Data Mining Web-Tool Prototype Using Monte Carlo Simulations

Diego Santamaría
Álvaro de Ramón

School of Engineering
Blekinge Institute of Technology
Box 520
SE – 372 25 Ronneby
Sweden
This thesis is submitted to the School of Engineering at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Software Engineering. The thesis is equivalent to 20 weeks of full time studies.

**Contact Information:**

**Author(s):**
Diego Santamaria  
Address: Nuño Rasura 1 1º H, 09400 Aranda de Duero, Burgos, Spain  
E-mail: diego.santamaria.arribas@gmail.com

Alvaro de Ramón Salto  
Address: Avda. Paralela 10 3ºB, 28220 Majadahonda, Madrid, Spain  
E-mail: c0rs4r10@gmail.com

**External advisor(s):**
Lawrence E. Henesey  
Company/Organization name: TTS Port Equipment AB,  
Address: Kämpegatan, 3 SE-411 04 Göteborg, Sweden  
Phone: +46 31 725 79 00

**University advisor(s):**
First name, last name: Lawrence E. Henesey  
Department: Department of Systems and Software Engineering

School of Engineering  
Blekinge Institute of Technology  
Box 520  
SE – 372 25 Ronneby  
Sweden  
Internet : www.bth.se/tek  
Phone : +46 457 38 50 00  
Fax : + 46 457 271 25
ABSTRACT

Facilitating the decision making process using models and patterns is viewed in this thesis to be really helpful. Data mining is one option to accomplish this task. Data mining algorithms can show all the relations within given data, find rules and create behavior patterns. In this thesis seven different types of data mining algorithms are employed.

Monte Carlo is a statistical method that is used in the developed prototype to obtain random data and to simulate different scenarios. Monte Carlo methods are useful for modeling phenomena with significant uncertainty in the inputs.

This thesis presents the steps followed during the development of a web-tool prototype that uses data mining techniques to assist decision-makers of port planning to make better forecasts using generated data from the Monte Carlo simulation.

The prototype generates random port planning forecasts using Monte Carlo simulation. These forecasts are then evaluated with several data mining algorithms. Then decision-makers can evaluate the outcomes of the prototype (rules, decision tress and regressions) to be able to make better decisions.

Keywords: Monte Carlo, Data mining, prototype, simulation.
CONTENTS

ABSTRACT ...........................................................................................................................................I

CONTENTS ...........................................................................................................................................II

1 INTRODUCTION .............................................................................................................................. 1

2 AIM OF THE PROTOTYPE ............................................................................................................... 3
  2.1 INDUSTRIAL BACKGROUND ................................................................................................. 3
  2.2 AIMS – THE PROBLEM ......................................................................................................... 3
  2.3 RELATED WORK .................................................................................................................... 4
  2.4 RESEARCH QUESTIONS AND EXPECTED OUTCOMES ...................................................... 4

3 METHODS ....................................................................................................................................... 6
  3.1 INTRODUCTION ..................................................................................................................... 6
  3.2 DATA MINING ....................................................................................................................... 6
    3.2.1 Introduction .................................................................................................................... 6
    3.2.2 Data mining tasks .......................................................................................................... 6
      3.2.2.1 Classification ........................................................................................................... 7
      3.2.2.2 Estimation ................................................................................................................ 7
      3.2.2.3 Prediction ................................................................................................................ 8
      3.2.2.4 Affinity Grouping or Association Rules .................................................................. 8
      3.2.2.5 Clustering ................................................................................................................ 8
      3.2.2.6 Profiling ................................................................................................................... 9
    3.2.3 Data mining methodology .............................................................................................. 9
    3.2.4 Data mining algorithms ............................................................................................... 10
      3.2.4.1 Inferring rudimentary rules .................................................................................... 11
      3.2.4.2 Statistical modeling ............................................................................................... 11
      3.2.4.3 Divide-and-conquer: Constructing decision trees .................................................. 12
      3.2.4.4 Mining association rules ....................................................................................... 13
      3.2.4.5 Linear models ........................................................................................................ 13
      3.2.4.6 Instance-based learning ........................................................................................ 13
      3.2.4.7 Clustering ............................................................................................................... 14
    3.2.5 Weka .................................................................................................................................. 14
  3.3 MONTE CARLO ......................................................................................................................... 16
    3.3.1 Introduction .................................................................................................................... 16
    3.3.2 Origin of the method ...................................................................................................... 16
    3.3.3 Monte Carlo simulation ............................................................................................... 18
    3.3.4 Monte Carlo simulation using Microsoft Excel ............................................................. 19
      3.3.4.1 Introduction ............................................................................................................ 19
      3.3.4.2 What is Monte Carlo simulation? ......................................................................... 19
      3.3.4.3 RAND function in Excel ..................................................................................... 20
  3.4 MOTIVATION OF THE METHODS ......................................................................................... 20

4 THE PROTOTYPE ............................................................................................................................ 23
  4.1 WHY? CONTRIBUTION? ........................................................................................................ 23
  4.2 SOFTWARE METHOD ............................................................................................................. 23
    4.2.1 Requirements of the prototype ...................................................................................... 25
    4.2.2 Validation, Verification and Testing ............................................................................ 25
  4.3 DECISIONS ............................................................................................................................. 25
  4.4 SYSTEM REQUIREMENTS & TECHNOLOGIES ............................................................... 26
  4.5 WEB MAP ............................................................................................................................... 29
  4.6 QUICK GUIDE TO THE WEB-TOOL .................................................................................... 30
    4.6.1 Login into the system .................................................................................................... 30
    4.6.2 Main user page .............................................................................................................. 31
    4.6.3 New User Forecast ....................................................................................................... 33
    4.6.4 Data Mining Data Selection ........................................................................................ 34
1 INTRODUCTION

Data mining is becoming an important topic nowadays. More and more books can be found, web-pages related to it and also more job offers in newspapers. Currently, many companies have huge database systems with vast amounts of under-exploited data, which could be later formed into “information”. By using the Data Mining technique, this may assist users in locating or identifying patterns than can be useful as “information” in large data repositories that are often not first acknowledged or known by users. Applying Data Mining can assist users to locate information that can be used for making decisions and help to create knowledge.

Data mining has some advantages, such as helping companies to differentiate between the useful and useless data. For example, a company may find hidden predictive information obtained from a large database that is very useful rather than first focusing on information that is “perceived” to be important. This “hidden” data can help decision-makers in many ways.

Typical applications of data mining are usually related to business, financial or marketing problems. Market segmentation, customer practices, fraud detection are areas where data mining is also used. In [22] also some examples about combating terrorism, games and science engineering (bioinformatics, genetics, medicine, education, and electrical power engineering) are found. For example some related papers are [42- 46].

Monte Carlo simulations, explained in 3.3 , are used in the thesis to generate fake data to be able to study them in different scenarios because all the possible value for the variables cannot be tested, due to this prototype is dealing with hundreds of variables. It will take a lot of time to gather all the data in real time from the scenarios by the users. So it used to make an estimation of all the possible scenarios. Once all the information is obtained, it can be compared with the real scenarios and users can make decision or evaluate the accuracy of the fake scenario. Monte Carlo simulations are used in different science areas, for example in molecular research, nuclear area, risk analysis in business, aerodynamic testing, and of course, as it is a mathematical model, to approach the solution of mathematical equations. Examples can be found in [39, 40 and 41].

This thesis presents a web-tool prototype developed to help decision-makers in port planning business environments to reach a better decision. Decisions are mainly based in forecasting, planning, scheduling, etc. The tool was suggested by port experts, TTS Port Equipment and Moffat & Nichol (consulting firm located in London, UK).

Why port planning is important? Due to the global globalization, the distances now seem to be less than before. People are travelling more and all the infrastructures are growing up. For example, the number of twenty-foot equivalent unit containers (TEUs) shipped world-wide was increased from 39 million in 1980 to 356 million in 2004; the growth is projected to be an annual growth 10% till 2020 [15]. This is just an example; a 10% annual growth rate is a really big rate. So port planning is important, since a lot of money is involved on it. If we look to other type of markets, for example the growth for the Chinese economy is estimated in a 6% [37]. And China is one of the countries that are expected to grow more in the next years.

The tool provides the necessary mechanism to help decision-makers to find patterns and models in the gathered data in different scenarios using different data mining algorithms.
In this document in chapter 2 the aim of the prototype will be introduced. What was the industrial background, how was the environment before this prototype and why this thesis was done. In chapter 3 a definition of the methods, data mining and Monte Carlo will be found. Also why this thesis is focused in both methods, why they are used. In chapter 4 is the time of the prototype itself. A brief introduction of the prototype can be found, the software method which was used, the decisions that were taken to accomplish it, a quick guide and how to evaluate the results that can be obtained. In chapter 5 a real simulation is conducted. How to make a Monte Carlo simulation and how can be obtained and evaluated all the results by data mining. And then in the following chapters the discussion, conclusions and future work of this thesis.
2 AIM OF THE PROTOTYPE

2.1 Industrial background

The company that was interested in the web-tool prototype was TTS Port Equipment AB, located in Goteborg Sweden. TTS Port Equipment offers ports as well as ferry and Roll-on Roll-off ship operators cost and time efficient solutions for handling passengers and cargo in the interface between vessel and shore.

![Figure 1: logo of the company TTS](image)

It is a subsidiary of TTS, which is focuses on creating and supplying innovative systems and equipment to the Marine and Offshore industries. With a worldwide workforce of 1000, TTS has over 40 years experience in the marine industry. The group has subsidiaries in Norway, Sweden, Finland, Germany, USA, China, South Korea, Italy, Vietnam, Canada, Singapore and Czech Republic.

2.2 Aims – the problem

The main request from TTS Port Equipment AB was to develop a data mining web-tool prototype to help decision-makers in port planning business environments make forecasts for the next few years. Data mining was required to find patterns and models. Another request was to use the Monte Carlo method to simulate different forecasts and use data mining to evaluate them.

The aim for their participation was to learn how to make better forecasts and thereby maximize their profit, using the patterns, models and the information provided by the prototype. Also the ease of planning would be increased.

Why is this thesis so important? Why is it done? Decision-makers before they were using just an excel file, a spreadsheet. Only with that file and their experience decision-makers had to make the forecasts and estimate future values. It was a really difficult task, because they didn’t have that much information and it was a risky activity. The company decided then to take a step forward and use data mining. They wanted to know some predicted information about the future. Get some rules or decision tress to help decision-makers, providing them with more information. Also they decided that the needed variables to take all the information were too many, so with Monte Carlo simulations different scenarios will be evaluated without the user help. Automatically, just the scenario will be created and then with data mining algorithms the results will be showed. Then decision-makers will also have some random scenarios to know more about the relation of the variables. Always focused on make better forecasts and budget estimations.
2.3 Related work

Data mining is becoming more important, as was said before. Companies are starting to hire employees with data mining knowledge, just taking a look at the newspapers’ advertisements. The companies have started to store amounts of data due to the extremely inexpensive price of the storage devices, for which data mining is required.

A web-tool using data mining techniques that supports different types of data was not found to our knowledge. The prototype that is developed and evaluated in this thesis was based on a specific situation, no tool was found with the same functionality.

Weka, a data mining workbench was used [13]. Weka is an open source collection of machine learning algorithms for data mining tasks. The algorithms can be either applied directly to a specific dataset or from Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. This web-tool prototype is based in Weka workbench using the algorithms provided by Weka to apply them to the data.

Some research papers about data mining in industry were found. For example, in [14] the data mining issues and requirements within a company are introduced. That paper highlights the current important trends of data mining tools and methodologies useful for industry. But, for our knowledge, papers using data mining and Monte Carlo were not found. Tons of papers about data mining were found, as well about Monte Carlo simulation, explained in the introduction, but not about both in the same paper.

2.4 Research questions and expected outcomes

The research questions were:

1. How can data mining be useful as the basis for a decision support tool in port planning?
2. How can we evaluate and analyze data generated via Monte Carlo simulation in testing data mining techniques?

Some other questions can also be, for example, how can a web-based tool be implemented? In section 4.2, in the requirements section all the requirements for this thesis are explained.

Doing a generalization, in this thesis it can be learned how to use Monte Carlo simulations, and also a small overview about data mining, how the results that data mining provides are and how to evaluate them. The key issue is to help decision-makers, so also can be learned how to work thinking in another people, in the final users of the prototype, what could be useful for them and what not.

