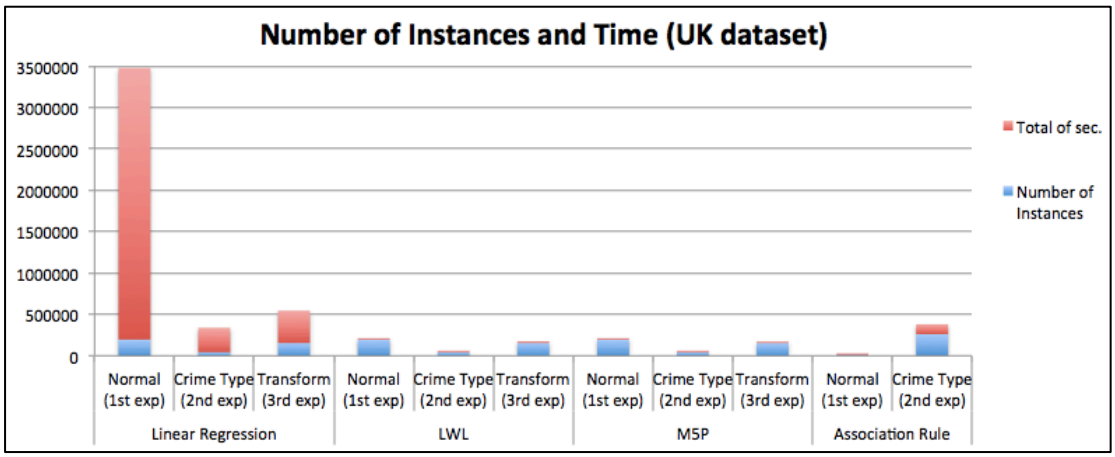


Figure 6-18: Instances and Time of every experiment for the UK dataset.

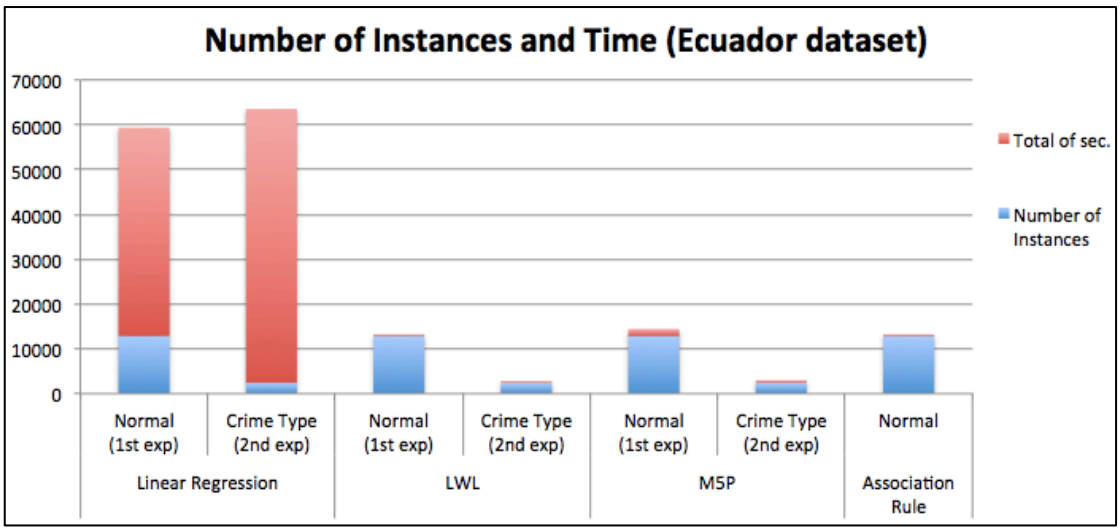


Figure 6-19: Instances and Time of every experiment for the Ecuador dataset.

7. Evaluation and Conclusion

7.1. Project Management.

The CRISP-DM standard described in Section 3.1, was selected to guide us through the process of executing this data mining project. Even though, it was used only five from the six stages, it helped organizing and structuring this project to improve the results and achieve the main objective of this project.

As chapter 4 described, some of the stages from the CRISP-DM, such as data preparation, modelling and evaluation were repeated, in order to execute different experiments and obtain more accurate models.

7.2. Problems Encountered and Recommendations.

During the execution of the project, we encountered several problems listed below:

- ***Project Time vs Analysing Data Time:*** The amount of time spent on the analysis of the data will depend on the algorithms, amount of data, and the computer features. Due to the three months to execute this project and the huge amount of data for the UK, it became challenging to wait, analyse and write about the results of every experiment. Therefore, it is important to start analysing the data early.
- ***Be organized:*** It was necessary to be organize to complete this project within the three months. The creation of a Gantt chart at the beginning of the project, keeping a logbook with the tasks to be done, creating a mind map for each of the section were a few of the documents created to successfully execute the project.

- ***Create copies of your work:*** Unfortunately, the laptop used to execute the project and write this report was having problems and therefore, it was difficult to continue with the work. On the other hand, a copy of the work was made on an external hard disk and using the computers of the Forensic Lab helped to continue the work.

7.3. Evaluation and Conclusions

This report presents the analysis of the datasets of the UK and Ecuador using four algorithms (LWL, LR, M5P, and Apriori) in four different experiments to create models that will predict crime.

The training-set, and 10-fold cross validation methods were used to evaluate each model and obtain their measures. The performance of the models were analysed through the measures; correlation coefficient, MAE, and RMSE.

Contrarily to the LWL algorithm, it was found that both the Linear Regression and M5P algorithms predict significantly well. However, M5P has shown better accuracy due to the low error values (MAE, RMSE).

