UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Facultad 1

Departamento de Ciencias Básicas

MODELO PARA EL ANÁLISIS DE DATOS DE PROCESOS DE FORMACIÓN DEL PROFESIONAL

Trabajo final presentado en opción al título de

Máster en Informática Aplicada

Autor: LIC. JOSÉ HILARIO QUINTANA ÁLVAREZ

Tutores: Dr C. PEDRO YOBANIS PIÑERO PÉREZ

Dr C. OLGA LIDIA MARTÍNEZ LEYET

Ciudad de La Habana, Julio del 2011

Agradecimientos

Agradezco a la Revolución por brindarme la oportunidad de formarme y de ser útil a la Patria

A Marisela Lastra por su Amor y apoyo constante.

Agradezco también a mis tutores, el doctor Pedro Piñero Pérez por su apoyo constante y la doctora Olga Lídia Martínez Leyet por sus sabios consejos

Agradezco a todas mis compañeras y compañeros de la Universidad de las Ciencias Informáticas y a los profesores que me instruyeron y educaron.

DECLARACIÓN JURADA DE AUTORÍA Y AGRADECIMIENTOS

Declaro por este medio que yo: José Hilario Quintana Álvarez, con carné de identidad 61052402085, soy el autor principal del trabajo final de maestría MODELO PARA EL ANÁLISIS DE DATOS DE PROCESOS DE FORMACIÓN DEL PROFESIONAL, desarrollada como parte de la Maestría en Informática Aplicada y que autorizo a la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo.

Y para que así conste, firmo la presente declaración jurada de autoría en Ciudad de La Habana a los 8 días del mes de Septiembre del año 2011.

José Hilario Quintana Álvarez

Resumen

En esta investigación se propuso un modelo para el análisis de los datos educacionales de procesos de formación del profesional, que se compone de las informaciones resultantes del tratamiento de conjuntos de datos provenientes del curso regular diurno de la carrera Ingeniería en Ciencias Informáticas, mediante técnicas de minería de datos y la estadística inferencial.

La propuesta integró datos de diversas fuentes y momentos del curso, con formatos numéricos o nominales que fueron combinados con un enfoque integrador atendiendo a los objetivos de condensar la información, incluir indicadores relativos a los conocimientos, hábitos y valores, revelar patrones no triviales y facilitar la toma de decisiones de los decisores del proceso docente educativo. A modo de validación, se generó un modelo con información diagnóstica y resultados de evaluaciones parciales reales de los estudiantes del primer año de la carrera de ingeniería en ciencias informáticas de la facultad 1 de la Universidad de las Ciencias Informáticas del curso 2009-2010 con vistas a lograr la predicción del avance académico que manifiestan entre el primer y segundo corte evaluativo. Se comparó el modelo propuesto con la forma tradicional de realizar el análisis de los datos educacionales. Se ofrecieron valoraciones relacionadas con las ventajas que ofrece este modo de proceder con respecto a la eficiencia del procesamiento, las cualidades para la clasificación, el pronóstico y la socialización de reportes.

Palabras claves: modelo, minería de datos, procesamiento, toma de decisiones

Abstract

This research suggested a model for the analysis of educational data of process of professional formation composed by the integration of the information's resulting of the treatment of data sets from daytime regular course. The proposal integrated data from various sources and moments of the course, with numeric or nominal formats which were combined in a comprehensive approach, with the aims of condensing information; including indicators related to the skills, habits and values; revealing non-trivial patterns and facilitating the decision-making of the educational teaching process decision-makers. In order to validate the proposal, a model was generated, with diagnostic information and results of actual partial assessments of the first year students in School Number 1, at the University of Informatics Sciences, in the academic year 2009-2010, in order to fulfill the prediction of the academic progress these students manifested in the period between the first and second non final evaluative reports. The model suggested is compared with the traditional way to proceed with the analysis of educational data. Besides, assessments related to the advantages of this way of proceeding with regard to the efficiency of the processing, the qualities for the classification, prospects and the socialization of reports are also offered.

Keywords: model, data mining, processing, decision-making.

ÍNDICE

INTRODUCCIÓN

La formación del profesional es el proceso en el que los sujetos desarrollan el compromiso social profesional, la flexibilidad ante la cultura, la trascendencia en su contexto, toda vez que elevan su capacidad para la reflexión divergente y creativa, para la evaluación crítica y autocrítica, para solucionar problemas, tomar decisiones y adaptarse flexiblemente a un mundo cambiante. (1)

El artículo 3 de la resolución 210 del 2007 que norma el trabajo metodológico en el Ministerio de Educación Superior de Cuba (2) establece que el modelo de formación de la educación superior cubana es de perfil amplio y se sustenta en dos ideas rectoras fundamentales:

- La unidad entre la educación y la instrucción, que expresa la necesidad de educar al hombre a la vez que se instruye.
- El vínculo del estudio con el trabajo, que consiste en asegurar desde el currículo el dominio de los modos de actuación del profesional, en vínculo directo con su actividad profesional.

