ISSN 1870-4069

Development of a Framework for the Use of a Tool for Machine Learning and Data Mining

Jacqueline Ramos Landeros, Ma. De Lourdes Margain Fuentes

Ingeniería en Sistemas Estratégicos de Información, Universidad Politécnica de Aguascalientes, Aguascalientes, Mexico

{up130142, lourdes.margain}@upa.edu.mx

Abstract. The "Universidad Politécnica de Aguascalientes" is an institution who's interested in quality education, ensuring the control of the subjects that are imparted in a long term in the mayor. Taking this on account, the design and development of a Moodle platform focused on data mining was planted, this subject is currently given to ninth quarter students. In order to re-inforce the knowledge that the course has to offer, it was decided to utilize as a support tool "Weka", it is classified as an "intelligent software". It provides a sustenance in different areas, such as, Marketing, Manufacturing, Health, Energy, Finance, Medicine, inter alia, for its application it must be taken in account the type of assignment it is wished to realize. The purpose of this course is for the student to obtain the necessary knowledge in this field and during this process the material developed can be of use as a tool for this.

Keywords: Data Mining, Automatic Leaning, Business Intelligence, Knowledge Discovery in Databases, Software Weka, Software Engineering.

1 Introduction

At present the task to improve the Access to information that is given to us by the companies is gaining more strength, especially in modern business, where process based in the recourse of information extraction is mainly required, whereby because of the huge workplace and the amount of information, that is called Data Base, is necessary to count on new methods of data processing, new technologies that facilitate the process of search and extraction of knowledge to service companies in the taking of decisions that benefit its performance.

A way to achieve said results are situations or states in which an enterprise pretends to achieve data mining, it is of great importance in the working world, because it allows the obtaining of knowledge based on the data that are found stored, through the process of implicit information extraction, previously known and potentially useful.

The objective of the course is to facilitate the support material for both students and teachers in the learning of the data mining subject, a compilation of a varied set of information, in combination with one of its most famous tools (Weka).

pp. 127–139; rec. 2016-08-11; acc. 2016-10-08

2 Design Foundations

In order to model and develop the course a Learning Management System (LMS) was used, it provides a backing for both educators and pupils in the visual teaching and learning process also it is used to create, endorse, administrate, store, distribute and manage the activities in a visual way.

The task to create the contents for the courses is developed using a Learning Content Management System (LCMS).

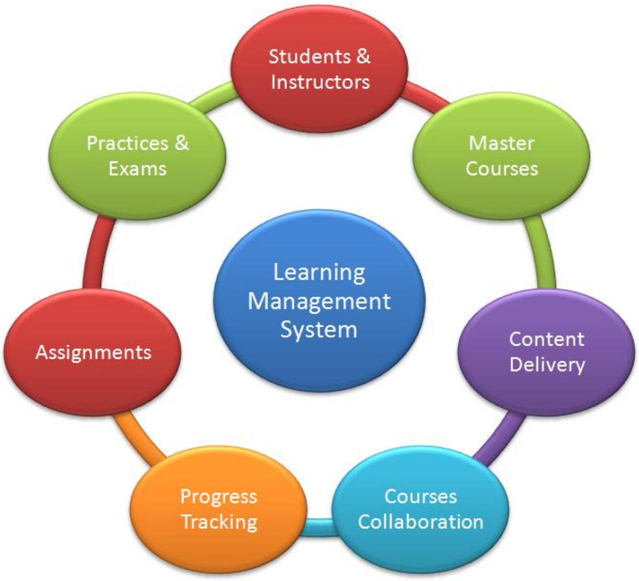


Fig. 1. Learning Management System (LMS) Model.

Students that learn on their own and at their rhythm are alone and are completely independent, while the e-learning courses facilitated or directed by an instructor offer different levels of support from tutors, instructors and collaboration between scholars. The e-learning courses often employ both approaches, but in order to be brief and practical, it is easier to examine them in separated ways.

2.1 Moodle Platform

The use of the e-learning platform of Moodle compared with other systems implies that it is made based on the constructivist social pedagogy, where communication has a relevant space in the way of knowledge construction. Being the objective of generating an enriching learning experience [1].