Talking about the transparency of the thesis, the results from the prototype could be validated by experts in port planning from a U.S. based consulting firm and from experts from Sweden. They will use the tool, creating new forecasting and they will compare with their own experience to check the accuracy of the results and provide some feedback. Maybe one way is to try with old forecasting in the prototype, comparing with the present values.

With the focus in novelty, no relevant research could be found about the main concept of the developed prototype. The research questions were centered about the main functions of the prototype and the possible outcomes, at a really top level. It is a
new idea. It was new that data mining uses Monte Carlo simulations to evaluate port planning. Nothing was found about these two topics together. It is just new, but some examples about other different areas were found in [39-46].

Is this project relevant? The results from the prototype were expected to be presented as either tables or graphs. Decision makers in port planning companies are waiting for these results and information. This thesis was done to help them to make better decisions, providing with much more information than before and some predictions.
3 METHODS

3.1 Introduction

After the first view of what is the thesis about, the two main concepts of it, data mining and Monte Carlo will be introduced and defined. The user has to know what these topics are and why they are so important.

An approach of why data mining and Monte Carlo should be used will be introduced in section 3.4, as well as why they should be used together.

3.2 Data mining

3.2.1 Introduction

Data mining is the analysis of observational data sets to find unsuspected or unexpected relationships and to summarize them in new ways that are both understandable and useful to the data owner [9]. Observational data is data was collected for some purpose other than just the data mining analysis. The relationships derived through a data mining activity are often called models or patterns. Some examples can be linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series.

In other words, data mining prepares, surveys, and explores the data to find out all the hidden information in it. Usually not only the data is the most valuable aspect, but the information hidden in its relationships and dependences.

3.2.2 Data mining tasks

Many problems of intellectual, business or economic interest can be expressed in terms of these six tasks, as explained in [11].

- Classification
- Estimation
- Prediction
- Affinity grouping
- Clustering
- Description and profiling

The first three are examples of directed data mining, where the aim is to find the value of a target variable. Affinity grouping and clustering are undirected tasks where the goal is to reveal structure in data without considering a particular target variable. Profiling may be either directed or undirected [11].
3.2.2.1 Classification

Classification is one of the most common data mining tasks. The world is always classifying, categorizing and grading, in order to understand and communicate. Classification is the activity of examining the features of a presented object and assigning it to one of a predefined set of classes. Usually the objects to be classified can be represented by records in a database table or file, so the act of classification is just adding a new column with a class code. The main goal is to build a model of some kind that can be applied to unclassified data in order to classify it.

Some examples of classification are for example [11]:

- Classifying credit applicants as low, medium, or high risk.
- Choosing content to be displayed on a Web page.
- Determining which phone numbers correspond to fax machines.

In these examples, there are a limited number of classes and is expected to be able to assign any record into one or another of them.

Decision trees and nearest neighbor are techniques well suited to classification. Neural networks and link are also useful for classification under certain circumstances [11].

3.2.2.2 Estimation

Classification deals with discrete outcomes: yes or no; or different types of diseases, etc. Estimation deals with continuous values. Given some input data, estimation will give a value for some unknown continuous variable such as income, height, or credit card balance.

In practice, estimation is often used to perform a classification task. If for example a credit card company wants to sell advertising space in its billing envelopes to a ski boot manufacturer, a classification model with all of its cardholders into skier or nonskier classes can be used. Another approach is to build, which assigns a “propensity to ski score” to each cardholder. This should be a value from 0 to 1 indicating the probability that the card holder is a skier [11]. After that the classification task establishes a threshold score. Anyone with a score greater than or equal to the threshold is classes as a skier, and anyone with a lower score, is not considered to be a skier.

With the estimation approach, the individual records can be ordered by rank according to the estimate, which is a great advantage. In the example of the ski boot manufacturer, the classification approach can be used to find skiers. Then with this poll of skiers, using the tendency to ski score, the company can focus just in the most likely candidates from all the skiers found with the classification approach. Some examples of estimation tasks are [11]:

- Estimating the number of children in a family
- Estimating a family’s total household income
- Estimating the lifetime value of a customer
- Estimating the probability that someone will respond to a balance transfer solicitation.
Regression models and neural networks are well suited to estimation tasks. Survival is well suited to estimation tasks where the goal is to estimate the time to an event.

### 3.2.2.3 Prediction

Prediction is the same as classification or estimation. It only differs in that the records are classified according to some predicted future behavior or estimated future value. In prediction there is only one way to check the accuracy of the classification: wait and see. The primary reason for treating it separately from classification and estimation is because prediction has additional issues about the temporal relationship of the input variables or predictors to the target variable.

Any technique used for classification and estimation can be adapted for be used in prediction, using training examples where the variable to be predicted is already known with historical data of those examples. With the historical data it is possible to build a model that explains the current observed behavior. Once the model is applied to current inputs, the result is a prediction of future behavior.

Examples of prediction tasks [11]:

- Predicting which customers will leave within the next 12 months
- Predicting which telephone subscribers will order a new service such as voice mail.

### 3.2.2.4 Affinity Grouping or Association Rules

The objective of affinity grouping is to decide which things go together. An example is a shopping cart at the supermarket, which is the task at the heart of market basket analysis [11]. Retail chains can use affinity grouping, then for example the items often purchased together will be seen together on store shelves or in a catalog.

Affinity grouping can also be used to find cross-selling opportunities. Affinity grouping is one simple approach to generating rules from data. If for example, two items, cat food and kitty litter, occur together frequently enough, this association rules can be generated [11]:

- People who buy cat food also buy kitty litter with probability P1.
- People who buy kitty litter also buy cat food with probability P2.

### 3.2.2.5 Clustering

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters [11]. The difference between clustering and classification is that clustering does not depend on predefined classes. In classification, each record is assigned to a predefined class.

In clustering, there are no predefined classes and no examples. The records are grouped together based on self-similarity. It is decision of the user to determine what meaning, if any, to attach to the resulting clusters. For example clusters of symptoms
could indicate different diseases. Clusters of customer attributes might be different market segments.

Clustering is often used before some other form of data mining or modeling. For example, clustering could be the first step in a market segmentation effort. Instead of trying to find a one-size-fits-all rule for “what kind of promotion do customers respond to best,” maybe a first division of the customer base into clusters or people with similar buying habits could be done and then ask the kind of promotion which works best for each cluster.

3.2.2.6 Profiling

Sometimes the purpose of data mining is simply to describe what is going on in a complicated database, just to increase the understanding of the people, products, or processes that produced the data. A good description of a behavior will suggest an explanation for it as well. The sentence “women support Democrats in greater numbers than do men” can trigger large amounts of interest and further study on the part of journalists, sociologists, economists, and political scientists, not to mention candidates for public office [11]

Decision trees are a powerful tool for profiling customers (or anything else) with respect to a particular target or outcome. Association rules and clustering can also be used to build profiles.

3.2.3 Data mining methodology

A typical data mining process is composed of the following steps, as mentioned in [22]:

- **Selection of the target data**, the dependent variables, as well as the objective variables. The sampler of the available registers will be probably required also.

- **Analysis of the data properties**, especially histograms, scattergrams, atypical value presence and lack of data. In this step preprocessed data will be obtained.

- **Transformation of the input data set**. It will be accomplished using different techniques, according to the previous analysis. The objective is the preparation of the data for applying the most suitable data mining technique in each case.

- **Select and apply data mining**. After a review of the data and data mining tasks the model has to be built. It is important to choose the most suitable model for our data: classification, estimation, prediction, ruler, clustering or profiling.

- **Evaluate the results** and comparing them with previous data or stored data to validate the model.
If the final model does not pass the evaluation, the process could be started from the beginning, or if the expert considers appropriate, maybe from another step, not the first one. This feedback would be done as long as the evaluation does not reach the expected values or the model is not correct.

Once a statistical or data mining model has been produced by the steps above, the next phase begins of deploying the model in operational systems. Deployment usually consists of three different activities. First, data is given a weight using the statistical or data mining model produced on a periodic basis: daily, weekly, monthly, or perhaps in real time. Second, these weights are deployed into operational systems and are also used as the basis for various reports. Third, on a periodic basis, say monthly, a new model is built and compared to the existing model. If required, the old model is replaced by the new model.

### 3.2.4 Data mining algorithms

As is suggested in [12], one of the best lessons is that simple ideas often work very well, and the adoption of a “simplicity-first” methodology when analyzing practical datasets should be applied. Different types of datasets can be found to use with data mining algorithms, for example:

- A single attribute does all the work and the other may be independently and equally to the outcome.
- There may be a few independent rules that govern the assignment of instances of different classes.
- Class values may not be provided: the learning is unsupervised.
- A simple logical structure, involving just a few attributes that can be captured by a decision tree.

In the infinite variety of possible datasets there are many different kinds of structure that can occur, and a data mining tool that is looking for one class of structure
may completely miss regularities of a different kind. All the examples given before lead to a different machine learning method well suited to it.

In the next paragraphs a small overview about data mining algorithms will be showed, following [10] and [12].

3.2.4.1 Inferring rudimentary rules

This is one of the easiest algorithms; it will find very simple classification rules from a set of instances. It is called 1R for 1-rule. It generates a one-level decision tree. 1R is a simple, cheap method that comes up with good rules for characterizing the data. Despite its simplicity, simple rules frequently achieve surprisingly high accuracy. It is always a good plan to try the simplest things first [12].

The idea of the method is to create rules that test a single attribute and branch accordingly. Each branch fits to a different value of the attribute. Obviously, the best classification to give each branch is to use the class that occurs most often in the training data. Then the error rate is easily determined, just counting the errors occurring in the training data. Figure 3 shows the algorithm 1R in pseudo code [12].

3.2.4.2 Statistical modeling

The 1R method uses a single attribute as the basis for its decision and then it chooses the one that works best. This statistical modeling is also a simple technique, but it uses all attributes and allows them to contribute to the decisions that are equally important and independent of one another, given the class [12].

This method is called Naïve Bayes, based on Baye´s rule and “naively” assumes independence. Naïve Bayes works very well when tested on actual datasets, particularly when it is combined with some attribute selection procedures that eliminate redundant, and therefore no independent, attributes [12].

Bayes’ rule says that if you have a hypothesis $H$ and evidence $E$ that are relevant on that hypothesis, then

$$\Pr[H|E] = \frac{\Pr[E|H] \Pr[H]}{\Pr[E]}.$$ 

The notation that $\Pr[A]$ denotes the probability of an event $A$, and $\Pr[A|B]$ denotes the probability of $A$ conditional on another event $B$ [12].
### 3.2.4.3 Divide-and-conquer: Constructing decision trees

Following [11], a decision tree is a structure that can be used to split a large collection of records into successively smaller sets of records by applying a set of simple decision rules. In each successive division, the resulting sets members become more and more similar to each other. The familiar division invented by Carl Linnaeus in the 1730s is a good example of this [11]. Also an example can be found in Figure 4.

A decision tree model consists of a set of rules for dividing a large heterogeneous data into smaller, more homogeneous groups regarding to a particular target variable. A decision tree may be grown automatically by applying decision tree algorithms to a model set comprises of preclassified data. The target variable is usually categorical and the decision tree model is used to [11]:

- Calculate the probability that a given record belongs to each of the categories
- Classify the record by assigning it to the most suitable class.
- Estimate the value of a continuous variable, even if there are other more suitable techniques to accomplish that task.

Figure 4: Example of a binary decision tree classifies catalog recipients as likely or unlikely to place an order. [11]
3.2.4.4 Mining association rules

One request of associating rules is the clarity and utility of the results which are in the form of rules about groups of products. Association has an intuitive request because it expresses how tangible products and services group together. An example of rule like, “if a customer purchases three-way calling, then that customer will also purchase call waiting” is clear, simple and full of significance. Even better, it suggests a specific course of action, such as putting them into a single package [11].

While association rules are easy to understand, they are not always useful. These are examples of real rules generated from real data [11].

- Wal-Mart customers who purchase Barbie dolls have a 60 percent likelihood of also purchasing one of three types of candy bars. Forbes on September 8, 1997 [11]
- Customers who purchase maintenance agreements are very likely to purchase large appliances.
- When a new hardware store opens, one of the most commonly sold items is toilet bowl cleaners

These three examples bring into relief the three common types of rules produced by association rules: the actionable, the trivial, and the inexplicable.