Desde el punto de vista de la planeación estratégica, la Universidad de las Ciencias Informáticas (UCI) tiene como misión convertirse en una: "Universidad innovadora de excelencia científica, académica y productiva, que forma de manera continua profesionales integrales comprometidos con la patria, soporte de la informatización del país y la competitividad internacional de la industria cubana del software". (3)

El objetivo general es: " Formar ingenieros en Ciencias Informáticas con sólidas competencias sustentadas en una concepción científica y dialéctico-materialista del mundo, que estén comprometidos con su patria y que actúen como profesionales responsables, honestos, honrados, creativos, modestos, solidarios y con ética revolucionaria en el campo de la Informática. Tendrán fuerte espíritu crítico, autocrítico y de auto superación durante toda la vida.

Serán capaces de aplicar con ética conocimientos económicos, estéticos, de protección al medio ambiente y de seguridad informática para contribuir al desarrollo socio-económico y a la defensa de la sociedad socialista cubana. Estarán preparados para asumir la docencia en cualquiera de los temas relacionados con la Informática en cualquiera de los niveles educacionales del país. Estarán preparados para, mediante su integración en equipos como miembro o como líder, participar de forma decisiva en los diferentes planes para la informatización de la sociedad cubana, siendo además, portadores y promotores de una cultura general integral" (4)

En la documentación elaborada por el Colectivo de Carrera de la UCI, con vistas a la implementación del plan C modificado, se concibe "una integración más completa de las asignaturas" y "se refuerza la creación de habilidades de trabajo en grupos". (3)

Con tales premisas, desde el curso 2009-2010 comenzaron a realizarse en la Universidad, un grupo de transformaciones que permitieran lograr una mayor integración de los procesos formativos, productivos e investigativos.

En ese sentido se estructuró el proceso formativo en un ciclo básico para los tres primeros años de la carrera y un ciclo profesional para los dos restantes años.

Para potenciar el aprendizaje desarrollador en el proceso formativo, era necesario manejar más información educacional de los estudiantes para mejorar la toma de decisiones. No bastaba con los reportes de evaluaciones y asistencias utilizados hasta el momento.

Previo al inicio del curso 2009-2010 y en el marco del proyecto "Modelo de integración de la Formación, Producción e Investigación (MIFPI)", elaborado por el Centro de Innovación de la Calidad Educativa (CICE) de la UCI, se diseñó la estrategia de caracterización integral de los estudiantes que ingresaban en los ciclos básico y profesional.

Un heterogéneo grupo de profesores de experiencia, dirigentes, investigadores, psicólogos, sociólogos e informáticos, dirigidos por los doctores José Lavandero García y Olga Lidia Martínez Leyet diseñó los instrumentos de evaluación diagnóstica y de selección de las herramientas apropiadas para ejecutar la tarea.

Se deseaba caracterizar a cada estudiante en cuanto a los atributos sociales, motivacionales, del desarrollo cognitivo y de competencias. Tras el diseño de las baterías de encuestas, se utilizaron herramienta avanzadas para la aplicación de los diagnósticos, su procesamiento y socialización.

En la actualidad, el CICE ha definido la realización de investigaciones sistemáticas para evaluar la calidad de los procesos formativos durante todo el curso. Los diagnósticos a los estudiantes que inician los ciclos básicos y profesional tienen un importante peso en estos estudios.

La doctora Olga Lidia Martínez Leyet, directora del CICE, en informe al Consejo Universitario acerca del diseño y aplicación, de la estrategia de caracterización de los estudiantes en la UCI (5) , resaltó los objetivos a seguir, la necesidad de la sistematización de la actividad diagnóstica, cuales atributos a monitorear y cómo utilizar los resultados del diagnóstico. En cuanto a los resultados de la aplicación del diagnóstico, mostró caracterizaciones de dos estudiantes de la universidad, uno de primer año y el otro de cuarto año, La caracterización de la facultad 2 a través de las estadísticas de sus 224 estudiantes de primer año en cuanto a los atributos de interés.

La directora del CICE identificó las siguientes debilidades del proceso de caracterización:

- Dificultades en el acceso generalizado a los resultados, lo que no ha favorecido su utilización en el trabajo cotidiano de los profesores, tutores y profesores guías.
- Es insuficiente aún la preparación que deben adquirir todos los implicados para poder hacer un uso efectivo de los resultados en el diseño de estrategias de intervención pedagógica.