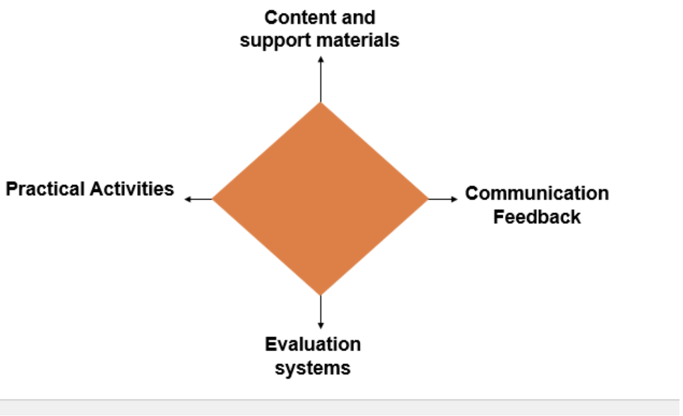


Fig. 2. Moodle work system.

The philosophy planned by Moodle includes a constructive approximation based on the social environment of the education, emphasizing that students (and not only teachers) may contribute in the educative experience in many methods.

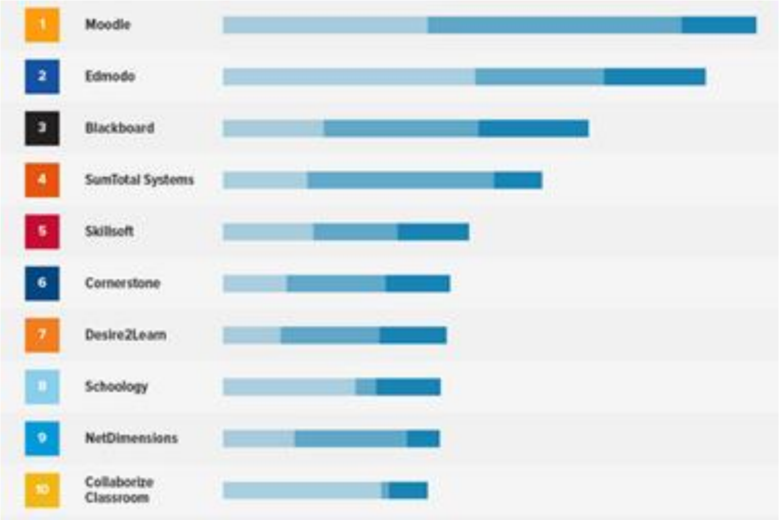


Fig. 3. Comparison between the most used LMS platforms [2].

It can be seen in the previous image how the Moodle platform is the most used among the different systems that may be found nowadays, due to its stability and intuitive interface that permits the user to develop the learning process successfully.

2.2 Methodology

The model that was used as a methodology for the development of the course was the ADDIE model. Is a framework that lists generic processes that instructional designers and training developers use [2]. It represents a descriptive guideline for building effective training and performance support tools in five phases (Fig. 4).

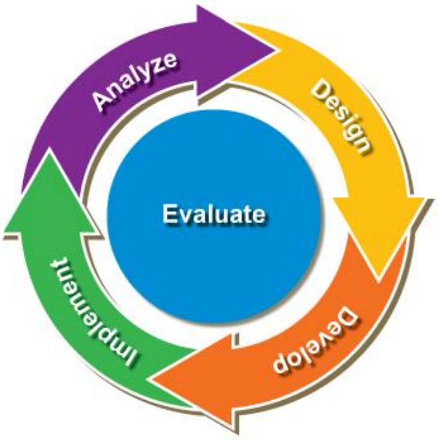


Fig. 4. ADDIE Model [3].

Analysis phase: The analysis phase clarifies the instructional problems and objectives, and identifies the learning environments, learner´s existing knowledge and skills.

Dsign phase: The design phase deals with learning objectives, assessment instruments, excercises, content, subject matter analysis, lesson planning, and media selection.

Development phase: In the devolpment phase, instructional designers and developers crate and assemble content assets blueprint in the design phase.

Implementation phase: The implementation phase develops procedures for training facilitators an learners.

Evaluation phase: The evaluation phase consists of two aspects: formative and summative, the first one is present in each stage of the ADDIE process, while summative evaluation is conducted on finished instructional programs or products.