3.2.4.5 Linear models

The methods focused in decision trees and rules work most with nominal attributes. They could be extended to numeric attributes incorporation numeric-value test into the decision tree or by prediscretizing numeric attributes into nominal ones. However, data mining has methods to work naturally with numeric attributes.

Linear regression: when the outcome or class is numeric and all the attributes are numeric, linear regression is a technique to be considered [12]. It is an essential method in statistics. It expresses the class or outcome as a linear combination of the attribute, with predetermined weights, which are calculated from the training data.

3.2.4.6 Instance-based learning

In instance-based learning the training examples are stored word by word, and it uses a distance function to determine which member of the training is the closest to an unknown test instance. When this instance has been located, the prediction for the test instance of the nearest one is done. Only one problem will remain, defining the distance function which is not very difficult to define if the attributes are numeric [10]. Although there are other possible choices, most learners use Euclidean distance.
3.2.4.7 Clustering

Clustering techniques can be applied when there is no class to be predicted but rather when the instances can be divided into natural groups. Clustering requires different techniques to the classification and association learning methods we have considered so far.

There are different ways in which the result of clustering can be expressed. The groups that are identified may be exclusive so that any instance belongs in only one group. Or they may be probabilistic, whereby an instance belongs to each group with a certain probability. The choice among these possibilities should be found using the provided clustering mechanisms. However, these mechanisms are rarely known, and clusters are still something to discover and usually the choice is dictated by the clustering tools available.

The classic clustering technique is called \textit{k-means} and this clustering method is simple and effective [12]. First it needs to know how many clusters are being sought: this is the parameter $k$. Then $k$ points are chosen at random as cluster centers. All instances are assigned to their closest cluster according to Euclidean distance metric. Next the mean of the instances in each cluster is calculated, this is the “means” part. These means will be the new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers, until the same points are assigned to each cluster in consecutive rounds [12].

3.2.5 Weka

Weka is found in [12] and [13]. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools, which includes a collection of tools for completing many data mining tasks. It is designed so that existing methods on new datasets can be tried in flexible ways. It is open-source software issued under the GNU General Public License.

It is considered relevant because it is used in the prototype. This workbench is required to implement and to use the data mining algorithms with the data generated and obtained from the prototype. Also the outcomes of the prototype follow the Weka format. As the prototype was developed in PHP, a workbench with command line access was required. Weka supports this access and also requires the transformation of the input files into a specific format, ARFF [23]. A screenshot of Weka is showed in Figure 6.

Weka is used in this prototype from command line, to be able to obtain the results and show them online. But Weka also has implemented a graphical user interface (GUI) where more graphs and data can be evaluated. Weka needs the input files in a special format, called arff [11]. An example of this file can be found in Figure 5, it is quite different to an excel file.
To use Weka offline, the user has to install it and then use the GUI, this prototype provides the ARFF file of the Excel file that the user uploaded and of the Monte Carlo simulation, in Figure 20. Using this file into Weka, trees can be showed and more data mining algorithms can be applied to the input data.
3.3 Monte Carlo

This method is a deterministic method or statistic number uses to approach complex mathematical expressions hard to evaluate.

3.3.1 Introduction

The Monte Carlo’s casino gave the name to the method due to be the capital of the gambling games. The name and the systematic development of Monte Carlos’s methods are from 1944 and the method is improving because of the development of the computers [16].

The use of Monte Carlo’s methods as an investigation tool comes from the development of the atomic bomb during the Second World War. That work was about simulating probabilistic hydrodynamic problems that are about random behaviors.

At the beginning of all of these investigations, John Von Neumann and Stanislaw Ulam improved this Russian roulette and the methods of task divisions [16]. However, the systematic development of all these ideas had to wait till 1948 with the work of Harris and Herman Kahn [16].

Approximately that same year Enrico Fermi, Metropolis and Ulam obtained good results using the Schrödinger’s equation to capture neutrons talking in a nuclear level [16].

The Monte Carlos’ method gives approximated solutions to a huge variety of mathematical problems leaving the possibility of experimenting with sampling pseudo-random numbers for the computer.

It is possible to apply this to every stochastic and deterministic problem. Apart from the numerical methods that are based in evaluating n points in an m-dimensioned space to give an approximated solution, the Monte Carlos’ method has an absolute error of the estimation that decreases following the proportion [16]:

\[
\frac{1}{\sqrt{N}}
\]

According to the theorem of central limit that shows, in general conditions, the distribution of the sum of random variables trends to a normal distribution.

3.3.2 Origin of the method

The invention of Monte Carlo’s method is given to Stan Ulan and John von Neumann. Ulam was playing a card game during the time he was ill in 1946 [16]. He realized that it is much simpler to play having a general idea of the solitaire testing multiple times with the cards and counting the proportions of the results instead of computing all the possibilities of combination. After that he decided to apply his work about diffusion of neutrons, because try to solve some equations like integer-differential.

The main idea was to try by mental thoughts, so thinking in thousand of possibilities and in each stage determine by chance, using distributed random numbers.
as its own possibilities, that it would occur and summarize all the possibilities and have an idea of the physical process.

They could use computers that started to be available to do the numerical proofs and in fact avoid the physical experiment. Ulam shared that information with von Neumann, after that von Neumann got exited and soon he started to develop the new possibilities in a systematic method. [1]
3.3.3 Monte Carlo simulation

Taking a look to the diagram below it is possible to understand what is the simulation, how the simulation works.

Figure 7: Conceptual map of Monte Carlo simulations. [16]
3.3.4 Monte Carlo simulation using Microsoft Excel

3.3.4.1 Introduction

To understand how the Monte Carlo simulation is using Excel, it is important to talk about the different areas of which some have been explained before, like the Monte Carlo method and the simulation. As it is shown in the diagram below it is possible to use the simulation using both discrete and continuous variables, all the data used in the simulation is discrete. It will be explaining further along which random function is used and how the variables are generating using Excel.

Figure 8: Content diagram of Monte Carlo simulation using Excel [38]

The Monte Carlo (MC) simulation is, as Figure 8 shows, a technique that combines statistical concepts (random sampling) with the capacity that computer has to generate pseudo-random numbers and automates calculus.

Since its ascendancy, Monte Carlo simulations have been implemented to an infinite scope like an alternative to the reliable mathematical methods or included as the only medium of esteeming that solves complex problems. Nowadays it is possible to find models that use Monte Carlo simulation in computing field, business-related, economic, industrial and even social fields [2, 3]. In other words, Monte Carlo simulation is manifested in all ambits that random behavior is really important.

There are many authors that have wagered to use spreadsheets to simulate using MC [4, 5, 6]. The power of these spreadsheets dwells in their universality, user-friendliness, the ability to recalculate values and specially the possibilities that gives talking about “What-if analysis”. Moreover the latest versions of Excel incorporate its own programming language; Visual Basic for Applications (VBA). Within it is possible make real applications of simulation bound for final user. In the market, there is a lot of Add-Ins for Excel designed to simulate MC. The most popular are @Risk, Crystal Ball, Insight.xla, SimTools.xla, etc [18, 19, 30].

3.3.4.2 What is Monte Carlo simulation?

This simulation is a quantitative technique that uses statistics and uses computers to imitate the behavior of non-dynamic real systems using mathematical models (all in all, the processing of systems that its status changes over the course of time it is better
to use discrete events or simulate continuous systems). It is necessary to understand mathematical model like a numerical representation performed according to the abstraction of a phenomenon in real life.

The key of Monte Carlo simulation is to create a mathematical model of the system, process or desired activity to analyze, identifying the inputs of the model such that it is possible to determine the global behavior of the system with random values. After the identification of the inputs or random variables it is possible to carry out the experiment that consists in two steps:

1. Generate with computer random samplings (concrete values) for those inputs.
2. Analyze the behavior of the system using the values calculated before.

### 3.3.4.3 RAND function in Excel

Spreadsheets and Excel (like any other standard programming language) are able to generate pseudo-random numbers from a uniform distribution between 0 and 1. That type of numbers is the basic element to develop any simulation using a computer.

Using Excel it is possible to get a pseudo-random number by means of RAND function. These numbers have two properties:

1. Each time that the function is used; any number between 0 and 1 has the same probability to be generated (that is why it is called uniform distribution).
2. The different numbers generated are statistically independent with respect to one another (in other words, the value generated in each instant does not depend on numbers previously calculated.

The RAND function is volatile which means that each time that there is any change or modification in any input of the model, all cells that contains the function will be recalculated.

### 3.4 Motivation of the methods

A deep research was done to try to find some related work with Monte Carlo simulations and data mining. No articles were found, at least to our knowledge. But some ideas are showed in this section, like why they are useful or why they should be used.

First of all, why **data mining** should be used? In [31] some characteristics are exposed. Data mining analyzes customer’s data to build a predictive model that can tell, for example, which customers are at grave risk of leaving a contract and then make a direct marketing campaign to save these valuable customers.

In [32] a deeper explanation is found. Data mining technology can generate new opportunities, given databases with sufficient size and quality, providing these capabilities:
- **Automated prediction of trends and behaviors.** Data mining can automates the process of finding predictive information in large databases. For example in marketing, data mining uses past data of promotional mailings to maximize return on investment in future mailings.

- **Automated discovery of previously unknown patterns.** Data mining tools can explore databases and identify hidden patterns. An example could be detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms. When data mining tools are implemented on high performance parallel processing systems, analysis of massive databases can be done in minutes. Larger databases, in turn, yield improved predictions.

Changing the focus to **Monte Carlo simulations**, in [33] can be found an interesting article about the use of Monte Carlo. Forecasting the future, for any kind of project, can be a complex task. Usually a number of potential risks that may occur can dramatically change all the forecast. When producing a forecast some assumptions about the future have to be made. The Monte Carlo simulation method is a very valuable tool for planning project schedules and developing budget estimates.

When a good estimation forecast or decision with significant uncertainly has to be made, Monte Carlo has to be considered. Dr. Sam Savage, an authority on simulation says “Many people, when faced with an uncertainty succumb to the temptation of replacing the uncertain number in question with a single average value. I call this the flaw of averages, and it is a fallacy as fundamental as the belief that the earth is flat.” Most business activities, plans and processes are too complex for being analyzed analytically. When the situation involves uncertainly in many dimensions, Monte Carlo can be found surprisingly effective [35].

In business is more important, because it involves risks, which is the loss of money, reputation or even stakeholders confidence. So big questions are highlighted: for example, if a new product is launched, what are its chances of failure? Is it possible to calculate its probability of success?

Some related work was found, for example in Rubinstein [34], where he proposed that Monte Carlo simulation is appropriate when it is impossible or too expensive to obtain data, the system is too complex, the analytical solutions is difficult to obtain and when it is impossible or too expensive to validate the experiment with a mathematical model.

How Monte Carlo help in the job or career can was found in [35]. If the success depends on making good forecasts or managing activities with some uncertainty, Monte Carlo can be really useful. As Dr. Sam Savage warns, “Plans based on average assumptions will be wrong on average.” How can that be possible?

- **Go Beyond the Limits of 'What If' Analysis.** Exploring thousands of combinations for the 'what-if' factors and analyzing all the range of possible outcomes, much more accurate results can be obtained, with only a little extra work.

- **Know What Factors Really Matter.** Some tools enable a quick identification of the high-impact factors in a model, using sensitivity analysis across thousands of Monte Carlo attempts. Hours can be needed to identify these factors in the ordinary ‘what if’ analysis.

- **Give Yourself a Competitive Advantage.** In deal negotiations or just a simple market competition having a realistic idea of the probability of the
outcomes enabling to strike a better bargain, choosing the price that yields the most profit or benefit.

- **Be Better Prepared for Executive Decisions.** The higher in an organization, the more deals with uncertainty.

Some specific reasons for using these methods in this thesis are:

- Monte Carlo is not human. Is not afraid of increase some values. Maybe an increase of 5% in a value is really risky for a decision maker, but not for Monte Carlo. And maybe increasing this value really interesting and unexpected results can be obtained.

- In the case of the thesis, almost eight hundred variables are evaluated and simulated. The time to obtain all these variables in a forecast of 5 years it is around 4.000 values. The time to get them could be really long. Using Monte Carlo simulation, this step can be summarized and simulated.

- Is not possible to test all the different possible scenarios. Using Monte Carlo is possible to make comparisons between the simulated scenario and the real scenarios. Users can upload personal scenarios and then use some Monte Carlo scenarios to compare the influence of the variables and also the sensitivity of them.