En cuanto al monitoreo del proceso de caracterización a los estudiantes de primer año, en el 2009, señalaba lo siguiente:

- Solo un 5% de los encuestados ha reconocido que conocen y están utilizando los resultados del diagnóstico.
- No se están utilizando estos resultados como premisa para los proyectos educativos en las brigadas, en 4to año es un trabajo que aún está en fase de proyección, no hay todavía resultados concretos.
- No se están utilizando sus resultados todavía por todos los implicados, en el proceso de rediseño didáctico de las asignaturas.

El autor de esta investigación identifica como un problema, que la información resultante haya quedado dispersa en varios reportes (caracterización de cada uno de los estudiantes, caracterización de la facultad, caracterización de la Universidad) por lo que se dificultó el análisis de los datos.

Del análisis del informe se concluye que la integración de la información diagnóstica recabada con el resto de los datos que emanan constantemente del proceso docente fue insuficiente, por lo que existe un problema con la gestión de la información educacional que debe influir grandemente en la calidad de la toma de decisiones.

Se precisa cruzar la información diagnóstica inicial con la que continuamente genere el proceso docente, para poder dar seguimiento al aprendizaje de los estudiantes, de ahí que existe la necesidad de ampliar el modelo de gestión de la información educacional, heredado del accionar práctico de universidades que contribuyeron a la fundación de la UCI.

Una contribución a la solución a la problemática mencionada está en el mejoramiento de los modelos para el análisis de datos de los procesos. Usualmente, los datos recogidos se almacenan en tablas, que son resumidas mediante estadísticas descriptivas con el fin de permitir su interpretación y socialización. La información pasa bruscamente de un nivel particular a uno general que impide visualizar las fortalezas y debilidades de los tipos de estudiantes que tenemos.

Aunque la estadística descriptiva juega un importante papel en la elaboración de modelos de gestión, debe tenerse en cuenta que:

- Destruye la información particular en aras de la general.
- La aplicación de técnicas de tipo univariadas a datos multivariados[1] omite la posible relación entre variables.
- Es difícil extraer de sus resultados estrategias de trabajo individualizadas para cada tipo de personalidad a pesar de la necesidad de hacerlo en el enfoque histórico cultural.

[1] Datos con múltiples variables.

- Aunque reducen el formato de la información, es difícil la interpretación de datos ubicados en tablas diferentes.

Ante las limitaciones del uso casi absoluto de la estadística descriptiva se impone la aplicación de otros métodos que faciliten el descubrimiento de información en los datos disponibles y que posibiliten realizar investigaciones potencialmente útiles para la toma de decisiones.

Existe una rama de la Inteligencia artificial especializada en la búsqueda de información no trivial encerrada en grandes bases de datos, la denominada Minería de datos. ¿Pudieran utilizarse sus técnicas y procedimientos en la modelación de datos para el análisis de procesos?¿Que cambios se habrán producido a nivel internacional en la generación de modelos para el análisis de datos educacionales?

Una pionera del procesamiento de la información educacional, la investigadora francesa Agathe Merceron señaló hace poco: "Hasta donde sepamos, existen pocas herramientas dedicadas a encontrar información pedagógicamente relevante sobre el trabajo del estudiante. Muchas herramientas únicamente producen estadísticas sobre notas, buenas o malas respuestas o errores"(6). Con tales palabras se puede constatar la existencia de la problemática en cuanto al diseño de este tipo de modelos y la preocupación por solucionarla.

Desde el año 1995 en adelante, se realizan numerosas investigaciones en el mundo, con el empleo de herramientas estadísticas y de visualización por una parte y de la minería web (7) por el otro, en el marco del boom de los sistemas de enseñanza a distancia.

Aun así, en muchos de los resultados alcanzados por la naciente comunidad de investigadores de minería de datos educacionales se sobredimensiona el uso de variables de tipo conductistas, lo que limita el alcance de esas investigaciones, según los fundamentos de la pedagogía cubana.

En esta investigación se busca contribuir a la solución del siguiente **Problema Científico:** ¿Cómo integrar las técnicas modernas de minería de datos y la estadística inferencial al modelo de análisis de datos tradicional, para mejorar la toma de decisiones [2] en el proceso de formación del profesional durante el ciclo básico en la Ingeniería en Ciencias Informáticas?

El **Objeto de la investigación** es el Proceso de Minería de Datos.

El **Campo de la investigación** es el modelo para analizar información del proceso de formación del profesional.

El **Objetivo general de la investigación** consiste en elaborar un modelo para el análisis de datos educacionales del proceso de formación mediante la utilización de técnicas de minería de datos y la estadística inferencia, para mejorar la toma de decisiones.