Furthermore, the models were also evaluated over the time taken to create the model, and they shown that the linear regression algorithm took more time to develop the models than the others algorithms, including the LWL, which goes against our main objective (to reduce the time of the analysis).Besides, the models created by Linear Regression and M5P were validated by comparing the expected and predicted value using a random instance.

On the other hand, the Apriori algorithm generated five association rules for the Ecuador datastet, which lead us to think that most of the crime committed in the Circuit “Rocafuerte” focus on the Subcircuit “Rocafuerte

1”. In contrast, for the UK dataset no association rules were found using the same algorithm (Apriori).

Finally, through the experiments executed, it is clear that the crimes of the Hampshire County from the UK, and The canton of Machala from Ecuador can be predicted using data mining techniques. The police Departments of the UK and Ecuador can benefit from this project by using the models developed, which will decrease the amount of time on the analysis of the historical data.

7.4. Future studies.

As the results shown, the M5P algorithm developed the most accurate models in all the experiments, and due to the lack of time we were not able to improve them. Therefore, for a future work it may be interesting to search for the particular records where the models do not work, remove them and re-run all the experiments in order to improve them.

Bibliography

- Abernethy, M. (2010). *Data mining with WEKA, Part1: Introduction and regression.* Retrieved from http://www.ibm.com/developerworks/library/os-weka1/
- Agrawal, R., Imieliński, T., & Swami, A. N. (1993). Mining association rules between sets of items in large databases. *Sigmod Record.* DOI:10.1145/170035.170072
- Aljumah, A., Ahamad, M., & Siddiqui, M. (2012). *Application of data mining: Diabetes health care in young and old patients.* DOI: 10.1016/j.jksuci.2012.10.003
- Atkeson, C., Moore, A., & Schaal, S. (1997). *Locally Weighted Learning.* Retrieved from https://cs.brown.edu/courses/archive/2006-2007/cs195-5/extras/AMS.pdf.
- Bachner, J. (2013). *Predictive Policing: Preventing Crime with Data and Analytics.* Retrieved from http://lgdata.s3-website-us-east-1.amazonaws.com/docs/501/897499/nps67-100113-01.pdf
- Berry, M., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support.* New York: Wiley.
- Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2014). *WEKA Manual for Version 3-7-11.* Retrieved from http://prdownloads.sourceforge.net/weka/WekaManual-3-7-11.pdf?download
- Brossette, S., Sprague, A., Hardin, J., Waites, K., Jones, W., & Moser, S. (1998). *Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance.* DOI: 10.1136/jamia.1998.0050373
- Bruha, I., & Famili, A. (2000). Postprocessing in machine learning and data mining. *ACM SIGKDD Explorations Newsletter*, 2(2), 110-114.
- Çakır, A., Çalış, H., & Küçüksille, E. (2009). Data mining approach for supply unbalance detection in induction motor. *Expert Systems With Applications.* DOI: 10.1016/j.eswa.2009.04.006.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. Retrieved from ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.

- Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., & Chau, M. (2004). *Crime data mining: a general framework and some examples.* DOI: 10.1109/MC.2004.1297301
- Cunningham, S., & Holmes, G. (1999). *Developing innovative applications in agriculture using data mining.* Retrieved from https://perun.pmf.uns.ac.rs/old/radovanovic/dmsem/cd/install/Weka/doc/pubs/1999/99SJC-GH-Innovative-apps.pdf
- Data.police.uk (n.d.). *About data.police.uk.* Retrieved from http://data.police.uk/about/.
- Dolado, J. J., Rodriguez, D., Riquelme, J., Ferrer-Troyano, F., & Cuadrado, J. (2007). A two-stage zone Regression Method for Global Characterization of a Project Database. *Advances in Machine Learning Applications in Software Engineering,* 1.
- Dondanville, C., Zhang, X., & Lee, T. (2007). *Data Mining of Crime Research Information statistics Portal: The Experience and Lessons Learned.* DOI: 10.4018/978-1-59904-929-8.ch099.
- Du, H. (2010). *Data Mining techniques and applications: An introduction.* Andover: Cengage Learning.
- Englert, P. (n.d.). *Locally Weighted Learning.* Retrieved from http://www.ias.informatik.tu-darmstadt.de/uploads/Teaching/AutonomousLearningSystems/Englert_ALS_2012.pdf.
- Essays, UK. (2013). *Applications Of Data Mining In Insurance Sector Finance Essay.* Retrieved from http://www.ukessays.com/essays/finance/applications-of-data-mining-in-insurance-sector-finance-essay.php?cref=1
- Eumetcal. (n.d.a). *Correlation coefficient.* Retrieved from http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_cont_var/uos4/uos4_ko1.htm.
- Eumetcal. (n.d.b). *Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).* Retrieved from http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_cont_var/uos3/uos3_ko1.htm
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37-54.
- Guo, L. (2003). *Applying Data Mining Techniques in Property/Casualty Insurance.* Retrieved from: https://www.casact.org/pubs/forum/03wforum/03wf001.pdf
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The Weka Data Mining Software: An Update. *SIGKDD Explorations,* 11(1), 10-18. DOI: 10.1145/1656274.1656278