- Data mining is needed to find these hidden patterns. The user input file contains amounts of data, as well as the Monte Carlo simulation. Data mining algorithms will provide with different results to find the rules and patterns of the data.

- These simulated scenarios can be exported to a real case. Using these values, users can use this previous knowledge to execute their own forecasting. This adds up to make better decisions.
4 THE PROTOTYPE

Web-based tool that uses Monte Carlo simulation and data mining algorithms that pretends to facilitate the expert’s decisions and so, by use of models and patterns the experts will be able to predict which scenario is more desirable.

The tool allows the opportunity to the user upload a file that contains his own forecast with the gathered data in a scenario or simulates after three thousand iterations and give the report with best case, both cases the user can use data mining algorithm to study the results of the forecasts.

In the next paragraphs this prototype is explained in depth, with for example the reasons of developing it, the software model followed an overview of the prototype, how to install it or how to evaluate the possible results.

4.1 Why? Contribution?

One of the main objectives of this thesis is to help decision-makers; in this case, in port planning. The task to make the forecasting for five years was too complicated and took a lot of time: so many variables, all of them in an excel file and then to make decision they only had two or three different graphs. The thesis will help them providing with data mining results and more statistical information. And also simulations of different forecastings can be obtained. Then the needed time to create the file with all the variables disappears and more evaluations or simulations can be done and given to the user. Then users or decision-makers can be able to use more results, statistical information and data mining results to make a better decision and make a better forecasting having some knowledge about how the future could be.

4.2 Software method

![Diagram of the prototype method](image)

Figure 9: Example of the prototype method [36]

For this thesis, the prototype method was used. It is based in three different prototypes, three different iterations of the method.
In the first iteration, in Figure 10, the main characteristics of the prototype were introduced. The first idea was to get all the variables from a web form, but it added too much complexity for both user and prototype.

![First iteration of the prototype](image)

**Figure 10: First iteration of the prototype**

Different decisions were taken in the second iteration of the prototype, Figure 11, one was to change the way to get the information, using an excel file, which is user friendly and more easily to use. Also the first ideas of data mining were introduced and also the statistical reports were included. Some functionality was still not completed, but the main areas were clear and developed.

Then for the third iteration of the prototype, which can be observed in 4.6, Quick Guide of the web-tool, the rest of the functionality. Instead of only one data mining algorithm, seven were implemented. The statistical reports were refined, as well as the interface.

![Second iteration of the prototype](image)

**Figure 11: Second iteration of the prototype**
4.2.1  Requirements of the prototype

The requirements for this project were just a few; it was a really open design.

- It has to be a web tool. One of the main objectives is to be able to create
  the forecasts not only in one specific computer. Just open a browser and
  do it.
- It will implement data mining techniques. Developers will decide which
  algorithms and how many of them.
- The prototype will allow users to send their own forecasts and evaluate
  them in the system, obtaining data mining conclusions.
- The prototype will create simulated forecasts using Monte Carlo
  simulations and then data mining results have to be obtained from those
  simulations.
- The inputs are given. The user provided us with a list of the variables. It
  cannot be changed.

4.2.2  Validation, Verification and Testing

For the validation and verification first the experience of the developers is used.
As the experience in port planning of the developers is not really extended, also some
examiners and industry experts will examine the tool to give feedback.

Then the tool should be implemented in the company or started to be used. In that
moment, when historical data is available from different companies, specific Monte
Carlo simulations can be done. A better and more specific simulation for each user
could be done.

4.3  Decisions

In this prototype some key decisions were taken:

- PHP and MySQL are used in this prototype. PHP is a language to produce
dynamic web pages. Is needed in the prototype to deal with databases and
to create the forecasts. PHP was found maybe the easiest language to learn
and to implement. Due to that reason, it was the chosen language for the
prototype.
- The statistical results are available in pdf format. Pdf is one of the most
common file formats. Everybody has a pdf reader and can also be stored
easily. Users can store the results in an easy way.
- For the Monte Carlo simulations, 3000 simulations are done. Tests were
made with fewer simulations, and with this number of simulations seemed
to be more accurate, also thinking the time it needs to be executed. It is the
best accurate/time proportion.
- The information is taken from a excel file. Excel files are a really easy
way to get data. First the stakeholders have familiarity with these kinds of
files, and the file has the same format as the file they were using before this prototype. So users will not change their way of make the forecasts.

- Seven different data mining algorithms are used. After some tests of the results obtained with all the data mining algorithms provided by Weka, seven algorithms were found to be really useful. It was seen useful to give results from different algorithms, because maybe one algorithm cannot find patterns in the data, but maybe it is possible to find them with another one.

- Some of the functionalities of the prototype, creating the pdf files and simulations were realized with vba in excel files. Excel is a really powerful application, and it was less difficult to implement those features inside the excel files and just launch the functions inside of the files when the prototype opens the files.

### 4.4 System requirements & technologies

This section of the report reviews the system requirements, installation of the applications and the technology used for the development of the prototype.

**System requirements**

1. Windows 2000/XP/Vista (home or professional) as operating system.
2. XAMPP 1.6.7 [24] all-in-one solution that provides:
   i. Apache HTTPD 2.2.9 + Openssl 0.9.8h
   ii. MySQL 5.0.51b
   iii. PHP 5.2.6
   iv. PHP 4.4.8
   v. phpMyAdmin 2.11.7
   vi. FileZilla FTP Server 0.9.25
   vii. Mercury Mail Transport System 4.52
3. Organize folders and files showed in the Figure 12.
4. Microsoft Office 97 or later (recommended 2003). It is necessary to install the application inside the registered folder otherwise it is no possible to execute Excel and then impossible to run the simulator.
5. Ghostscript [25]
6. Bullzip PDF Printer 5 [26]. The PDF printer should be configure to do not prompt any message box like save as, change preferences, default directory to print the files and avoid any kind of error messages.
7. Java Runtime Environment (JRE) version 1.6 [28]
8. Weka 3.4.13 [27] (the workbench to be able to apply data mining algorithms)
It is necessary to install the application inside the folder \textit{dmOperations} to guarantee a proper work of the tool

\textit{Installation}

After installing all the applications mentioned below it is necessary to copy some files in specific folders to conclude the installation process:

1. XLS file \textit{userFile\_MC\_v5\_releas2.xls} into the folder \textit{montecarlo}, that file is going to be the engine of the simulator
2. XLS files \textit{dummy.xls} and \textit{printer.xls}, into the folder that previously was installed Microsoft Office Excel
3. PHP files \textit{newForecastConfirm.php} and \textit{waitForFile.php} into the same folder than Microsoft Office Excel

\begin{figure}
\centering
\includegraphics[width=\textwidth]{files_structure.png}
\caption{Files structure of the prototype}
\end{figure}
Different technologies have been used to develop this prototype, listed below:

1. HTML-PHP that was used to get access to the database executes different commands like open XLS files, generate HTML files, execute Weka using the command line with the different algorithms of data mining, copy files to another directories or rename files.
2. VBA used to program macros in the XLS files like manage the files, print reports, analyze data and be the engine of the Monte Carlo simulation.
3. SQL to implement the database.
4.5 Web map

In Figure 13 the web map of the tool is showed. Also the php pages from where the arff, pdf, excel and txt files are obtained. In this map all the relations between the web pages of the tool can be observed. Some pages like error.html (to control the security of the prototype) and exit (to close the session) are related to every web page.

This figure could help possible tool changes or facilitate new versions in the future. While new code is written Figure 13 shows all the interactions inside the prototype.

![Web map diagram]

Figure 13: Web map of the prototype.
4.6 Quick guide to the web-tool

4.6.1 Login into the system

To access to the tool, the prototype provides this login screen (Figure 14). The user will have to introduce the user name and the password. If the user does not have yet a user name and password, the prototype provides the link “Create a new account” which links to Figure 15.

![Figure 14: Screenshot – Login](image)

![Figure 15: Screenshot – New User](image)
4.6.2 Main user page

This is the main screen that the user will find in the system after being logged in into the system.

In the left column the green button links always to this main page. Then the name of the logged user is shown and the door icon is the button to log out of the system and in the right column the two main aspects of the tool are found: The Statistics and Monte Carlo.

On the right column all the forecasts the user has made are stored and also the choice to do new ones is available.

![Figure 16: User options: creating new forecasts](image)

In Figure 16 are explained the two different options a user has in the prototype:

- It has the title “The Statistics”. It is possible to see the results of old user forecasts. Also to create forecasts with a provided excel, fill it and upload it to the prototype to obtain the statistical and data mining results. At the end of this section the button “New user forecasting” can be found to create new forecasts. Clicking in the excel icon the excel file the user sent to the prototype can be downloaded. Then there is a name to differentiate each forecast. And after the name, three icons for the three possible activities of the prototype: data mining report, statistics report and delete forecast. Data mining report links to Figure 19. In that page analysis data can be selected. Statistics report refers to a pdf file summarizing the excel file sent by the user with more tables and graphs. Also the forecast can be deleted.

- It has the title “Monte Carlo”. It is possible to see the results of old Monte Carlo simulations or to create a new simulation. At the end of the section the button “New Monte Carlo Simulation” can be found. Monte Carlo does not have an excel file from the user. All the data will be simulated. The structure is the same. Data mining button links to Figure 20, only changing the title. Report button refers to a pdf file summarizing of the best case found with Monte Carlo simulation, as well as some graphs.

A screenshot of the main menu can be reviewed in Figure 16.
Figure 17: Screenshot – Main user page
4.6.3 New User Forecast

In this screenshot, the user can add a new forecast. First, a name for differentiating all the forecasts is required. After that the system asks for the excel file. Clicking on the excel icon (the red square), the user can get the excel file that has to be edited with the expected forecast.

When the excel file is ready, the user just has to upload it. And then clicking on the button “Next”, the system saves the file and generates some of the required outcomes. If the operation was successful, a screen will appear.

Figure 18: Screenshot – New user forecast
4.6.4 Data Mining Data Selection

By clicking in the data mining icon from “The Statistics” table in Figure 17, the input of the data mining algorithms can be selected, between:

1. The scenarios. The system evaluates data mining only for the cases the user inserted to compare.
2. The output metrics. The system evaluates data mining to the excel file the user uploaded.

In Figure 19 the user selects the data to which data mining will be evaluated.

Figure 19: Screenshot – data mining user selection
4.6.5 Data mining results

This page is the same for all the data mining buttons found in Figure 17. Only the title differs from the rest. It can be reached to apply data mining to:

1. User scenarios
2. User output metrics
3. Monte Carlo simulation

First of all, clicking in the ARFF icon, the generated file can be viewed. This file is the one needed by Weka. It is possible to download the file to use it with all the functionality of Weka if the user has it installed.

Three different types of data mining algorithms were selected, considering that the data is numeric: Rules, Trees and Regressions. There are three rule algorithms, two tree algorithms and two regressions. Clicking on the icon tagged with “txt”, the result of that specific algorithm can be obtained.

Figure 20: Screenshot – Data mining results
4.6.6 New Monte Carlo simulation

The simulation is maybe the most fragile operation of the prototype. The simulation generates fifteen thousand rows. It is an operation which takes time, depending on the server, from thirty seconds to a few minutes. The system has to wait until all the operations are done. In Figure 21 the system asks the user to give the simulation a name, and then it can be identified later in the main page.

Figure 21: Screenshot – New Monte Carlo simulation

Figure 22: Screenshot – New Monte Carlo simulation working
The Figure 22 is the screen made to let the user know that the simulation is being done. The time to finish the simulation depends on the server. A fast server maybe can only take thirty seconds. But if it tried at home, in a laptop or an old computer, it could take three or four minutes. During this time the user has to wait, not touch any button of the system. It is a really fragile operation, generating fifteen thousand rows and then generating all the data mining results of the better year and the pdf file summarizing all the information.

When Figure 23 is showed, all the operations will have finished. If the user wants to see the results, back in the main page, in Figure 17, the new simulation will be found there.

Figure 23: Screenshot – New Monte Carlo simulation done
4.6.7 Log out

When the user logs out of the system, Figure 24 will appear just to ensure that the session was closed correctly.

![Figure 24: Screenshot – User logged out](image)

4.6.8 Security warning

The system is safe. It is prepared to show, the warning alert when someone is trying to access to it without being logged into it, as shown in Figure 25.

![Figure 25: Screenshot – Security](image)
4.7 Results obtained from the tool

The prototype is focused on obtaining data mining conclusions from the forecast the user uploaded or from the Monte Carlo simulation. The forecasts are always made for five years.