---

[2] Toma de decisiones: proceso mediante el cual se realiza una elección entre varias alternativas o formas para resolver un problema actual o potencial, para ello, se requiere conocer, comprender y analizar pues una mala o buena elección suele tener consecuencias.

Y como **objetivos específicos** los siguientes**:**

1.Valorar el estado del arte relacionado con la aplicación de la minería de datos y la estadística inferencial en la modelización de datos educacionales.

2.Seleccionar los requisitos idóneos para la reelaboración del modelo de análisis de datos educacionales actual, que permita resolver la situación problémica.

3.Elaborar un modelo de análisis de datos que integre la estadística descriptiva del modelo tradicional con las técnicas de minería de datos y de la estadística inferencial.

4.Validar la eficiencia del nuevo modelo para el análisis de datos, cuando se procesa información educativa real del primer año de la facultad 1 de la UCI.

Para darle solución al problema científico señalado anteriormente, se plantea como **Hipótesis de Investigación** la siguiente**:**

La aplicación de un modelo para el análisis de datos de procesos de formación del profesional que incluya técnicas de minería de datos educacionales favorecerá el proceso de toma de decisiones pedagógicas.

Población y muestra

Para evaluar la validez de la propuesta se seleccionó como población al colectivo de estudiantes, de primer a tercer año, de la Facultad 1 de la UCI y como muestra intencionada, a las cinco brigadas de primer año que comenzarían a transitar por el nuevo plan de estudios correspondiente a la cohorte 2009-2010.

Se decidió trabajar con la información de una facultad de la universidad, por constituir una unidad en la que se concentran los procesos fundamentales de la gestión del conocimiento docente.

La población consta de 498 estudiantes del ciclo básico de la Facultad 1, distribuidos en cinco brigadas de primer año, siete brigadas de segundo año y seis brigadas de tercer año.

Se tomó como muestra a los 140 matriculados en primer año lo que conforma un 28,11% de la población.

En el desarrollo de la investigación se utilizaron métodos científicos para la obtención, procesamiento y arribo a conclusiones, a continuación se detallan los más significativos:

Métodos teóricos:

- Análisis Histórico- Lógico: La modelación de procesos ha experimentado notables avances que debemos incorporar dialécticamente al proceso docente educativo.
- Análisis y síntesis: Para estudiar las principales técnicas de modelación y procesamiento de la información diagnóstica en general y de la minería de datos educacionales en particular, así como su posible aplicación en el proceso docente educativo de la Universidad de las Ciencias Informáticas.

- Inducción deducción: Para poder transitar del conocimiento de casos particulares a un conocimiento más general, que refleja lo que hay de común en los fenómenos individuales y viceversa, es el caso del procesamiento realizado.
- Modelación: Para reflejar de manera más eficiente el proceso educativo recogido por la información educacional.
- El enfoque de sistema**:** La gestión del conocimiento educacional tiene como reto en la UCI poder integrar datos de procesos formativos, investigativos y docentes que a su vez tienen componentes docentes y axiológicas.

Métodos empíricos

- Aplicación de encuestas a profesores y directivos de la universidad, para conocer sus opiniones y valoraciones sobre la modelación de los procesos de formación del profesional que se realizan actualmente y las posibles propuestas de mejoras.
- Procesamiento de datos educacionales mediante técnicas de minería de datos, de una muestra de estudiantes de primer año de la UCI pertenecientes al curso 2009-2010.
- Aplicación de métodos estadísticos para valorar**:**
    - La eficiencia del procesamiento de la información educacional que implica el modelo que se propone.

Importancia y actualidad del tema de la Investigación

La importancia y actualidad del tema radica en que disponer de información precisa es importante para la toma de decisiones en la UCI y en cualquier otro lugar.

La **novedad** de esta tesis está dada por:

- El diseño de un modelo integrado para el análisis de datos educacionales de procesos de formación del profesional.

Este modelo se distingue por:

- Utilizar datos relevantes para la toma de decisiones.
- Por el uso de técnicas y herramientas de minería de datos educacionales.
- Por los recursos humanos y competencias necesarias para trabajar con el modelo lo que permite que el análisis de los datos de los procesos formativos sea realizado por los profesores y otros decisores.

Aporte teórico: A partir de la sistematización realizada sobre los modelos aplicados en la minería de datos, se logró el diseño del modelo para el análisis de datos que incluye técnicas estadísticas multivariadas, de generación de reglas borrosas y de agrupamiento.

EL modelo está diseñado para reflejar la evolución de los estudiantes dentro del proceso docente educativo de manera que, siempre relacione la información nueva con la que se procesó anteriormente.