- Hinman, H. (2013). *9 Data Mining Challenges From Data Scientists Like You.* Retrieved from http://1.salford-systems.com/blog/bid/305673/9-Data-Mining-Challenges-From-Data-Scientists-Like-You.
- Hof, E. (2001). *Testing The Use Of Locally Weighted Learning Algorithm For Robot Navigation.* Retrieved from http://webee.technion.ac.il/control/info/Projects/Students/Hof/hof1.htm.
- Hussain, K. Z., Durairaj, M., & Farzana, G. R. (2012). Application of Data mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model. *International Journal of Computer Science and Information Technology & Security, 2*(1), 25-29.
- Jakkula, V. (2007). *Predictive Data Mining to Learn Health Vitals of a Resident in a Smart Home.* DOI: 10.1109/ICDMW.2007.57.
- Jermyn, P., Dixon, M., & Read, B. (1999). Preparing Clean Views of Data for Data Mining. *ERCIM Work. On Database Res,* 1-15.
- Jin, F., Wang, W., Xiao, Y., & Pan, Z. (n.d.). *Proposal of Crime Data Mining Project.* Retrieved from http://filebox.vt.edu/users/xykid/dataAnalysisProject/report.pdf
- Juca, M. (2010, April). Data mining with WEKA, Part 1: Introduction and regression [Web log post]. Retrieved from http://www.ibm.com/developerworks/library/os-weka1/
- Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms.*
- Kramer, S. (n.d.). *M5P.* Retrieved from http://www.opentox.org/dev/documentation/components/m5p.
- Malathi, A. & Santhosh, S. (2011). *An Enhanced Algorithm to Predict a Future Crime using Data Mining.* Retrieved from http://www.ijcaonline.org/volume21/number1/pxc3873335.pdf.
- Mathbits (n.d.). *Correlation Coefficient.* Retrieved from http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm.
- McCue, A. (2007). *Data Mining and predictive Analysis: Intelligence Gathering and Crime Analysis.* Retrieved from http://eds.b.ebscohost.com/eds/ebookviewer/ebook/ZTAwMHR3d19fMTczNTM4X19BTg2?sid=344a0b66-48e9-4811-8576-8c605e4fc653@sessionmgr110&vid=2&format=EB&lpid=lp_49&rid=0
- Moore, A. W., Atkeson, C. G., & Schaal, S. A. (1997). Locally weighted learning for control*. AI Review, 11*, 75-113.
- Moro, S., Laureano, R., & Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), *Proceedings of the European Simulation and Modelling Conference – ESM'2011* (pp. 117–121). Guimaraes, Portugal. EUROSIS.

- Nadali, A., Naghizadeh, E., & Nosratabadi, H. (2011). Evaluating the success level of data mining project based on CRIPS-DM Methodology by a Fuzzy Expert System. *Electronics Computer Technology (ICECT), 2011 3rd International Conference on, 6, 161-165.* DOI: 10.1109/ICECTECH.2011.5942073
- Nath, S. (2006). *Crime Pattern Detection using Data Mining.* DOI: 10.1109/WI-IATW.2006.55.
- Nordman, A. (n.d.). *Data Mining: Data And Preprocessing* [Lecture Notes]. Retrieved from the Department of Science and Technology website http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec2.pdf
- Office for National Statistics (2013). *Postcodes (Enumeration) (2011) to output areas (2011) to lower layer super output areas (2011) to middle layer super output areas (2011) to local authority districts (2011) E+W Lookup.* Retrieved from https://geoportal.statistics.gov.uk/geoportal/catalog/search/resource/details.page?uuid=%7B18444B52-47C2-40FE-8003-92230C344598%7D
- Oracle (n.d.). *Oracle Data Mining*. Retrieved from http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html.
- Ozgul, F., Atzenbeck, C., Celik, A., & Erdem, Z. (2011). *Incorporating data sources and methodologies for crime data mining.* doi:10.1109/ISI.2011.5983995.
- Quinlan, J. R. (1992, November). Learning with continuous classes. In *Proceedings of the 5th Australian joint Conference on Artificial Intelligence* (Vol. 92, pp. 343-348).
- Ranka, S. (2003). *Data Preprocessing* [Lecture Notes]. Retrieved from the Department of Computer & Information Science & Engineering website http://www.cise.ufl.edu/class/cis4930sp09dm/notes/dm2part2.pdf
- Riddel, P. (1998). *10-fold Cross Validation.* Retrieved from https://www.cs.auckland.ac.nz/~pat/706_98/ln/node119.html.
- SAS (n.d.). *Advanced Analytics Software*. Retrieved from http://www.sas.com/en_us/software/analytics.html.
- SIGKDD Website (2005). *The Weka team.* Retrieved from the SIGKDD Website: http://www.sigkdd.org/node/369.
- Skillicorn, D. (2009). *Knowledge discovery for counterterrorism and law enforcement.* Boca Raton: CRC
- Suryajaya, B., Aryani, F., Devarakonda, U., & Erwin, A. (2014). Research on E&P Efficiency metrics to support SKMIGAS Mission utilizing CRISP-DM Methodology. In F. God, B. Sowito, M. Bououdina & M. Chen. (Eds.), *Proceedings of the 2013 International Conference on Advances in Intelligent Systems in Bioinformatics*. Retrieved from http://www.atlantis-

press.com/php/pub.php?publication=intel-13&frame=http%3A//www.atlantis-press.com/php/paper-details.php%3Ffrom%3Dauthor+index%26id%3D11351%26querystr%3Dauthorstr%253DD%2526publication%253Dintel-13.