Focusing on the results the prototype gives itself about data mining, seven different measures were selected from three different algorithms to be applied to the data:

- Rules:
  - Conjunctive Rule
  - M5 Rule
  - ZeroR

- Trees
  - Decision stump
  - M5P

- Regressions
  - Additive regression
  - Linear regression

Why these algorithms? First of all, the user wants to find patterns, which fields are really influencing the others to make better forecasts in the future. The prototype is using the Weka classifier to obtain these results, to make a classification of them. And also the data is numerical, not nominal. Data mining has algorithms for numerical data and some others for nominal data. The prototype is showing the algorithms that give more useful information about the results and one of the main objectives was giving enough results, not only one. The user can compare different algorithms.

These algorithms are applied in the three different results pages of the prototype:

- User scenarios
- User output metrics
- Monte Carlo simulation

How to evaluate the results? In all the Weka outputs, first the result of the algorithm is showed, for instance the found rules or the leaves of the decision tree.

All the outputs of the seven algorithms will be shown in the next pages, but the more important aspects are:

- The first paragraphs always show the result of the algorithm. In the case of this prototype, always the result is the profit. From which values depends it or the function of the profit. Will be explained in next pages.
- To ensure the validation of the algorithm, the correlation coefficient has to be studied, in both “error on training data” and “cross validation”. This coefficient will help to ensure the accuracy, the validation of the algorithm. It will help to know if the result can be trusted or not.
o When it is 1, the correlation coefficient is perfect. Is a direct relation.
o When it is -1, the correlation is also perfect, but inverse.
o When it is 0, there is no relation.

The user will decide if 0.60 is accurate enough or he needs at least a 0.75 to think that was a good result. To decide the valid proximity to 1 to trust the result is up to the user, but it can be proposed, as an example, that if the value is between the direct or the indirect range, it could be trusted:

- **Direct relation**: $0.60 < \text{value} < 1$
- **Indirect relation**: $-1 < \text{value} < -0.60$
- Not clear the relation $-0.60 < \text{value} < 0.60$

Always the fields in bold in the next pages are the fields the user or the decision maker has to focus on.
4.7.1 Conjunctive rule

Taking a look at the bold texts, this is one of the conclusions of the rule. For experts in statistics, Weka always shows statistics about the operation.

The first lines are always the result, in this case, the profit will be 10124.85 $ if the \textit{averageRatePerMove} is greater or equal to 152.5. This is the main result of the algorithm.

Now that result has to be validated. All the results have to been validated; check the accuracy of the algorithm. For that task, the correlation coefficient has to be checked. As it was explained in the beginning of 4.7, how to evaluate the results, that variable has to be evaluated.

In this case in both cases, look at the highlighted lines, the value is 0.89 and 0.75. It is between the 0.60 - 1 interval, so it seems to be a good direct relation. This rule can be trusted.

\begin{verbatim}
Single conjunctive rule learner:
---------------------------------
(averageRatePerMove <= 152.5) => Profit = 10124.85

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

=== Error on training data ===
Correlation coefficient                  0.8924
Mean absolute error                   4196.696
Root mean squared error               4826.8575
Relative absolute error               50.0753 %
Root relative squared error           45.1783 %
Total Number of Instances             15

=== Cross-validation ===
Correlation coefficient                  0.7545
Mean absolute error                   5688.4926
Root mean squared error               7354.4669
Relative absolute error               68.2168 %
Root relative squared error           68.3564 %
Total Number of Instances             15
\end{verbatim}

Figure 26: Conjuctive Rule results
### 4.7.2 M5 Rule

This is another type of rule. The result obtained is:

\[
\text{PROFIT} = 0.0884 \times \text{revenuesThroughputTEUs} + 2110.9476
\]

The profit is dependent of the revenuesThroughputTEUs. So in this case decision-makers can change the value of that variable, to know how it will affect to the profit.

As was explained in the algorithm before, to ensure the accuracy of this algorithm, the correlation coefficient has to be evaluated. In this case, it is close to the limit given in 4.7 how to evaluate the results. The first value is 0.69 but the second is 0.57. In this case the result of this algorithm has to be used carefully, because maybe is not really accurate.

<table>
<thead>
<tr>
<th>M5 pruned model rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>(using smoothed linear models) :</td>
</tr>
<tr>
<td>Number of Rules: 1</td>
</tr>
</tbody>
</table>

**Rule: 1**

\[
\text{PROFIT} = 0.0884 \times \text{revenuesThroughputTEUs} + 2110.9476 \quad [5/71.433%]
\]

Time taken to build model: 0.05 seconds
Time taken to test model on training data: 0 seconds

```plaintext
=== Error on training data ===

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0.6998</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>65326.3167</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>70981.6113</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>84.0295 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>71.4335 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>5</td>
</tr>
</tbody>
</table>

=== Cross-validation ===

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0.576</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>69192.5486</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>89887.8245</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>71.2021 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>72.368 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>5</td>
</tr>
</tbody>
</table>
```

Figure 27: M5 rule results
4.7.3 ZeroR

ZeroR does not give rules. It gives a predicted value of the average, in this case, the profit of the forecast with the values given from the user or from the Monte Carlo simulation. In this case the algorithm predicts that the profit will be $29193.638$.

As can be evaluated, the correlation coefficient is 0 and -0.2, so in this case data mining seems that could not find a good predicted value. That is one of the reasons why the prototype gives more than one algorithm to evaluate the data. Sometimes they cannot find a trustful result.

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>8380.7704</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>10684.0212</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>100 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>100 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>15</td>
</tr>
</tbody>
</table>

--- Error on training data ---

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>-0.2923</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>8412.3313</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>10937.7247</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>100 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>100 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>15</td>
</tr>
</tbody>
</table>

Figure 28: ZeroR results
4.7.4 Decision stump

Decision stump will create a one-level binary decision tree. In this case, the result found by data mining was:

Other Revenue Per Move
- If is less or equal to 10.5, then the profit will be 10124.85
- If is greater than 10.5, then the profit will be 33960.83
- If is missing, then the profit will be 29193.638.

As before, decision-makers can iterate the value of other revenue per move to know how it affects the final profit.

Looking to the correlation coefficients, like in all the algorithms before to see the accuracy, it can be trusted, with 0.89 and 0.87 is inside of the valid range. It is also a direct relation.

<table>
<thead>
<tr>
<th>Decision Stump</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classifications</strong></td>
</tr>
<tr>
<td>OtherRevenuePerMove &lt;= 10.5: 10124.85 (= Profit)</td>
</tr>
<tr>
<td>OtherRevenuePerMove &gt; 10.5: 33960.835000000001 (=Profit)</td>
</tr>
<tr>
<td>OtherRevenuePerMove is missing: 29193.638 (= Profit)</td>
</tr>
</tbody>
</table>

Time taken to build model: 0.02 seconds
Time taken to test model on training data: 0 seconds

--- Error on training data ---

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0.8924</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>4196.696</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>4821.1658</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>50.0753 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>45.125 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>15</td>
</tr>
</tbody>
</table>

--- Cross-validation ---

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0.8772</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>4439.6906</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>5137.8163</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>53.2411 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>47.7536 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>15</td>
</tr>
</tbody>
</table>

Figure 29: Decision stump results
4.7.5 M5P

M5P generates a tree, in this case using the prototype will not show the graphical tree (only a text tree) but using the Weka GUI, it is possible to see the graphical tree.

As Figure 30 shows, data mining has found that `averageRatePerMove` is important for the profit and the tree indicates that. Following the tree and iterating with the value of average rate per move, the profit will be LM 1, LM 2 or LM 3.

- LM 1 means that the profit will be a function of:
  \[ \text{Profit} = 1513.6794 \times \text{averageRatePerMove} - 213148.7635 \]
  In this case average rate per move is less or equal to 152.5

- LM 2 means that the profit will be a function of:
  \[ \text{Profit} = 1484.6092 \times \text{averageRatePerMove} - 207354.8462 \]
  In this case average rate per move is greater than 152.5 but less or equal to 162.5

- LM 3 means that the profit will be a function of:
  \[ \text{Profit} = 1135.5969 \times \text{averageRatePerMove} - 150539.2556 \]
  In this case average rate per move will be greater than 162.5

In this case the correlation coefficients are 0.96 and 0.95, really close to 1, so in this algorithm the results can be trusted. They seem to be really accurate, showing a direct relation between average rate per move and the profit.
Options: -M 4.0

M5 pruned model tree:
(using smoothed linear models)

\( \text{averageRatePerMove} \leq 162.5 : \)
| \( \text{averageRatePerMove} \leq 152.5 : \text{LM1} (3/0\%) \)
| \( \text{averageRatePerMove} > 152.5 : \text{LM2} (6/0\%) \)
\( \text{averageRatePerMove} > 162.5 : \text{LM3} (6/0\%) \)

LM num: 1
\[ \text{Profit} = 1513.6794 \times \text{averageRatePerMove} - 213148.7635 \]

LM num: 2
\[ \text{Profit} = 1484.6092 \times \text{averageRatePerMove} - 207354.8462 \]

LM num: 3
\[ \text{Profit} = 1135.5969 \times \text{averageRatePerMove} - 150539.2556 \]

Number of Rules : 3

Time taken to build model: 0.08 seconds
Time taken to test model on training data: 0 seconds

=== Error on training data ===

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0.9628</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>2394.8846</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>2892.785</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>28.5759 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>28.5759 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>15</td>
</tr>
</tbody>
</table>

=== Cross-validation ===

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0.9569</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>2696.0406</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>3157.1083</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>32.0487 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>28.8644 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>15</td>
</tr>
</tbody>
</table>

Figure 31: M5P results
4.7.6 Additive regression

The additive regression gives 10 models using decision stump, which was used before also. Each iteration of the regression fits a model to the residuals left by the classifier on the previous iteration. Prediction is accomplished by adding the predictions of each classifier.

In each model, the value that the algorithms always predict is the profit. The interpretation is the same used in 4.7.4. But in this case the algorithm provides 10 different decision stumps.

In Figure 32, the result of the profit is shown with all the variables.

The statistics seem to be good, with a high correlation coefficient and low errors. The correlation is 0.99 in both cases. It is a really good direct relation. All the variables evaluated in this algorithm have a direct relation with the profit.

---

Additive Regression

ZeroR model

**ZeroR predicts class value: 29193.638 (= Profit)**

Base classifier weka.classifiers.trees.DecisionStump

10 models generated.

Model number 0

Decision Stump

Classifications

- **OtherRevenuePerMove <= 10.5 : -19068.788 (= Profit)**
- **OtherRevenuePerMove > 10.5 : 4767.197000000002 (= Profit)**
- **OtherRevenuePerMove is missing : 1.4551915228366853E-12 (=Profit)**

Model number 1

Decision Stump

Classifications

- **OtherRevenueOptional <= 4075000.0 : -3497.246666666667 (= Profit)**
- **OtherRevenueOptional > 4075000.0 : 5245.870000000002 (= Profit)**
- **OtherRevenueOptional is missing : 6.063298011819521E-13 (=Profit)**

Model number 2

Decision Stump

Classifications

- **averageRatePerMove <= 152.5 : 3497.246666666667 (= Profit)**
- **averageRatePerMove > 152.5 : -874.3116666666674 (= Profit)**
- **averageRatePerMove is missing : -5.456968210637569E-13 (= Profit)**

Model number 3

Decision Stump
Classifications

\[
\text{averageRatePerMove } \leq 167.5 : -480.57416666666694 \text{ (Profit)} \\
\text{averageRatePerMove } > 167.5 : 1922.296666666667 \text{ (Profit)} \\
\text{averageRatePerMove is missing } : -9.094947017729283E-14 \text{ (Profit)}
\]

*****************************************************************************
models 4 – 5 – 6 – 7 – 8
*****************************************************************************

Model number 9

Decision Stump

Classifications

\[
\text{averageRatePerMove } \leq 152.5 : 427.2706249999999 \text{ (Profit)} \\
\text{averageRatePerMove } > 152.5 : -106.81765624999998 \text{ (Profit)} \\
\text{averageRatePerMove is missing } : -1.5158245029548804E-14 \text{ (Profit)}
\]

Time taken to build model: 0.05 seconds
Time taken to test model on training data: 0.02 seconds

=== Error on training data ===

\begin{tabular}{ll}
\textbf{Correlation coefficient} & 0.9998 \\
Mean absolute error  & 197.6359 \\
Root mean squared error & 236.9786 \\
Relative absolute error & 2.3582 % \\
Root relative squared error & 2.2181 % \\
Total Number of Instances & 15 \\
\end{tabular}

=== Cross-validation ===

\begin{tabular}{ll}
\textbf{Correlation coefficient} & 0.9965 \\
Mean absolute error  & 663.0618 \\
Root mean squared error & 1002.915 \\
Relative absolute error & 7.882 % \\
Root relative squared error & 9.1693 % \\
Total Number of Instances & 15 \\
\end{tabular}

Figure 32: Additive regression results
4.7.7 Linear Regression

In the linear regression the objective is to fit an equation to a dataset. The simplest form is the linear regression, which uses the formula of a straight line (y=mx+b).