Aportes Prácticos: El procedimiento de aplicación y el software para la implementación. Su posibilidad de sociabilización.

Estructura de los capítulos de la tesis

En el capítulo 1 se realiza una valoración del estado del arte de la generación de modelos para el análisis de datos educacionales mediante la aplicación de técnicas de la minería de datos y de la estadística inferencial.

En el capítulo 2, se propone un modelo para el análisis de datos educacionales de los procesos formativos que resulta una ampliación de la manera tradicional de representar la información educacional en la UCI.

Se detallan los aspectos metodológicos necesarios para la elaboración del modelo mediante una combinación de ciertos algoritmos de minería de datos y la estadística inferencial. Se describen los algoritmos fundamentales del software prototípico **SystDiagnosPrognosxHQ**, encargado de generar el modelo de análisis de datos.

En el capítulo 3, se realiza la validación de la propuesta del capítulo 2 mediante la generación de un modelo con información diagnóstica y evaluativa real tomada de una muestra de estudiantes del ciclo básico de la facultad 1

Se realiza una comparación del modelo propuesto y el tradicional.

Finalmente se brindan conclusiones y recomendaciones, se detallan las referencias bibliográficas y se incluyen los anexos en formato digital.

1 Minería de datos aplicada en la elaboración de modelos de análisis de datos educacionales.

1.1 Modelos de la Minería de datos y la Estadística Inferencial

La palabra modelo se define en muchos diccionarios como una representación de la realidad por medio de sistemas formales cualesquiera. Este concepto no es tan simple como pudiera aparentar, tiene asociado tres dimensiones: Sintáctica, Semántica y Pragmática (8).

La Dimensión sintáctica se asocia con el rigor de las representaciones utilizadas (gráficas, ecuaciones, reglas). La dimensión semántica se vincula a la Pertinencia del modelo, es decir, a la relación entre el sistema formal y lo que se desea representar. La dimensión Pragmática, expresa la eficacia del modelo para lograr ciertos objetivos.

Un modelo no es un objeto matemático. Que esté construido con ecuaciones y variables significa que su sintaxis es matemática pero sus otras dos propiedades provienen de disciplinas que no son en absoluto matemáticas.

El concepto es bastante reciente, la "noción de modelo no fue utilizada ni manipulada antes de 1920 o 1930" (8), aunque si fue precedida por una "Matematización de lo real" o sea, procedimientos matemáticos para dar solución a problemas.

Por modelización se entiende a la actividad humana de elaborar modelos de sistemas y su historia pasa por la fusión paulatina de las aproximaciones a la verdad científica Hipotético-deductiva y la Inductiva.

Desde el punto de vista sintáctico, los modelos de minería de datos son el resultado de la aplicación de algoritmos del mismo nombre sobre bases de datos. Estos modelos pueden tomar forma tabular o gráfica. Por su parte, los modelos estadísticos, constituyen expresiones simbólicas en forma de igualdad o ecuaciones, aplicables a todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

En esta investigación se identificará como modelo tradicional para el análisis de datos, al conjunto de tablas de frecuencias y gráficos elaborados con técnicas estadística descriptiva que se utiliza con los objetivos de describir el desarrollo del proceso educacional y sustentar la toma de decisiones en el momento actual en la UCI.

1.2 Minería de datos y descubrimiento de conocimientos en bases de datos

El primer paso para la implementación de un buen algoritmo de procesamiento de la información debe ser el estudio del estado del arte relacionado con las herramientas que para tal fin existen y los algoritmos que permiten la extracción o descubrimiento de conocimientos en bases de datos (Knowledge Discovery in Data Bases o KDD, según sus siglas) (9).

"KDD es la extracción automatizada de conocimientos o patrones interesantes, no triviales, implícitos, previamente desconocidos, potencialmente útiles y predictivos de la información de grandes bases de datos" (9).

Los procesos KDD incluyen la clasificación mediante agrupamiento (Clustering), el reconocimiento de patrones, las predicciones y la detección de dependencias o relaciones entre los datos.

KDD posee 4 fases, ver figura 1.1 abajo:

1.Preprocesamiento de los datos y limpieza.

2.Aplicación de técnicas de minería de datos.

3.Descubrimiento de patrones.

4.Evaluación/Interpretación/Visualización del conocimiento resultante del procesamiento.