3 Organization and Components of the Course

The organization of the course twirls around of sections that divide the contents and activities based on its main function, impart the student the background necessary to the apprenticeship in the Data Mining field. It is observed in Fig. 4 the course structure.

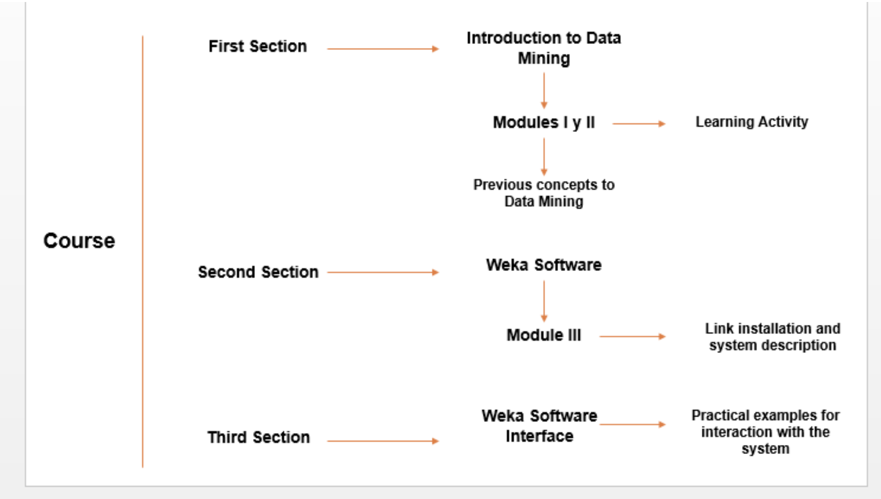

Course
First Section
Introduction to Data Mining
Modules I y II
Learning Activity
Previous concepts to Data Mining
Second Section
Weka Software
Module III
Link installation and system description
Third Section
Weka Software Interface
Practical examples for interaction with the system

Fig. 5. Course Structure.

3.1 Description of the Sections of the Course

First Section:

The first section was established specially for the student to comprehend the background concepts of Data Mining and therefore continue with the learning process, which was made in two modules, which are detailed next:

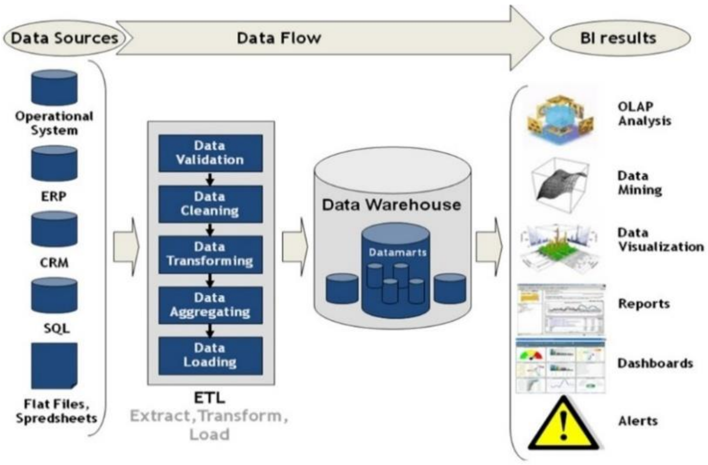


Fig. 6. BI Scheme [5].

Module I:

The first step to Data Mining is to comprehend the business, this means to determine the objectives in order to decide the goals of the Data Mining thus the Business Intelligent subject was investigated, and as a result BI scheme, it is an essential process that is obtained as a result of the sources and data flows to achieve the process of Data Mining.

Business Intelligence:

Business Intelligence is the ability to transform data in information, and information in knowledge, in a way that it can improve the decision making in business. [5]