In Figure 33 it is shown the result of the profit with all the variables. The profit expressed in a function of all the variables. Maybe decision-makers can be a useful application for this formula.

The statistics seem to be good, with a high correlation coefficient and low errors, 1 and 0.98, inside of the range of direct relation and really close to 1; really accurate.

```
PROFIT =
0.008 * revenuesThroughputTEUs +
0.0177 * revenuesThroughputMoves +
201.0223 * revenuesAverageRatePerMove +
221.054 * revenuesOtherRevenuePerMove +
0.0602 * revenueTotalContainerRevenue +
0.0361 * revenueTotalRevenues +
0.1416 * expensesVesselLaborExpenses +
0.2836 * expensesYardLaborExpenses +
0.0973 * expensesTotalLaborExpenses +
41.7064 * expensesM&RExpenses +
0.2901 * expensesSG&A_includingRent +
0.0738 * expensesTotalExpenses +
0.0499 * expensesEBITDA +
281.4558 * vesselsCalls +
-11459.4952 * SideHandler +
3300.6184 * AverageHourlyWagePerHour_vessel +
2468.2625 * VesselLaborCostPerMove +
5910.9603 * AverageHourlyWagePerHour_yard +
166.8323 * AcreageUsedTotal +
0.2513 * groundSlotsNeededRTG +
2.0134 * groundSlotsNeededTopPick +
-1.9565 * groundSlotsNeededSidePick +
-1.8002 * revenueTeusPerVesselCall +
0.0161 * revenueTeusPerBerth +
0.2861 * revenueTeusPerShipCranes +
0.0353 * revenueMovesPerVesselBerth +
0.7166 * revenueMovesPerVesselShipCranes +
-6524.8082 * vesselOperationsHours@berth +
117498.5124 * vesselOperationsBerthUtilizationRate +
3300.6189 * AdjustedWagePerHour-vessel&yard +
3311.8344 * yardOperationsGateHours +
-4904.2089 * yardOperationsGateProcessingTime(minutes/transaction) +
-32.7717 * yardOperationsLaneHoursRequired +
5910.9605 * adjustedWagePerHour +
-117402.2045 * percentOf40Containers +
-2168588.0295

=== Error on training data ===

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>0</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>0 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>0 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>5</td>
</tr>
</tbody>
</table>

=== Cross-validation ===

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>0.9838</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>59630.7568</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>97422.5144</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>16.8637 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>24.2576 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>5</td>
</tr>
</tbody>
</table>
```

Figure 33: Linear regression results

49
4.7.8 Statistics and reports

Looking at the main page, Figure 17, not only data mining reports are given as outcomes. The prototype also gives some reports of the forecasts, in pdf format. In Figure 17 they are surrounded with a circle.

4.7.8.1 Statistic report from the user file

In the statistics, the user can obtain a pdf with a small report about the forecast made, it contains the output metrics obtained from the user data and some graphs, like TEU Growth, Rate per Box, Total Expenses/Percentage revenues, Profit and EBITDA.

In the next figures there a review about that report.

---

**Summary Report**

<table>
<thead>
<tr>
<th></th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Revenues</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Throughput (TEUs)</td>
<td>$9,519</td>
<td>$12,714</td>
<td>$13,003</td>
<td>$15,123</td>
<td>$16,473</td>
</tr>
<tr>
<td>% Growth</td>
<td>0%</td>
<td>39.1%</td>
<td>3.9%</td>
<td>20.2%</td>
<td>3.4%</td>
</tr>
<tr>
<td>% of Revenues</td>
<td>0%</td>
<td>9.9%</td>
<td>9.1%</td>
<td>89.8%</td>
<td>19.9%</td>
</tr>
<tr>
<td>Yar Labor Expenses</td>
<td>$3,765</td>
<td>$6,155</td>
<td>$6,705</td>
<td>$6,705</td>
<td>$6,200</td>
</tr>
<tr>
<td>% Growth</td>
<td>0%</td>
<td>36.2%</td>
<td>-6.7%</td>
<td>40.8%</td>
<td>2.0%</td>
</tr>
<tr>
<td>% of Revenues</td>
<td>8.9%</td>
<td>7.7%</td>
<td>6.9%</td>
<td>7.9%</td>
<td>7.4%</td>
</tr>
<tr>
<td>**Total Labor Expenses</td>
<td>$13,023</td>
<td>$18,024</td>
<td>$18,024</td>
<td>$22,058</td>
<td>$23,379</td>
</tr>
<tr>
<td>% Growth</td>
<td>0%</td>
<td>-30.5%</td>
<td>0.2%</td>
<td>25.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>% of Revenues</td>
<td>31.0%</td>
<td>27.2%</td>
<td>26.8%</td>
<td>26.4%</td>
<td>26.4%</td>
</tr>
<tr>
<td>MSL Expenses (W.O. $)</td>
<td>$5,081</td>
<td>$6,120</td>
<td>$5,913</td>
<td>$5,108</td>
<td>$6,078</td>
</tr>
<tr>
<td>% Growth</td>
<td>0%</td>
<td>5.2%</td>
<td>5.2%</td>
<td>5.2%</td>
<td>5.2%</td>
</tr>
<tr>
<td>% of Revenues</td>
<td>11.8%</td>
<td>7.9%</td>
<td>7.9%</td>
<td>6.6%</td>
<td>6.6%</td>
</tr>
<tr>
<td>O&amp;M Labor Expenses (W.O.)</td>
<td>$13,164</td>
<td>$15,664</td>
<td>$15,857</td>
<td>$18,973</td>
<td>$19,175</td>
</tr>
<tr>
<td>% Growth</td>
<td>0%</td>
<td>13.5%</td>
<td>7.8%</td>
<td>15.8%</td>
<td>4.3%</td>
</tr>
<tr>
<td>% of Revenues</td>
<td>32.3%</td>
<td>33.7%</td>
<td>33.8%</td>
<td>25.5%</td>
<td>24.9%</td>
</tr>
<tr>
<td>**Total Expenses (W.O. $)</td>
<td>$31,572</td>
<td>$38,080</td>
<td>$38,463</td>
<td>$48,991</td>
<td>$48,994</td>
</tr>
<tr>
<td>% Growth</td>
<td>0%</td>
<td>22.2%</td>
<td>-1.3%</td>
<td>19.8%</td>
<td>3.0%</td>
</tr>
<tr>
<td>% of Revenues</td>
<td>75.5%</td>
<td>58.8%</td>
<td>56.7%</td>
<td>25.1%</td>
<td>54.7%</td>
</tr>
<tr>
<td>EBITDA (W.O. $)</td>
<td>$10,125</td>
<td>$18,100</td>
<td>$20,100</td>
<td>$38,159</td>
<td>$40,255</td>
</tr>
<tr>
<td>% Margin</td>
<td>24.6%</td>
<td>41.5%</td>
<td>41.9%</td>
<td>44.9%</td>
<td>45.3%</td>
</tr>
<tr>
<td>per Share</td>
<td>$0.51</td>
<td>$0.88</td>
<td>$0.81</td>
<td>$0.11</td>
<td>$0.13</td>
</tr>
<tr>
<td><strong>Capex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintenance</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
</tr>
<tr>
<td>Commission</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
</tr>
<tr>
<td>Expansion</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
</tr>
<tr>
<td><strong>Cumulative</strong></td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
<td>$ -</td>
</tr>
<tr>
<td>Profit (W.O. $)</td>
<td>$10,125</td>
<td>$18,100</td>
<td>$20,100</td>
<td>$38,159</td>
<td>$40,255</td>
</tr>
<tr>
<td>Key Operating Metrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vessel/Calls</td>
<td>235</td>
<td>250</td>
<td>310</td>
<td>362</td>
<td>362</td>
</tr>
<tr>
<td>Sells</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Post Office</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

---

Figure 34: Statistics user report, page 1.
In Figure 34 can be shown the outputs of the revenues (included the total revenues), expenses (included the total expenses) and the beginning of the key operation metrics.

The profit of the forecast (Total expenses - total revenues) is also calculated, is the main variable that the prototype uses to apply data mining. It is considered the most important variable for the company.

<table>
<thead>
<tr>
<th></th>
<th>9</th>
<th>9</th>
<th>9</th>
<th>9</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Deck</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>RMG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BSKs</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metrics</th>
<th>36.68</th>
<th>37.64</th>
<th>36.12</th>
<th>40.39</th>
<th>41.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Hourly Wage per</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
</tr>
<tr>
<td>hour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance in labor cost</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
</tr>
<tr>
<td>per hour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Hourly Wage per</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
</tr>
<tr>
<td>hour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year-on-Year change per</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
</tr>
<tr>
<td>MTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Gross Average            | 152.2 | 150.0 | 150.0 | 152.2 | 150.0 |
| Net Average              | 152.2 | 150.0 | 150.0 | 152.2 | 150.0 |
| Average Used (Total)     | 27.0  | 40.5  | 40.5  | 47.3  | 47.3  |
| Whisked                  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| RTS                      | 12.3  | 16.4  | 16.4  | 24.4  | 24.4  |
| Top Pick                 | 0.2  | 13.0  | 13.0  | 16.8  | 16.8  |
| Sick Pick (MTY Yard)     | 5.0  | 8.0  | 8.0  | 10.1  | 10.1  |
| RMS                       | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Steal                    | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Capacity Utilization     | 0.24  | 0.36  | 0.36  | 0.42  | 0.42  |
| Ground Slots Reserved    | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| Whisked                  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| RTS                      | 4.92  | 7.24  | 7.24  | 8.44  | 8.44  |
| Top Pick                 | 1.44  | 2.18  | 2.18  | 2.52  | 2.52  |
| Sick Pick (MTY Yard)     | 5.48  | 8.18  | 8.18  | 9.65  | 9.65  |
| RMS                       | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Steal                    | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Revenue TEUs per         | 1.533 | 2.000 | 1.666 | 1.666 | 1.666 |
| Vessel Call              | 172.000 | 256.000 | 256.000 | 301.000 | 301.000 |
| Ship Crane               | 172.000 | 129.000 | 129.000 | 155.000 | 155.000 |
| Revenue Mils per         | 772  | 1,033 | 258  | 257  | 947  |
| Vessel Call              | 100.000 | 150.000 | 168.000 | 175.000 | 175.000 |
| Ship Crane               | 100.000 | 75.000  | 76.000  | 87.500  | 87.500  |
| Vessel Operations        |       |       |       |       |       |
| Receiver Factor          | 1.05  | 1.05  | 1.05  | 1.05  | 1.05  |
| Total Mils per call      | 1.21  | 1.21  | 1.21  | 1.21  | 1.21  |
| Change                    | 1.05  | 1.05  | 1.05  | 1.05  | 1.05  |
| Hour @ berth              | 33   | 54   | 54   | 54   | 54   |
| Target hours @ berthamt   | 38   | 38   | 38   | 38   | 38   |
| berth Utilization Rate    | 60.0% | 48.5% | 48.5% | 53.0% | 53.0% |
| Adjusted Mils per hour    | 37.08 | 37.08 | 30.12 | 42.39 | 41.50 |
| % Growth                  | 0    | 3.0% | 3.0% | 3.0% | 3.0% |
| Vessel Operations         |       |       |       |       |       |
| Contact                  | 10   | 10   | 10   | 10   | 10   |
| Gate Processing Time (min)| 7    | 7    | 7    | 1    | 1    |
| Gate Failing Factor       | 1.10  | 1.10  | 1.10  | 1.10  | 1.10  |

Figure 35: Statistics user report, page 2.

In Figure 35 the second part of the output metrics are shown.
In Figure 36 the end of the output metrics and the beginning of the graphs part. In this page the EBITDA and its % margin are represented.