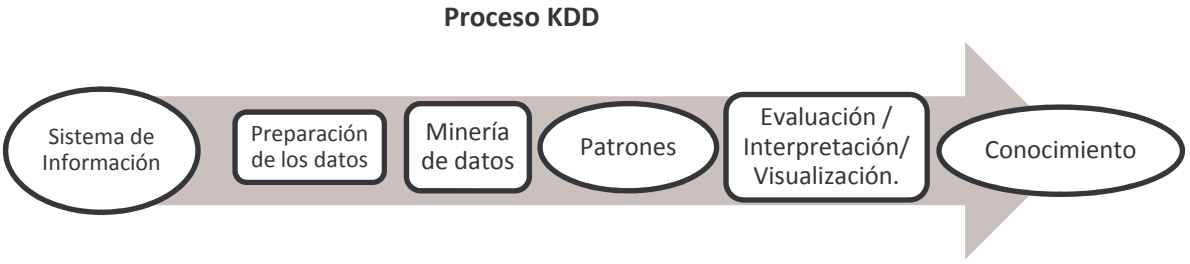


Figura 1.1: Fases del Proceso KDD. Fuente: (González, 2008)

Los datos a procesar contienen información sobre cada individuo u objeto sometido a investigación mediante múltiples variables aleatorias, interrelacionadas de forma tal que "sus diferentes efectos no pueden ser interpretados separadamente con algún sentido" (10).

Las técnicas de la minería de datos son el corazón del KDD y constituyen algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos con el fin de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

Se desea en la presente investigación, aprovechar las ventajas que aportan los modelos de minería de datos para reducir la redundancia de la información y por tanto, la complejidad del problema pues se orientan hacia la búsqueda de relaciones entre las variables o de las relaciones entre los individuos.

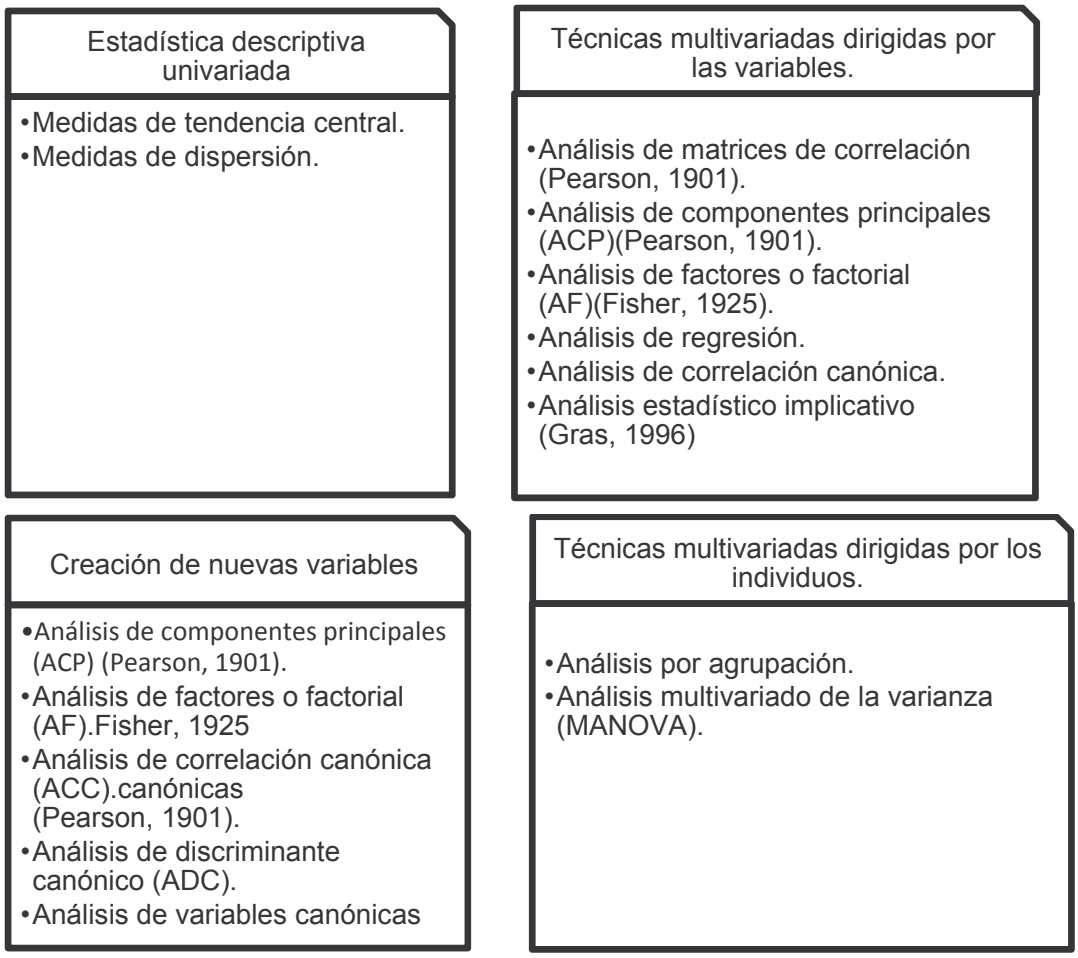
1.3 Técnicas Aplicadas en la Minería de Datos

Una herramienta de minería de datos no es más que una colección que agrupa un conjunto de técnicas o algoritmos de naturaleza:

1. Estadística
2. Lógico-Clasificatoria
3. De Inteligencia Artificial
4. De Otros tipos

A continuación se amplía la información sobre las mismas:

1.3.1 Técnicas de naturaleza Estadística:



El Análisis de Componentes Principales (ACP) y el Análisis Factorial son técnicas que crean nuevas variables no correlacionadas a partir de las variables originales.