- Stackoverflow. (2011, October 3). What does correlation coefficient actually represent [Web log post]. Retrieved from http://stackoverflow.com/questions/7631799/what-does-correlation-coefficient-actually-represent
- StatSoft. (n.d.). *STATISTICA | Data mining software.* Retrieved from http://www.statsoft.com/Products/STATISTICA/Data-Miner.
- Técnico Lisboa (n.d.). *Association Rules Apriori Algorithm.* [Lecture Notes]. Retrieved from the Técnico Lisboa website https://fenix.tecnico.ulisboa.pt/downloadFile/3779571250083/licao_9.pdf.
- The R Foundation (n.d.). *The R Project for Statistical Computing*. Retrieved from http://www.r-project.org/
- Thongtae, P., & Srisuk, S. (2008). *An analysis of data mining applications in crime domain.* DOI: 10.1109/CIT.2008.Workshops.80
- Tumpowsky, J. (2013). *Understanding Today's Customer: How Data Mining Can Help Insurers Increase Customer Satisfaction.* Retrieved from http://blogs.cisco.com/financialservices/understanding-todays-customer-how-data-mining-can-help-insurers-increase-customer-satisfaction/
- Wasilewska, A. (2014). *Apriori Algorithm* [Lecture Notes]. Retrieved from Department of Computer Science of the Stony Brook University website http://www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf
- Wang, Y., & Witten, I. H. (1997, April). *Inducing model trees for continuous classes.* Retrieved from http://www.cs.waikato.ac.nz/~ml/publications/1997/Wang-Witten-Induct.pdf.
- Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). *Learning to Detect Patterns of Crime.* Retrieved from http://web.mit.edu/rudin/www/WangRuWaSeECML13.pdf.
- Wikipedia. (n.d.). *Cross-validation (statistics).* Retrieved September 1, 2014, from http://en.wikipedia.org/wiki/Cross-validation_(statistics)
- Witten, I. (2013). *Classification boundaries* [Lecture Notes]. Retrieved from the University of Waikato website http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/slides/Class4-DataMiningWithWeka-2013.pdf.
- Witten, I., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* Amsterdan [etc.: Elsevier Morgan Kaufmann.
- Yale University (n.d.). *Linear Regression* [Lecture Notes]. Retrieved from the Department of Statistics website http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

- Yu, C., Ward, M., Morabito, M., & Ding, W. (2011). *Crime Forecasting Using Data Mining Techniques.* DOI: 10.1109/ICDMW.2011.56.
- Zhan, C., Gan, A., & Hadi, M. (2011). Prediction of Lane Clearance Time of Freeway Incidents Using the M5P Tree Algorithm. *IEEE Transactions on Intelligent Transportation Systems.* DOI:10.1109/TITS.2011.2161634
- Zubi, Z., & Mahmud, A. (2013). Using Data Mining Techniques to analyze crime patterns in the Lybian National Crime Data*. Recent Advances in Image, Audio and Signal Processing.* Retrieved from http://www.wseas.us/e-library/conferences/2013/Budapest/IPASRE/IPASRE-09.pdf

APPENDIX A: Project Specification Document

School of Computing
Postgraduate Programme

MSc in Forensic Information Technology

Project Specification

Ginger Viviana Saltos Bernal

Project Specification

1. Basic details

Student name:	Ginger Viviana Saltos Bernal
Draft project title:	Predicting crime using Data Mining
Course and year:	MSc. Forensic Information Technology
Client organisation:	University of Portsmouth
Client contact name:	PhD. Mihaela Cocea
Project supervisor:	PhD. Mihaela Cocea

2. Outline of the project environment

This project does not currently include a client, however, the Police Department of the United Kingdom and Ecuador can benefit with the development of it.

In order to determine a pattern of crimes and anticipate criminals, Police Departments have to analyse a large amount of historical crime data. But, the analysis can take much time and effort therefore, the aim of this project is to create models using data mining techniques, to detect patterns and predict crimes in a city of UK and Ecuador, so they can anticipate criminals by minimizing time of data analysis over historical crimes.

Data Mining is an efficient process of the knowledge extraction from a large collection of data, by using different techniques.

To achieve the aim of this project, the following objectives have been identified:

- Review previous research related to crime data mining.
- Analyse the model process that have been used in this project, focusing on the algorithms used to create the model
- Acquire datasets from the Police Departments of the UK.
- Acquire datasets from the Police Departments of the Ecuador.

- Design and implement models that could detect patterns and predict crimes using the WEKA software.
- Evaluate the outcomes and conclude over the results.

3. The problem to be solved

Crime is increasing all over the world; this is why governments invest so much money and time in the security field every year. Nowadays, security is based on the crime analysis from historical data, however the amount of data is growing so fast that, the analysis has become a problem, because it consumes too much time and effort making it an endless task.

Two models will be created based on the historical data collected by the Police Department of UK and Ecuador. The creation of these models will detect patterns that could help them predict and avoid crimes. They will reduce the time that crime analysts spend on finding similar characteristics between different types of crimes.

4. Breakdown of tasks

This project will be broken down into four tasks:

1. Research studies: To start the project it is necessary to research into previous studies in the field of predicting crime with data mining analysis. Because it is important to understand have been done and which could be the author approach.
2. Data Pre-processing: The historical data will be prepared for the next stage using Microsoft Excel to aggregate, reduce dimension, create new features, transform variables, and/or deal with missing values.
3. Creation of the Model: By analyzing the datasets with WEKA Software the models will be created. Different experiments will be applied to achieve the aim.
4. Evaluate the outcomes: On this task the outcomes will be examinated

to validate the models and their performance.

From the beginning of the first task a report will be written describing the development and outcomes of the models.

5. Project deliverables

The models developed in WEKA are the main artifacts of this project, which will be available in the attached disc. These models and their measurements will be explained in the following chapters.

A report including the description of the methodologies and explanation of the analysis results is one of the deliveries of this project.

6. Requirements

Although, the project does not have a direct client, it is important that the models have a high correlation coefficient and a lower relative error so they will be efficient and therefore, their prediction will be more accurate.