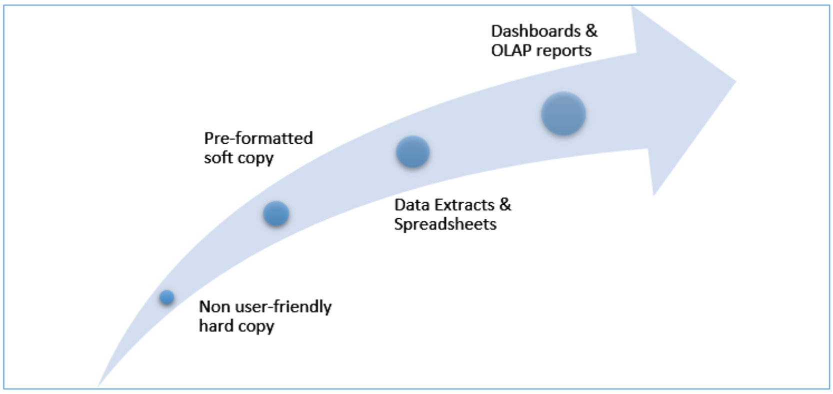


Fig. 7. The Evolution of Business Intelligence [5].

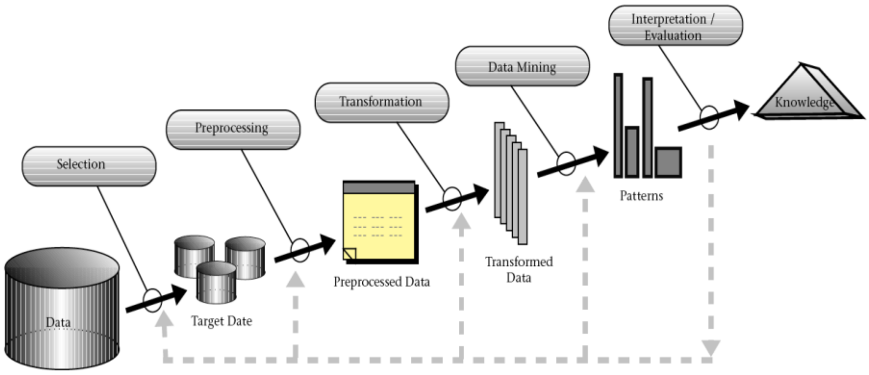


Fig. 8. KDD process.

Knowledge Discovery in Databases (KDD):

With the necessity of being able to handle large quantities of data, a study field arises, named Knowledge Discovery in Databases (KDD).

It refers to “the non-trivial process of discover knowledge and information potentially useful in the contained data inside an Information Warehouse. It is not an automatic process, it is an iterative process that explores thoroughly extremely large volumes of data to determine relations. It is a process that consist in using methods of Data Mining (algorithms) to extract (identify) what is considered as knowledge” [5]

Data Mining:

In simple terms it is about a data exploitation method and information extraction that transforms it into useful knowledge to help the decision making in an organization through the determination of patterns and models. It arises to comprehend the content in a Data Warehouse.

Module II

The second module is deeply detailed about the preparation of the data from its output up until its integration.

Output and Data Set Description

This topic is produced by the phase of data preparation, that will be used in order to model or the main Project analysis.

Describing the data set that will be used in the modeling and the main project analysis.

Data Selection and Recollection

Once the Data of interest are collected, an explorer (in this case the user) may decide what type of pattern will be found. The kind of knowledge that is desired to extract will marc clearly the technic of Data Mining used to decide which data will be used for the analysis.

Data Cleaning

It is a fundamental process in the data migration, its main objective is the quality of data obtained at the end of migration. Data Cleaning is particularly important when data come from heterogeneous sources (different sources) that may not share the same scheme of data or can´t represent the same real entity in different ways.

The process of cleaning constitutes great part of the estate of transformation of the data during Data Warehouse construction.

This duty includes the construction of data preparation operations such as the production of derivate attributes or the entry of new registers, or the transformation of values for existent attributes.

Outputs Derived Attributes

The derived attributes are new attributes that are built from one or more existing attributes in the same registry.
To clarify what’s said before, example: area=length*width

Generated Registry

In this step the newly generated registry are described.
Example: Create registry for women that bought lipstick color brown, thus there is no reason to obtain such registry in the raw data, but in order to model this might have sense to represent explicitly the fact that certain women have not made a purchase.

Data Integration

This is the method by which the information is combined in multiple tables or files to create new archive or values.

Data Format

Reformatting information refers to modifications made to the data that don`t change its meaning, but may be required by the modeling tool.