EBITDA is the difference between total operating revenues and total operating expenses.

The % Margin represents the percentage of the EBITDA value divided by the total revenues.
In Figure 37 two more graphs are represented. This first figure represents the relation between the Throughtput (TEUs) and its own growth percentage.

The second figure, Rate per box represents the relation between the rate per box and its own growth percentage.
Figure 38: Statistics user report, page 5.

In Figure 38 the figure defines itself; it is the relation between the total expenses and the percentage of revenues. It shows the percentage of the revenues represented by the total expenses.

The last diagram is a pie chart which could be considered interesting. Each slice represents the percentage of the profit of each year.

4.7.8.2 Report from the Monte Carlo simulation

Another report is created about the Monte Carlo simulation. If the result matches with the best case achieved with the simulation, it is the case that generates a better profit of the forecast. Also some statistical operations are applied to it. An example is shown in the next figures.
The Figure 39 summarizes the Monte Carlo simulation. For a better understanding of this figure, all the defined values are explained in page 66.
In Figure 40 can be shown the outputs of the revenues (included the total revenues), expenses (included the total expenses) and the beginning of the key operation metrics.
In Figure 41 more of the output metrics are shown.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Hourly Wage per hour (vessel)</td>
<td>$127.00</td>
</tr>
<tr>
<td>Net Hourly Wage per hour (vessel)</td>
<td>$176.09</td>
</tr>
<tr>
<td>Average Hourly Wage per hour (yard)</td>
<td>$76.96</td>
</tr>
<tr>
<td>YardHourly Labor cost per Move</td>
<td>$48.21</td>
</tr>
</tbody>
</table>

| Gross Average     | 254.79 |
| Net Average       | 190.24 |

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Used (Total)</td>
<td>23.23</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.00</td>
</tr>
<tr>
<td>RT4</td>
<td>24.10</td>
</tr>
<tr>
<td>RT5</td>
<td>0.18</td>
</tr>
<tr>
<td>Site Pick (MTY Yard)</td>
<td>4.20</td>
</tr>
<tr>
<td>RMG</td>
<td>0.00</td>
</tr>
<tr>
<td>ISed</td>
<td>0.00</td>
</tr>
<tr>
<td>Capacity Utilization</td>
<td>0.15</td>
</tr>
<tr>
<td>Ground Sets Needed</td>
<td>0.00</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.00</td>
</tr>
<tr>
<td>RT4</td>
<td>10.974.88</td>
</tr>
<tr>
<td>RT5</td>
<td>28.49</td>
</tr>
<tr>
<td>Site Pick (MTY Yard)</td>
<td>2.432.43</td>
</tr>
<tr>
<td>RMG</td>
<td>0.00</td>
</tr>
<tr>
<td>ISed</td>
<td>0.00</td>
</tr>
</tbody>
</table>

| Revenue Items per Vessel Call               | $1.376.00 |
| Vessel Call                                 | $230.978.40 |
| Berth                                       | $230.978.40 |

| Revenue Items per Vessel Call               | $832.54 |
| Vessel Call                                 | $128.005.18 |
| Berth                                       | $128.005.18 |

| Vessel Operations                           |      |
| Transit Time per call                       | 51.54 |
| Capacity                                    | 1.05 |
| Adjusted Wage per hour (vessel & yard)      | $127.00 |
| % Growth                                    | 3.00% |

| Yacht Operations                            |      |
| Gate Hours                                  | 9.25 |
| Gate Processing Time (minutes per transaction) | 7 |
| Gate Receiving Factor                       | 1.1 |
| Lane Hours Required                         | 134.31 |
| Move/Make Transaction Ratio                 | 1.2 |
| Weekly Yard/Peak Factor                     | 1.1 |
In Figure 42 are shown some of the output metrics and the generated inputs by Monte Carlo. These inputs are important for the user to know which scenario was simulated. This scenario can help to make better decisions. In Figure 43 the rest of the generated inputs are shown.
5 SIMULATION EXPERIMENT

The experiment realizes a Monte Carlo simulation, which means generate random data and analyze the results. The experiment consists of generating five years forecast three thousand times; giving the results of analyzing the generated data afterwards. Steps followed to make the experiment:

1. Create the model
2. Generating random inputs
3. Evaluating the model
4. Run simulation

5.1 Model

The model used is contained in the file userFile_MC_v5_Release2.xls that contains thousand of variables. The structure of the worksheets of the file is shown in the Figure 44.

![Figure 44: Worksheets structure of the Monte Carlo simulation](image)

GlobalInputs, OperationTemplate, CapacityTemplate worksheets contain all the inputs of the model. OutputMetrics and Analysis worksheets are the outcomes of the model. MC Sim is a temporal worksheet that is used to keep a record of all the results obtained during the simulation, Monte Carlo worksheet is the responsible to calculate the distribution and the best case in the results stored in MC Sim worksheet.
The first issue was to decide which variables are going to be randomized, not all of them are good candidates to be randomized; taking into account the variables listed in Table 1:

<table>
<thead>
<tr>
<th>INPUTS</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Moves per Customer</strong></td>
<td>Company 1</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Company 2</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Company 3</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td><strong>Management Projection</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average Rate Per Move</strong></td>
<td>Company 1</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Company 2</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Company 3</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td><strong>Other Revenue per move (optional)</strong></td>
<td>Company 1</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Company 2</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Company 3</td>
<td>% Growth</td>
<td></td>
</tr>
<tr>
<td><strong>Vessel Calls</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vessel Base Wage Per Hour</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Yard/Gate Base Wage Per Hour</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vessel Expense - Labor</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Capacity - Inputs</strong></td>
<td>Percent of 40' (or larger) Containers</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gross Acres</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model Data (User Input)</strong></td>
<td>Imports</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Throughput Distribution</strong></td>
<td>Exports</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empties</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transhipment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intermodal Imports</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intermodal Exports</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: variables randomized by Monte Carlo
5.2 How it works

This section is focused on how the experiment and how it works, the figure below shows the flow how the user starts the simulation in the server and which results are obtained.

Figure 45: Monte Carlo simulation in the web-tool prototype

The prototype launches Microsoft Office Excel to open the XLS file, which auto executes sequences of macros at the opening of the file, as Figure 45. Steps followed during the simulation:

1. Calculate and copy the results in MonteCarlo worksheet
2. If number of iterations are less than 3000 go to step 1.
3. Calculate the profit
4. Find the best case calculated by MC Sim worksheet
5. Export Analysis worksheet to a PDF file.
6. Export data to be used by data mining algorithms to a XLS file called dummy.xls
7. Quit the application.

Next, it is explained all these steps showing also the screenshots took during the simulation.
After the login the user access to his area, it is possible to difference two different sections *The statistics* and *Monte Carlo*, clicking in the link *New MonteCarlo Forecasting* the user access to the simulation.
It is possible to give the desired name to the simulation and after that press the button *Generate simulation*, in this moment is the beginning of the simulation, depending how powerful is the computer; taking that into account the process will take between 30 seconds or a couple of minutes.

As it could see all that occurs in the side of the user, now in the side of the server after the click of the user the application starts, so the Excel file starts copying all the results every iteration. In the figure above it is represented the moment of one of those iterations all the data calculated using the model with random data. As it is possible to see the outputs of five years, all that data is copy transpose in the cells next to the outputs. After three thousand iterations all the data is copied to *MonteCarlo* worksheet.

*MonteCarlo* worksheet contains all the results obtained from the simulation; an analysis of the data obtained after the iterations it is done by *MC Sim* worksheet (Table 2 and Table 3) calculating the distribution of all the values and taking the best case
afterwards. This analysis is going to be explained further along and taking into account the tables below:

### Histogram Plot

<table>
<thead>
<tr>
<th>Count</th>
<th>Bins</th>
<th>Scaled</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-369962</td>
<td>0</td>
<td>0.000266667</td>
</tr>
<tr>
<td>4</td>
<td>-323866</td>
<td>5.78503E-09</td>
<td>0.000733333</td>
</tr>
<tr>
<td>7</td>
<td>-277770</td>
<td>1.01238E-08</td>
<td>0.001333333</td>
</tr>
<tr>
<td>9</td>
<td>-231674</td>
<td>1.30163E-08</td>
<td>0.001833333</td>
</tr>
<tr>
<td>34</td>
<td>-185578</td>
<td>4.91727E-08</td>
<td>0.004033333</td>
</tr>
<tr>
<td>64</td>
<td>-139482</td>
<td>9.25605E-08</td>
<td>0.007866667</td>
</tr>
<tr>
<td>140</td>
<td>-93386</td>
<td>2.02476E-07</td>
<td>0.0172</td>
</tr>
<tr>
<td>254</td>
<td>-47290</td>
<td>3.67349E-07</td>
<td>0.034133333</td>
</tr>
<tr>
<td>500</td>
<td>-1194</td>
<td>7.23129E-07</td>
<td>0.067466667</td>
</tr>
<tr>
<td>799</td>
<td>44904</td>
<td>1.15556E-06</td>
<td>0.120733333</td>
</tr>
<tr>
<td>1031</td>
<td>91002</td>
<td>1.49109E-06</td>
<td>0.189466667</td>
</tr>
<tr>
<td>1119</td>
<td>137100</td>
<td>1.61836E-06</td>
<td>0.264066667</td>
</tr>
<tr>
<td>1164</td>
<td>183198</td>
<td>1.68344E-06</td>
<td>0.341666667</td>
</tr>
<tr>
<td>1180</td>
<td>229296</td>
<td>1.70658E-06</td>
<td>0.420333333</td>
</tr>
<tr>
<td>1115</td>
<td>275394</td>
<td>1.61258E-06</td>
<td>0.494666667</td>
</tr>
<tr>
<td>1073</td>
<td>321492</td>
<td>1.55183E-06</td>
<td>0.5662</td>
</tr>
<tr>
<td>1024</td>
<td>367590</td>
<td>1.48097E-06</td>
<td>0.634466667</td>
</tr>
<tr>
<td>896</td>
<td>413688</td>
<td>1.29585E-06</td>
<td>0.6942</td>
</tr>
<tr>
<td>799</td>
<td>459786</td>
<td>1.15556E-06</td>
<td>0.747466667</td>
</tr>
<tr>
<td>683</td>
<td>505884</td>
<td>9.87794E-07</td>
<td>0.793</td>
</tr>
<tr>
<td>607</td>
<td>551982</td>
<td>8.77878E-07</td>
<td>0.833466667</td>
</tr>
<tr>
<td>543</td>
<td>598080</td>
<td>7.85318E-07</td>
<td>0.869666667</td>
</tr>
<tr>
<td>431</td>
<td>644178</td>
<td>6.23337E-07</td>
<td>0.8984</td>
</tr>
<tr>
<td>333</td>
<td>690276</td>
<td>4.81604E-07</td>
<td>0.9206</td>
</tr>
<tr>
<td>299</td>
<td>736374</td>
<td>4.32431E-07</td>
<td>0.940533333</td>
</tr>
<tr>
<td>212</td>
<td>782472</td>
<td>3.06607E-07</td>
<td>0.954666667</td>
</tr>
<tr>
<td>181</td>
<td>828570</td>
<td>2.61773E-07</td>
<td>0.966733333</td>
</tr>
<tr>
<td>143</td>
<td>874668</td>
<td>2.06815E-07</td>
<td>0.976266667</td>
</tr>
<tr>
<td>97</td>
<td>920766</td>
<td>1.40287E-07</td>
<td>0.982733333</td>
</tr>
<tr>
<td>83</td>
<td>966864</td>
<td>1.20039E-07</td>
<td>0.988266667</td>
</tr>
<tr>
<td>50</td>
<td>1012962</td>
<td>7.23129E-07</td>
<td>0.9916</td>
</tr>
<tr>
<td>36</td>
<td>1059060</td>
<td>5.20653E-08</td>
<td>0.994</td>
</tr>
<tr>
<td>30</td>
<td>1105158</td>
<td>4.33877E-08</td>
<td>0.996</td>
</tr>
<tr>
<td>19</td>
<td>1151256</td>
<td>2.74789E-08</td>
<td>0.997266667</td>
</tr>
<tr>
<td>14</td>
<td>1197354</td>
<td>2.02476E-08</td>
<td>0.9982</td>
</tr>
<tr>
<td>7</td>
<td>1243452</td>
<td>1.01238E-08</td>
<td>0.998666667</td>
</tr>
<tr>
<td>6</td>
<td>1289550</td>
<td>8.67754E-09</td>
<td>0.999066667</td>
</tr>
<tr>
<td>5</td>
<td>1335648</td>
<td>7.23129E-09</td>
<td>0.9994</td>
</tr>
<tr>
<td>5</td>
<td>1381746</td>
<td>7.23129E-09</td>
<td>0.999733333</td>
</tr>
<tr>
<td>3</td>
<td>1427844</td>
<td>4.33877E-09</td>
<td>0.999933333</td>
</tr>
<tr>
<td>0</td>
<td>1473942</td>
<td>0</td>
<td>0.999933333</td>
</tr>
</tbody>
</table>