El Análisis de Componentes Principales (ACP) permite investigar la existencia de instancias atípicas y validar las suposiciones de normalidad de la distribución de los datos. Las Componentes Principales pueden utilizarse como entradas en programas para la representación gráfica y agrupamiento.

El Análisis Factorial (AF) busca investigar si las variables respuesta exhiben patrones de relaciones entre sí, de modo que se puedan dividir en subconjuntos de variables tales que, las variables de un subconjunto están fuertemente correlacionadas entre sí y bajamente correlacionadas con las de otro subconjunto.

Los “Factores de Clasificación” tienen una mejor interpretación que los componentes principales del ACP y generalmente se presentan en una menor cantidad.

El Análisis Discriminante (AD) se usa principalmente para clasificar individuos o unidades experimentales en dos o más poblaciones definidas de manera única mediante la generación de reglas discriminantes del tipo:

$$FDi = \beta_{1i}X_1 + \beta_{2i}X_2 + \cdots + \beta_{mi}X_m \quad (1)$$

Donde i =1,2,…, k; j =1,2,…, m

Los parámetros β_{1i}, β_{2i},…, β_{mi} permiten determinar aquellas variables predictivas con mayor poder discriminante.

El Análisis Discriminante Canónico (ADC) es un procedimiento con el que se crean nuevas variables que contienen toda la información útil para la clasificación que se dispone en las variables originales por lo que conduce a reglas más sencillas.

La Regresión Logística (RL) permite modelar la probabilidad de que una instancia caiga en un grupo particular, con base a la información medida en la propia unidad, esto puede utilizarse con fines de discriminación. Se recomienda la aplicación de esta técnica en conjuntos de datos donde coexistan variables discretas y continuas.

El Análisis por Agrupación (AA) se usa para clasificar individuos o instancias en subgrupos definidos de manera única. Es una técnica muy popular que será tratada posteriormente.

El Análisis Multivariado de la Varianza (MANOVA) es una generalización de la prueba ANOVA (técnica usada para comparar las medias de varias poblaciones en una sola variable medida) por lo que permite la comparación de dos o más poblaciones diferentes sobre un número grande de atributos.

El Análisis de Variables Canónicas (AVC) crea nuevas variables en conjunción con los análisis MANOVA lo que ayuda a determinar donde ocurren las diferencias importantes entre las medias de dos o más poblaciones.

El Análisis de Correlación Canónica permite una vez divididas las variables respuestas en dos grupos determinar, si se pueden utilizar las variables de uno de los grupos para predecir las variables del otro grupo. De ser posible, “este análisis intenta resumir la relación entre ambos conjuntos de variables mediante la creación de nuevas variables a partir de cada uno de los grupos de variables” (10)

El análisis estadístico implicativo (11) desarrolla una metodología estadística no simétrica para estudiar las relaciones entre variables usando la idea de la lógica implicativa del álgebra booleana y se basa en el hecho de que los conocimientos, se forman a partir de hechos y de reglas que relacionan a los hechos o a las reglas mismas. Esos conocimientos se van formando en estructuras de manera progresiva.