7. Legal, ethical, professional, social issues

Although the aims of the project are to detect patterns, and predict future crimes so, police officers can avoid them. It needs to be noted that it can also help criminals to predict where the police is making rounds, to avoid them and still commit the crimes. To avoid this, the models will only be on the attached CD.

There are several constraints on the execution of this project.

- The size of the dataset from the UK is too big to be handled in a common computer; therefore, the analysis will be executed in the supercomputer of the Institute of Cosmology and Gravitation (ICG).
- The amount of time to develop the models for both countries is short, considering the size of the datasets and, the time that takes to analyse

it. To overcome this, a Gantt chart will be created and included in the final report.

- To respect the citizens privacy, there will not be analysed any personal data. However, it might be interesting to include data such as gender and age, in order to analyse the trend of crimes.
- The crime data from Ecuador is not published, though the Defense Minister of this country signed a release form.

8. Facilities and resources

To achieve the aim of this project, the author will collect the datasets from the Police Departments of UK and Ecuador. Before the analysis, the datasets will be pre-processed using Microsoft Excel to improve the performance on the analysis.

After the pre-processing stage the datasets will be analysed with the Data Mining Software WEKA, and because of the amount of historical crime data, the project will also required the use of the super computer from the Institute of Cosmology and Gravitation (ICG), therefore, the software will be installed on it.

To access this supercomputer, an active Internet connection will be required, because the author will connect to it, mostly via SSH through her own laptop and, for graphics display through the NoMachine Software. The University computers can also be used to connect to the supercomputer when it is necessary.

During the project development, the author will hold regular progress meetings with the supervisor to verify its progress and clarify ideas.

9. Project plan

- Background Research
  - Crime Prediction using Data Mining
  - Data Mining Techniques
- Model Creation Process

    - UK's Data
        - Preparation of Data
        - Modelling and Evaluation
    - Ecuador's Data
        - Preparation of Data
        - Modelling and Evaluation
- Writing Chapters
    - Introduction
    - Criminal Data Mining Review
    - Data Mining Crime Requirements
    - Methodology
    - Implementation
    - Experiment Results
- Evaluation and Conclusion

10. Project mode

Registration mode	Full Time
Project mode	Full Time
Planned submission deadline	September 2014

11. Signatures

	Signature:	Date:
Student		
Client		
Project supervisor		

APPENDIX B: Ethical Examination Checklist

PJE40 and PJS40

Ethical Examination

Undergraduate Final Year Projects

University of Portsmouth

School of Computing
Faculty of Technology

This document describes the issues that you need to consider before you start your investigations. This is particularly important where your work may involve other people (human subjects) for the collection of information as part of your project work.

The examination takes the form of a checklist of 12 questions. Each question has come guidance notes.

Consider each question is turn and check the box for Yes or No.

You are then asked to write a short entry explaining the reason for your reply.

For example:

	Yes	No
6. **Are you in a position of authority or influence over any of the human subjects in your study?**	☐	X
Comments: *Although all the human subjects will be staff members at the University of Portsmouth, none of them are in my own department or area and none are subordinate to me in a management structure. Therefore I can see no way that I could have undue influence over them They could take part completely voluntarily.*		

If a grey box is ticked then your project ideas need to be looked at more closely, and you MUST discuss this matter with your project Supervisor.

The final sections deal with Information Sheet(s) and Informed Consent, and you must attach any extra documents concerning these (where relevant) to this Ethical Examination at time of the submission of your Initial Report.

Ethics Information: 12-point Checklist

1. Will the human subjects be exposed to any risks greater than those encountered in their normal lifestyle? *For example: could the study induce psychological stress or anxiety; is more than mild discomfort or pain likely to result from the study; will the study involve prolonged or repetitive activities?* *Investigators have a responsibility to protect human subjects from physical and mental harm during the investigation. The risk of harm must be deemed to be no greater than in their normal lifestyles.* **Comments:** No harmful activities	**Yes** ☐ **No** ☒
2. Will the human subjects be exposed to any non-standard hardware or non-validated instruments? *Human subjects should not be exposed to any risks associated with the use of non-standard equipment: anything other than pen-and-paper, or typical interactions with desktop, laptop PC's, tablet PC's, PDA's or mobile phones are considered non-standard (for example, using a VR room) nor should they be subjected to non-validated instruments e.g. unscrutinised questionnaires.* **Comments:**	**Yes** ☐ **No** ☒
3. Will the human subjects voluntarily give consent? *If the results of an evaluation (for example) are likely to be used beyond the term of the project (for example, software is to be deployed or data is to be published), then signed consent is necessary. A separate consent form should be signed by each human subject. Return of a consent email can constitute written consent if this has been made clear to the human subject.* *Otherwise verbal consent is sufficient and should be explicitly requested in the introductory script (Information Sheet).* **Comments:**	**Yes** ☒ **No** ☐
4. Will any financial, or other, inducements (other than reasonable expenses and compensation for time) be offered to human subjects? *The payment of human subjects must not be used to coerce them against their better judgement, or to induce them to risk harm beyond that which they risk without payment in their normal lifestyle.*	**Yes** ☐ **No** ☒