Table 2: MC Sim worksheet: histogram plot
### Summary Statistics

<table>
<thead>
<tr>
<th>Central Tendency (Location)</th>
<th>Sample Size (N): 15000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean: $312.168.74</td>
<td>Median: $278.868.15</td>
</tr>
<tr>
<td>StdErr: $2.005.23</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spread</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>StDev: $245.590.13</td>
<td></td>
</tr>
<tr>
<td>Max: $1.473.945.98</td>
<td>Q(.75): $461.999.36</td>
</tr>
<tr>
<td>Min: ($369.961.15)</td>
<td>Q(.25): $128.113.26</td>
</tr>
<tr>
<td>Range: $1.843.907.13</td>
<td>IQ Range: $333.886.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Shape</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness: 0.68144339</td>
<td></td>
</tr>
<tr>
<td>Kurtosis: 0.432104477</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantiles, Percentiles, Intervals</th>
<th>90% Interval</th>
<th>95% Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q(.05):</td>
<td>($23.435.50)</td>
<td>Q(.025): ($68.713.52)</td>
</tr>
<tr>
<td>Q(.95):</td>
<td>$764.398.00</td>
<td>Q(.975): $870.266.06</td>
</tr>
<tr>
<td>Alpha (a):</td>
<td>0.05</td>
<td>Q(a/2): ($68.713.52)</td>
</tr>
<tr>
<td>% Interval: 95%</td>
<td>Q(1-a/2): $870.266.06</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probabilities</th>
<th>Pr( y &gt; 0 ): 93%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr( y &lt; 350000 )</td>
<td>= 60.84%</td>
</tr>
<tr>
<td>Pr( y &gt; 1000000 )</td>
<td>= 0.91%</td>
</tr>
<tr>
<td>Pr( 350000 &lt; y &lt; 1000000 )</td>
<td>= 38.25%</td>
</tr>
<tr>
<td>Alpha (a):</td>
<td>= 0.6175</td>
</tr>
</tbody>
</table>

### Table 3: MC Sim worksheet: summary statistics

Concepts listed in the previous tables:

- **Sample Mean**: The average of the sampled data. An estimate of the true population mean.

- **Standard Error**: The uncertainty associated with the estimated mean. The standard error is an estimate of the standard deviation of the sample means for repeated MC simulations. It can be used to calculate confidence intervals for the true population mean.

- **Median**: The "value in the middle" or the 50th percentile. For an even number of points, the median is the average of the middle two points.

- **St Dev**: Sample Standard Deviation
- **Interquartile Range**: Q(0.75) - Q(0.25) or Q3 - Q1

- **Skewness**:
  - Positive: skewed to the right (longer right-hand tail)
  - Negative: skewed to the left (longer left-hand tail)

- **Kurtosis**: Peakedness, compared to a Normal distribution.

Steps followed in the simulation:

1. **Sample size**: The sample size N (showed in Table 2) is the amount of generated data after all the iterations.

2. **Central tendency**: The sample mean and median statistics give the central tendency, which is taken like the best case.

3. **Standard derivation, range and quartiles**: The standard derivation and range show spread is the data that was obtained; standard derivation is given using the STDEV function on Excel. The range gives to the calculation the measure between the maximum and the minimum value. After that is representing the four quartiles Q1 first quartile (25%), Q2 the median value (50%), Q3 third quartile (75%) and Q4 Maximum value.

4. **Shape**: The skewness describes if the distribution is going to have a right-hand tail or is going to be into the left. The Kurtosis indicates how flat is going to be shown the distribution, depends of how positive or negative is this value.

5. **Confidence interval**: The standard calculation is obtained using the formula below:

   \[ \text{StErr} = \frac{s}{\sqrt{n}} \]

   Using the standard error it is calculated the confidence intervals for the true population mean, which is calculated as:

   \[ 95\% UCL = \text{Mean} + 1.96 \times \text{StErr} = \overline{y} + 1.96 \times \frac{s}{\sqrt{n}} \]

   \[ 95\% LCL = \text{Mean} - 1.96 \times \text{StErr} = \overline{y} - 1.96 \times \frac{s}{\sqrt{n}} \]

   Upper Confidence Limit (UCL) and Lower Confidence Limit (LCL)
6. **Cumulative distribution**: Analyzing the results is shown by the histogram below, in order to visualize the uncertainty in profit [29]

![Cumulative distribution histogram](image)

Figure 46: Graph of the estimated cumulative distribution from the analysis worksheet.

7. **Percentile and rank**: As it is shown in the Table 3 into the section *Probabilities* using the PERCENTIL and PERCENTRANK function it is possible to estimate the cumulative probability based in the Monte Carlo simulation. The percentiles shown in the Table 3 into the section *Quartiles, Percentiles, Intervals* $Q (\alpha/2)$ and $Q (1-\alpha/2)$ are useful to calculate the central interval.

After finishing the analysis with all the outputs, profits and input variables used in the best case obtained from the distribution shown below in figure 45, the prototype makes the report, a PDF file. The name of the report is the id given to the user followed with MC. The report has been commented in the section 4.7.8.2
Returning to the side of the user it will be show the next screenshot:

![Image of a simulation tool](image)

After that the user will be back at the main window, and it is possible to see the name given to the simulation realized before as it is shown in the next figure into a red line discontinued rectangle.
In both cases statistics and MonteCarlo it is possible to apply the algorithm of data mining, but focusing in the simulation and clicking on the data mining icon the tool shows a new screen, as it is shown in figure below. Into this new screen the user can access to the file in ARF format necessary to use with Weka.

Also the user can choose between one of seven algorithms. After click on one of them it will be shown the different results. All those results were commented in the section 4.7 and forward.
The figures below show the results obtained with data mining. Clicking over each algorithm the results can be shown. To evaluate and interpret them they are deeply explained in section 4.7.


6 DISCUSSION

This thesis focuses on a web-tool prototype and is not a research paper about a new topic. This web tool is something new, something that was not made before, but it uses existing technologies and methods.

The results seem to be as expected in the beginning. It cannot be said that the results are expected because data mining is really dependent on the data. If the user generates an excel forecast with wrong data, the data is incomplete, or data mining algorithms cannot find a relation between the data, it will not be really useful. That is one of the main reasons of implementing seven different data mining algorithms in the prototype. The user will be able to try seven different outputs to find relationships between the data.

Why these seven? As it was explained before, a set of random test were made to evaluate with which algorithms more useful information was obtained. Date mining is usually more focused on nominal data, so a lot of algorithms could not be used in this prototype. Using the algorithms only with numerical data, these seven gave interesting information, just to help decision-makers in the company to make a better forecast.

The prototype needs to be established in the company and start to operate with the customers. The prototype should save at the beginning the historical data to be able to make a review of the data mining outputs. The accuracy of the data mining evaluations has to be evaluated. With data mining, selecting the input data is a really important step. In this prototype all the data from the user is evaluated by data mining; the obtained results are maybe not so accurate because of that. Only some variables should be evaluated by data mining. But this should be applied once the prototype is working; each company is different and uses different data.

Some papers were found comparing data mining and control charts [20]. Due to that paper, the prototype also provides statistical information about the forecasts. Using simple statistics, the prototype lets the users compare statistical data and data mining outputs.

Is a good combination combining data mining and Monte Carlo? They can be combined in a really useful way. As was explained before, Monte Carlo can be useful because it will not see like a dangerous activity to increase the value of one variable. Maybe a decision-maker will be afraid of that, because it is a key value and increase the money of that is a dangerous activity for the company. Also with Monte Carlo less time is needed to take all the values of the variables (in this prototype more or less two hundred variables are used). When the simulation is done, a validation is needed to know if it was a good simulation. Data mining is required for that step. Data mining will help to evaluate is the simulation was good, accurate or trustful. Decision-makers can use data mining results to obtain maybe unexpected rules or decision trees that can be useful for future predictions.
7 CONCLUSIONS

After this thesis was done, we can conclude that data mining seems to be a really useful and important technique nowadays and we think that it will be even more used in the future, because the companies will store more and more data. It is possible to obtain large amount of information of your stored data. That information can really help to predict the future. Monte Carlo is a good way to generate random data. Using the generated data it is possible to create a model of your interests. Then Monte Carlo and data mining related together can try to predict with accuracy. Using the results from the thesis, with the validation of the correlation coefficient, some rules and tree can be trusted and used to predict with more accuracy.

One of the questions that were addressed in this report is how data mining can be useful as the basis for a decision support tool in port planning. Before this prototype was developed, the decision-makers had to make forecasts using a simple spreadsheet. When making those forecasts, the predictions were made by just increasing the value of the previous year by a certain percentage. This type of data gathering takes a lot of time, since all the variables have to be known. This will be a tough task when talking about a scenario with approximately one thousand variables. That is why data mining is a useful way to find patterns and models in gathered data. With reference to patterns and models, data mining gives us rules, decision trees, linear regressions, etc. supported by statistics that helps to make more accurate and reliable decisions.

The amount of time that would be saved using data mining will reflect as cost saving to the company. Data mining can help to give much accurate predictions, because a better prediction can be done is decision-makers have more information about the future, this is what data mining algorithms provide, estimations and predictions of values.

The other question was how we can evaluate and analyze data generated via Monte Carlo simulation in testing data mining techniques. After obtaining the results of the simulation we can know the best case and the corresponding inputs. All the inputs are evaluated using data mining in the same way, even if the data is given by the user. Just the process of obtaining the input data is different. We evaluate these results by following the seven algorithms explained in section 4.7. To evaluate them we provide the data mining outputs. Evaluating all of these outcomes, the user of the prototype will be the one who can decide if the Monte Carlo simulation was accurate enough. Also, we give some statistical results as graphs and tables, which the user can use to analyze the results.

We think this prototype is the first step in developing a really helpful application not only in terms of cost saving, but also in helping decision-makers to simplify their daily tasks. We hope this prototype could become a much advanced tool with more functionality.

We think this thesis was really important and useful for us because it was our first step in these topics. Data mining seems to be one of the most important topics of the future.
8  **FUTURE WORK**

Some of the points that could be improve in different areas:

*Clustering and scalability*

Clustering is becoming a trend nowadays, so for a future work it would be recommended to move the system to into a cluster computing system, in order to be able to make bigger calculus that otherwise takes a lot of time using only one computer. Upgrading the system as mentioned, it is possible to have a large amount of users at the same time simulating forecasts. The system will also be scalable, which is one of the most desirable properties that a system should have.

*Dashboards report online*

This technique is definitely a great chance to review the XLS files with dynamic data, graphics, and the best thing is that it is not needed to have any software installed on the computer clients, just a web browser, so it is an ideal way to share information and show all the data easy, clear and fast.

It could be done using some of the programs mentioned before [18, 19, 30], as well as using one of the new features of Microsoft Office Excel 2007. However, none of them are open source; it is needed to get the license to use them so that means to raise the price of the tool.

*Visual Weka*

Next step is parsing the output files to represent the data with graphics or any kind of diagram such as decision trees, histograms, etc.

*Continue studying*

There are some input non-random variables that will be interesting to study to see the behavior of the system; it would be a good idea try to play with all those inputs and change to random inputs but always trying to simulate realistic values. That means find good random numbers to use in the simulation.
We would like to thank Dr. Lawrence E. Henesey (Lefteris) most sincerely for being unique. We are also grateful to our beloved families (special dedication to the memory of those who have left us) and closer friends; more especially to Beatriz and Sandra because they have shared our suffering. Thank to the people who have lived in Minerva for the unconditional support and an unforgettable year. The last and not less important; thanks to Piraten and all the institutions that made this possible.
REFERENCES


[38] Monte Carlo simulation with excel. E-math project. Supported by the Secretaria de Estado de Educacion y Universidades (MECD) UOC (www.uoc.edu)