Second Section (Module III)

Software Weka:

On the third module an explanation in the functionality of the Weka software is developed, nowadays different tools exist for the visualization and algorithms for the analysis of data joined in a user graphical interface to easily access its functions making use of the processing of information manifested in Data Bases, but after making said investigation, a conclusion was made that this tool is the most reliable in the market and at the same time it is at full disposal for the user.

“Weka is a software tool for the automatic learning and Data Mining designed on base Java and developed at the university of Waikato in New Zealand in 1993, this tool with the acronym (Waikato Environment for Knowledge Analysis) provides license distribution GNU-GLP or free software” [5].

Application Fields

Weka has a group of technics that can be applied with success to multiple fields, such as Marketing, Manufacturing, Health, Energy, Finance, Medicine, among other, for its application it must be taken in account the king of task is being planned to realize.

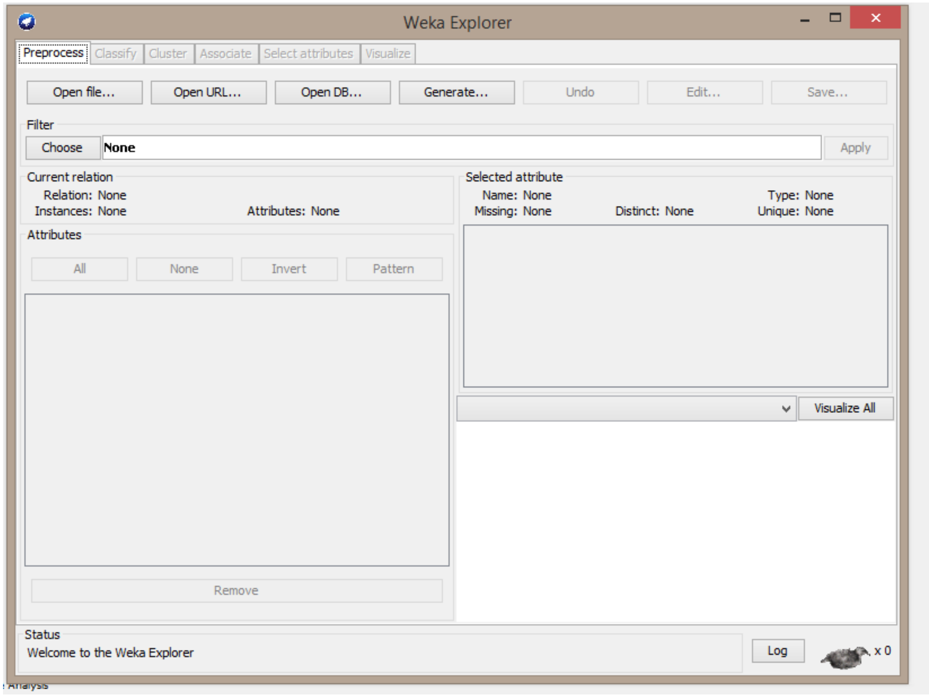
- Classification methods based on neuronal nets.
- Numeric methods manipulation over data (financial statistics).

- Classification methods based on vector support machines.
- Meta classifiers.
- Methods of decision trees implementation.
- Probability estimation methods.

Third Section (Software Interphase)

In this section a paragraph for the installation of the software is included. The sections of the software are:

- *Pre-process:* Here are included all the tools and filters to charge and manipulate the data that are going to be used.
- *Cluster:* Here various methods of grouping are integrated.
- *Associate:* includes the technics used for the association rules.
- *Select Attributes:* It contains the access options to the different technics used to reduce the amount or number of attributes.
- *Visualize:* This section allows to do the studies of behavior, using the visualization technics included in Weka.