Comments:	
5. Does the study involve human subjects who are unable to give informed consent (for example: children under 18, people with learning disabilities, unconscious patients)? *Parental consent is required for human subjects under the age of 18. Additional consent is required for human subjects with impairments, and people assessed to be lacking in mental capacity. If consent is gained from a person other than the human subject themselves e.g. a parent, then written consent* ***must*** *be obtained.* **Comments:** No subjects under the age of 18 were required.	**Yes** ☐ **No** ☒
6. Are you in a position of authority or influence over any of your human subjects? *A person in a position of authority or influence over any human subject must not be allowed to pressurize them to take part in, or remain in, any study.* **Comments:**	**Yes** ☐ **No** ☒
7. Are the human subjects being provided with sufficient details of the study at an appropriate level of understanding? *All human subjects should be able to understand the information provided in any documentation and/or verbal information they receive about the experiment or study. They have the right to withdraw at any time during the investigation, and they must be able to contact the investigator after the investigation. They should be given the details of both student and supervisor as part of the debriefing. This information should be in the introductory script (Information Sheet).* **Comments:**	**Yes** ☒ **No** ☐
8. After the study, will human subjects be provided with feedback about their involvement and be able to ask any questions they may have about this involvement? *If the human subjects request further information, the investigator must provide the human subjects with sufficient details to enable them to understand the nature of the investigation and their part in it.*	**Yes** ☒ **No** ☐

Comments:	
9. Will the human subjects be informed of the true aims and objectives of the study? *Withholding information or misleading human subjects is unacceptable if human subjects are likely to object or show unease when debriefed. It must be clear to human subjects if information is being withheld in order to elicit a true response. This should precede any analysis of the data.* **Comments:**	**Yes** [x] **No** [ ]
10. Will the data collected from the human subjects be made available to others (where appropriate and only in relation to this research study), and be stored, in an anonymous form? *All human subject data (hard-copy and soft-copy) should both be stored securely and, if appropriate made available, in an anonymous form. Making human subject data available to a third party may be relevant where a student is taking part in a wider research project eg. for a member of the University staff, in which case anonymity of human subject data must be preserved.* **Comments:** No subject names recorded.	**Yes** [x] **No** [ ]
11. Will the study involve NHS patients, staff, or premises? *If yes, then an application must be made to the appropriate external NHS Local Research Ethics Committee (LREC). For projects other than postgraduate research studies, the length of time for gaining external approval may not fit into a project timescale.* **Comments:** No NHS items needed	**Yes** [ ] **No** [x]
12. Will the study involve the investigator and/or any human subject, in activities that could be considered contentious, morally unacceptable, or illegal? *If yes, then further approval must be sought. For example: a project involving the study of pornography on the web will fall into this category. It is possible that the project may not be allowed to proceed.* **Comments:** No morally unacceptable or illegal activities involved	**Yes** [ ] **No** [x]

Please attach the following:

- **Any Information Sheet(s) or introductory script(s) that the investigator has created for the benefit of the human subjects in the study.** *(See http://www.btinternet.com/~trking for examples of Information Sheets that set out details of a research study for human subjects).*

- **Any documentation that the investigator has created to gather informed consent from the human subjects. This may be an Informed Consent Form, or a form of wording used to get verbal consent.** *(See http://www.btinternet.com/~trking/icf.htm for an example of an Informed Consent Form for research study with human subjects).*

==========

By signing this form, I AGREE to abide by the decisions made in the above points.

If at any time during my project, my answers would change from a white box to a grey box, then I MUST seek re-approval for my project. I understand that if I do not do so, then it is possible that I may FAIL the project component of my course.

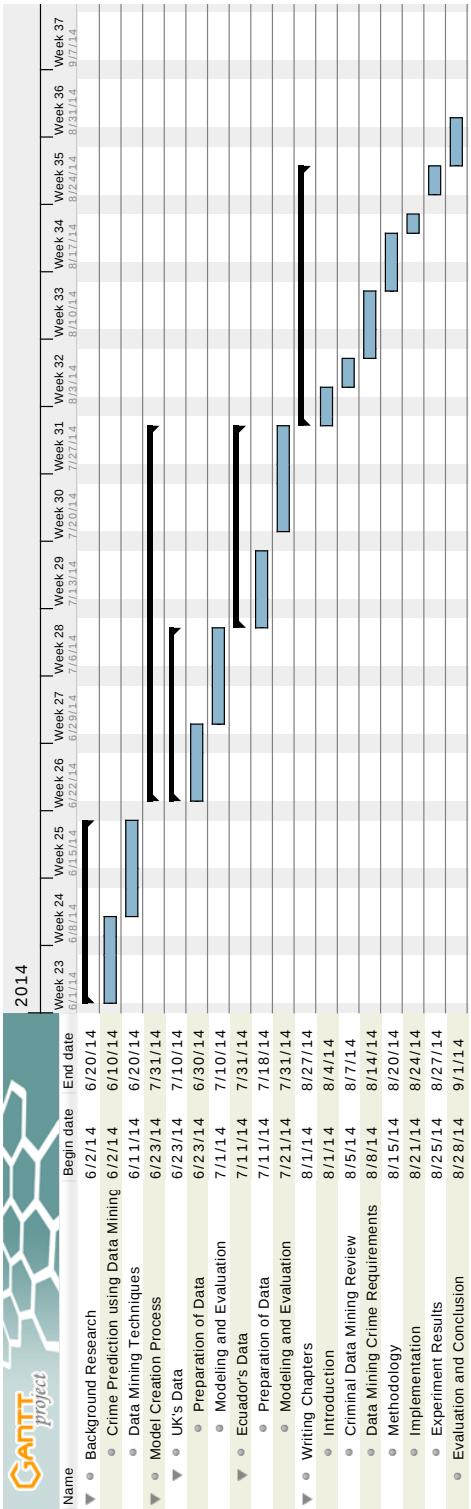
Student name: …Ginger Viviana Saltos Bernal …… Jupiter number: …UP707137……

Student signature: …….. Date ……

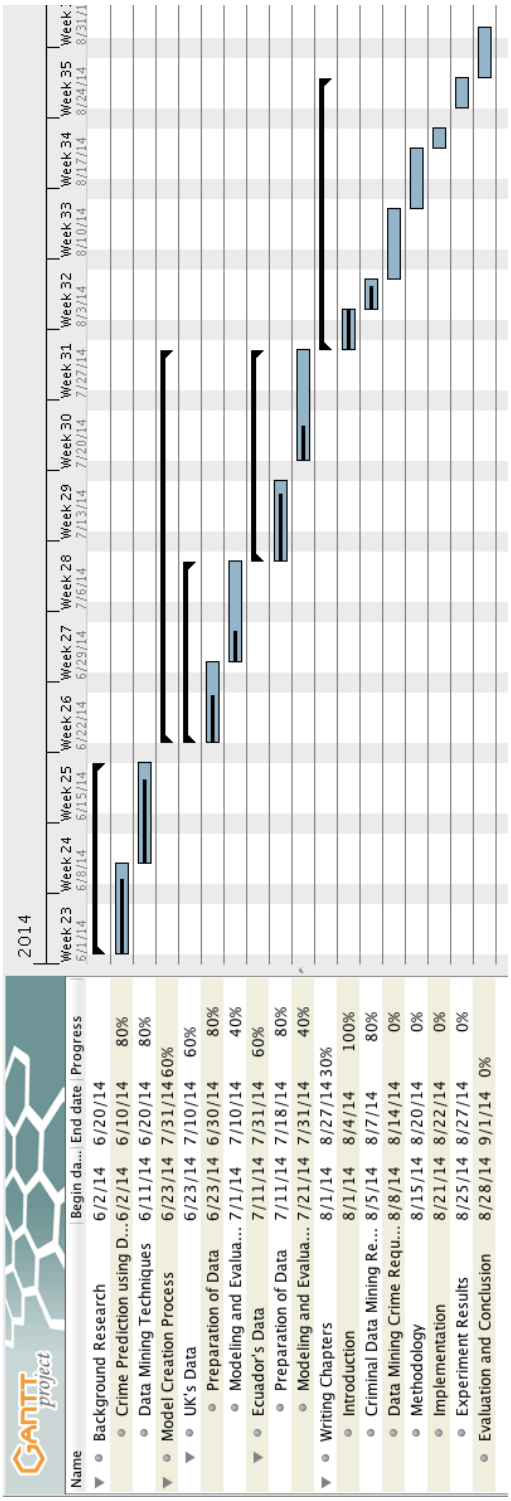
==========

Supervisor signature: ……. Date ……..