Filter: None

Relation: None
Instances: None
Attributes: None

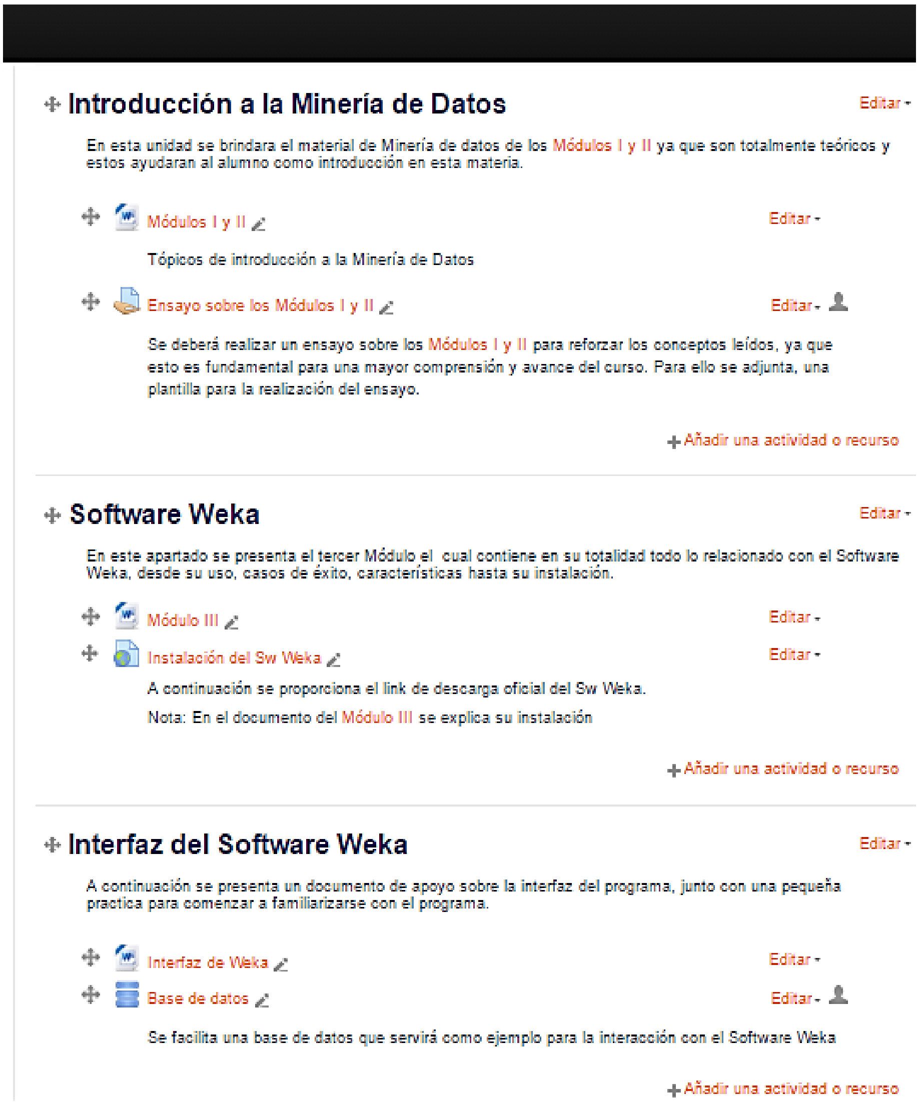
Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

Fig. 9. Weka Software Interphase.

4 Results

To implement the course on the Moodle LMS platform the following results are obtained.



Introducción a la Minería de Datos

En esta unidad se brindara el material de Minería de datos de los Módulos I y II ya que son totalmente teóricos y estos ayudaran al alumno como introducción en esta materia.

Módulos I y II

Tópicos de introducción a la Minería de Datos

Ensayo sobre los Módulos I y II

Se deberá realizar un ensayo sobre los Módulos I y II para reforzar los conceptos leídos, ya que esto es fundamental para una mayor comprensión y avance del curso. Para ello se adjunta, una plantilla para la realización del ensayo.

Software Weka

En este apartado se presenta el tercer Módulo el cual contiene en su totalidad todo lo relacionado con el Software Weka, desde su uso, casos de éxito, características hasta su instalación.

Módulo III

Instalación del Sw Weka

A continuación se proporciona el link de descarga oficial del Sw Weka.

Nota: En el documento del Módulo III se explica su instalación

Interfaz del Software Weka

A continuación se presenta un documento de apoyo sobre la interfaz del programa, junto con una pequeña practica para comenzar a familiarizarse con el programa.

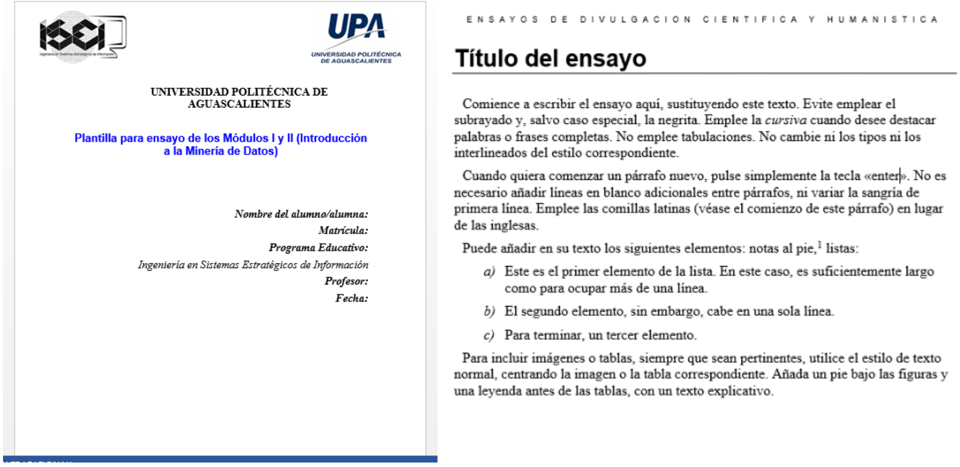
Interfaz de Weka

Base de datos

Se facilita una base de datos que servirá como ejemplo para la interacción con el Software Weka

Fig. 10. Course implemented in the Moodle platform.

4.1 Exercises and Material



UNIVERSIDAD POLITÉCNICA DE
AGUASCALIENTES

Plantilla para ensayo de los Módulos I y II (Introducción a la Minería de Datos)

Nombre del alumno/alumna:
Matrícula:
Programa Educativo:
Ingeniería en Sistemas Estratégicos de Información
Profesor:
Fecha:

ENSAYOS DE DIVULGACION CIENTIFICA Y HUMANISTICA

Título del ensayo

Comience a escribir el ensayo aquí, sustituyendo este texto. Evite emplear el subrayado y, salvo caso especial, la negrita. Emplee la *cursiva* cuando desee destacar palabras o frases completas. No emplee tabulaciones. No cambie ni los tipos ni los interlineados del estilo correspondiente.

Cuando quiera comenzar un párrafo nuevo, pulse simplemente la tecla «enter». No es necesario añadir líneas en blanco adicionales entre párrafos, ni variar la sangría de primera línea. Emplee las comillas latinas (véase el comienzo de este párrafo) en lugar de las inglesas.

Puede añadir en su texto los siguientes elementos: notas al pie,[1] listas:

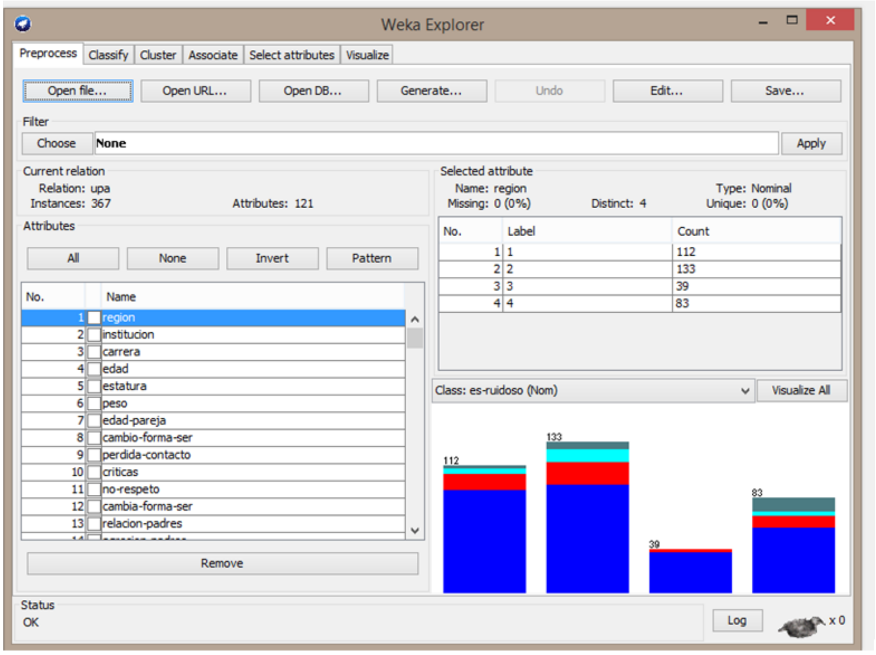
a) Este es el primer elemento de la lista. En este caso, es suficientemente largo como para ocupar más de una línea.

b) El segundo elemento, sin embargo, cabe en una sola línea.

c) Para terminar, un tercer elemento.