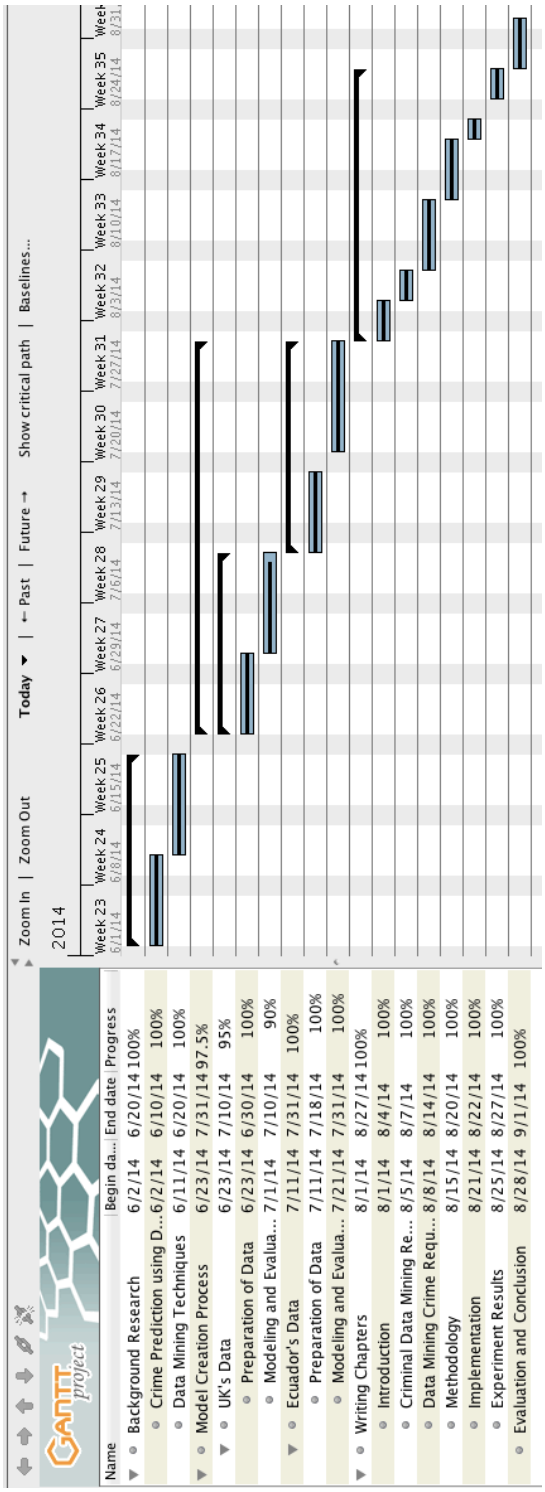
APPENDIX C: Initial Gantt Chart



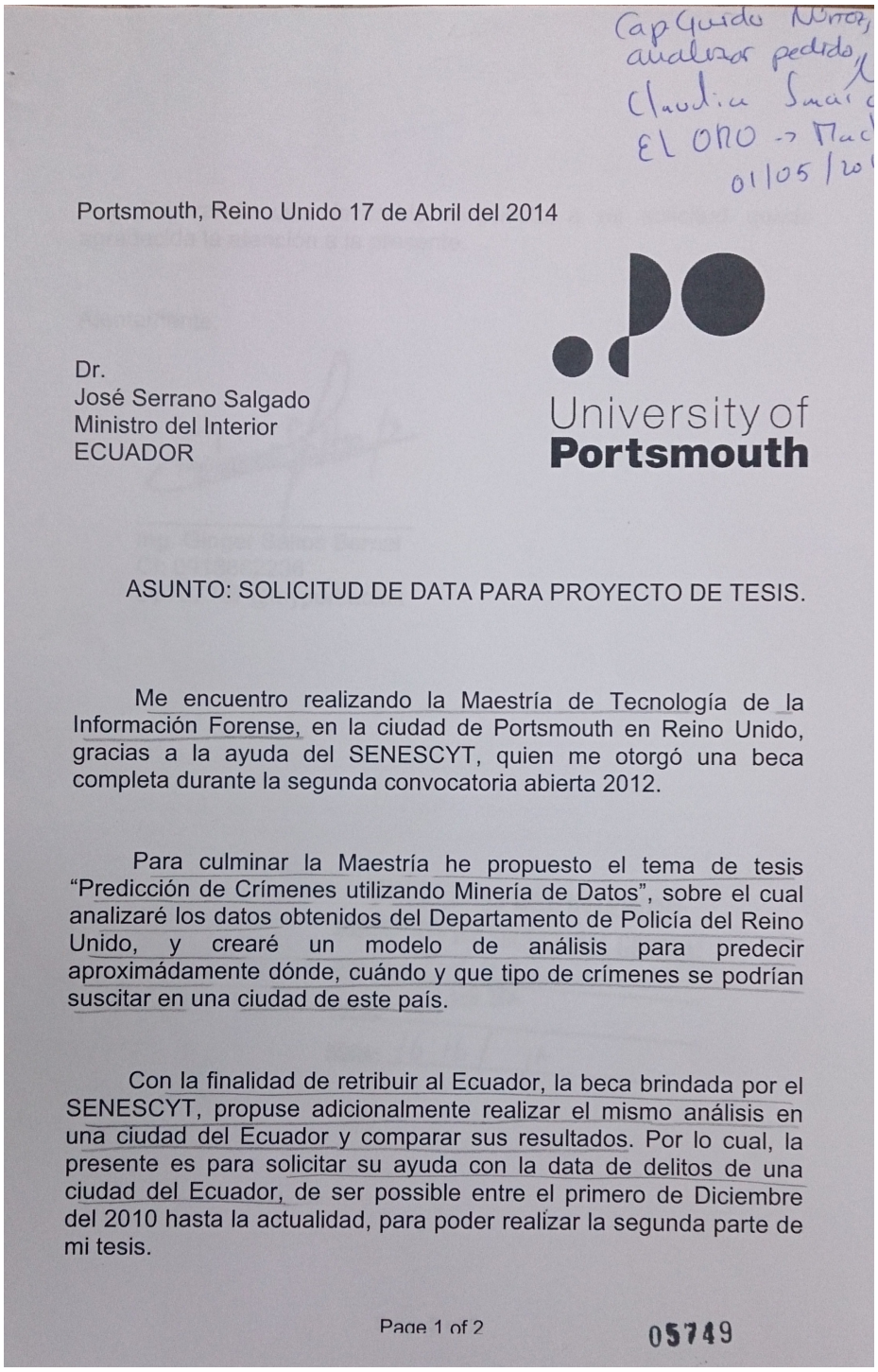
APPENDIX D: Intermedian Gantt Chart



APPENDIX E: Final Gantt Chart



APPENDIX F: Ecuadorean Release Form



Cap Guido Nitor,
analizar pedido
Claudia Sma
EL ORO -> Mac
01/05/20

Portsmouth, Reino Unido 17 de Abril del 2014

Dr.
José Serrano Salgado
Ministro del Interior
ECUADOR

University of
Portsmouth

ASUNTO: SOLICITUD DE DATA PARA PROYECTO DE TESIS.

Me encuentro realizando la Maestría de Tecnología de la Información Forense, en la ciudad de Portsmouth en Reino Unido, gracias a la ayuda del SENESCYT, quien me otorgó una beca completa durante la segunda convocatoria abierta 2012.

Para culminar la Maestría he propuesto el tema de tesis "Predicción de Crímenes utilizando Minería de Datos", sobre el cual analizaré los datos obtenidos del Departamento de Policía del Reino Unido, y crearé un modelo de análisis para predecir aproximádamente dónde, cuándo y que tipo de crímenes se podrían suscitar en una ciudad de este país.

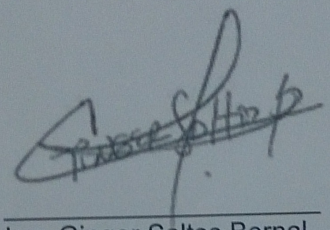
Con la finalidad de retribuir al Ecuador, la beca brindada por el SENESCYT, propuse adicionalmente realizar el mismo análisis en una ciudad del Ecuador y comparar sus resultados. Por lo cual, la presente es para solicitar su ayuda con la data de delitos de una ciudad del Ecuador, de ser possible entre el primero de Diciembre del 2010 hasta la actualidad, para poder realizar la segunda parte de mi tesis.

Page 1 of 2

05749

Esperando una favorable respuesta a mi solicitud quedo agradecida la atención a la presente.

Atentamente,



Ing. Ginger Saltos Bernal
CI: 0918862236
Up707137@myport.ac.uk

MINISTERIO DEL INTERIOR
DIRECCIÓN DE SECRETARIA GENERAL
RECIBIDO: [illegible]
FECHA: 17 ABR 2014
HORA: 16:16