Para incluir imágenes o tablas, siempre que sean pertinentes, utilice el estilo de texto normal, centrando la imagen o la tabla correspondiente. Añada un pie bajo las figuras y una leyenda antes de las tablas, con un texto explicativo.

Fig. 11. Essay template.



None

Current relation
Relation: upa
Instances: 367
Attributes: 121

Selected attribute
Name: region
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	1	112
2	2	133
3	3	39
4	4	83

Attributes

No.	Name
1	region
2	institucion
3	carrera
4	edad
5	estatura
6	peso
7	edad-pareja
8	cambio-forma-ser
9	perdida-contacto
10	criticas
11	no-respeto
12	cambia-forma-ser
13	relacion-padres

Class: es-ruidoso (Nom)

Status
OK

Fig. 12 Database for practice.

5 Conclusions

At the end of the realization of this material, it has been concluded that the main core of it is that the students become capable of develop pan autodidact way of learning, because in college is of great importance the search of knowledge trough the own ways of the autodidact learning. In order to achieve this it is necessary to count on the didactic material that facilitates the search of the subject in which it is going to be worked.

The methodology exposed in this material is focused in providing a line of solution in the definition of advanced terms in the Data Mining field, so that through its knowledge and application using different tools that this branch has to offer it is desired to achieve an objective which is knowledge extraction.

The information of the course has the objective of providing assistance as a tool of support in future works of the students, whereby it will be necessary to improve the content as the information is upgraded.

Acknowledgements. The development of this course would not have been possible without the help from the Dr. Ma. De Lourdes Margain Fuentes (Academic Program Director).

References

1. Sancho, J. B.: La Plataforma Educativa Moodle. Available at: http://www.fvet.uba.ar/postgrado/Moodle18_Manual_Prof_1.pdf (2007)
2. Puzziferro, M., Shelton, K.: A model for developing high-quality online courses: Integrating a systems approach with learning theory. Journal of Asynchronous Learning Networks, pp. 12, 119–136 (2008)
3. Wang, S. K., Hsu, H. Y.: Using the ADDIE model to design Second Life activities for online learners. TechTrends, 53(6), pp. 76–81 (2009)
4. Dillenbourg, P.: Collaborative learning: Cognitive and computational approaches. Advances in learning and instruction series. Elsevier Science, Inc., PO Box 945, Madison Square Station, New York, NY 10160-0757 (1999)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp. 10–18 (2009)
6. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine, 17(3), pp. 37 (1996)
7. Liu, H., Motoda, H.: Feature extraction, construction and selection: A data mining perspective. Springer Science & Business Media, 453 (1998)
8. Black, E. W., Dawson, K., Priem, J.: Data for free: Using LMS activity logs to measure community in online courses. The Internet and Higher Education, 11(2), pp. 65–70 (2008)
9. Macfadyen, L. P., Dawson, S.: Mining LMS data to develop an early warning system for educators: A proof of concept. Computers & education, 54(2), pp. 588–599 (2010)
10. Davies, J., Graff, M.: Performance in e-learning: online participation and student grades. British Journal of Educational Technology, 36(4), pp. 657–663 (2005)
11. Nagel, L., Kotzé, T. G.: Supersizing e-learning: What a CoI survey reveals about teaching presence in a large online class. The Internet and Higher Education, 13(1), pp. 45–51 (2010)

12. Davies, J., Graff, M.: Performance in e-learning: online participation and student grades. British Journal of Educational Technology, 36(4), pp. 657–663 (2005)
13. Romero, C., Ventura, S., García, E.: Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), pp. 368–384 (2008)
14. Macfadyen, L. P., Dawson, S.: Mining LMS data to develop an early warning system for educators: A proof of concept. Computers & education, 54(2), pp. 588–599 (2010)