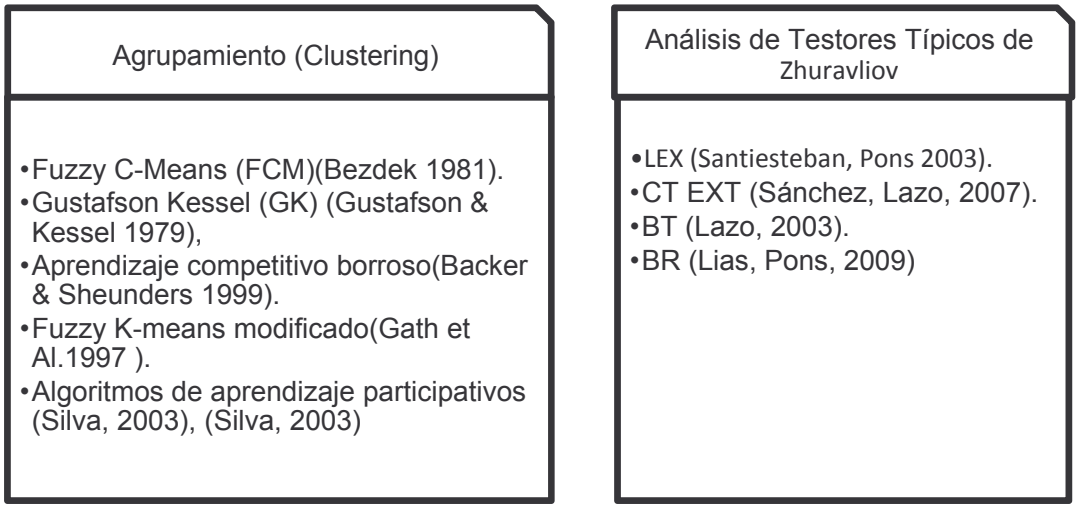
Entre las herramientas de naturaleza estadística más utilizadas en tareas de Minería de Datos se encuentran: R, SPSS Clementine (software) y CHIC (Clasificación jerárquica implicativa).

1.3.2 Conclusiones parciales sobre las técnicas estadísticas:

Del análisis de la naturaleza y diversidad de las técnicas estadísticas puede concluirse que:

- Se requieren una buena preparación de los encargados de aplicarlas y de interpretar sus resultados. En ocasiones existe divorcio entre los expertos en la estadística y los que deben idear los experimentos e interpretarlos.
- Antes de aplicar cualquier técnica, es necesario un análisis sobre el cumplimiento o no de las premisas que la sustentan. Muchos tipos de prueba supone la linealidad de la relación entre variables, esto no siempre se cumple.
- Este grupo de técnicas posibilita la creación de un modelo exploratorio de los datos de interpretación relativamente fácil debido a que se busca reducir la dimensión, determinar las dependencias entre variables y facilitar la representación tabular y gráfica.
- Lamentablemente, los modelos que se elaboran para sustentar investigaciones científicas no se generalizan con facilidad en la práctica educativa cotidiana.

1.3.3 Técnicas de naturaleza Lógica-Combinatoria



Estas técnicas tienen como objetivo final la clasificación de los datos (primer paso para la comprensión de un fenómeno complejo).

Los algoritmos de agrupamiento buscan optimizar una función objetivo de la forma:

$$J_m(X,U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ki}^{m}.d_{ki}^{2} \qquad (2)$$

Donde c es la cantidad de clusters, m una constante real mayor que 1; V es el conjunto de centroides v^0_1, v^0_2,…,v^0_k ; U es la matriz de membrecía cuyas celdas μ_{ki}^{m} ; expresan el tipo de pertenencia (dura o borrosa) del elemento x_k al conjunto centrado por v_i, y la distancia entre dos tuplas x_k y v_i se representa por $d_{ki} = \|x_k - v_i\|_A$.

Nota: Si $\|x\|_A = x^T.A.x$ y $A_{p\,x\,p}$ es la matriz identidad resulta que **d** es la distancia euclidiana

El análisis de Testores Típicos (reductos, rasgos relevantes) (12) , busca reducir la dimensión de los datos y encontrar los rasgos que inciden en un problema de manera determinante.

Un testor (Zhuravliov 1966) es un conjunto de rasgos que permite diferenciar entre dos clases, por lo que ningún objeto de la clase T_0 se confunde con objeto alguno de la clase T_1. Un Testor se llama Irreducible o Típico, si al eliminar cualquiera de dichos rasgos deja de ser testor para (T_0, T_1).

1.3.3.1 Conclusiones parcial sobre las técnicas lógico combinatorias.

El empleo de algunas técnicas de este grupo, otorgaría cualidades positivas al modelo que se desea construir y que se relacionan con:

- Poder tratar datos multivariados de naturaleza educacional
- Permitir la clasificación de datos con información ruidosa mediante la utilización de lógica borrosa o mediante el análisis de Testores.
- Los productos del procesamiento tienen una buena interpretabilidad debido a la considerable reducción de la dimensión y a la no destrucción de las variables originales.
- Permiten identificar los diferentes grupos de estudiantes atendiendo a la similitud de características individuales y de ese modo, brindarles una atención más personalizada.

1.4 Técnicas de Inteligencia artificial

Redes neuronales Artificiales.	Sistemas híbridos neuroborrosos.
•Perceptrón Multicapa (Rumelhart, Hinton y Willians, 1986). •Support Vector Quantization (SVQ).(C.Cortez & V. Vapnik,1995) •Mapas Autoorganizados de Cohonen (T. Cohonen, 1990).	•FALCON (C. T. Lin and C. S. Lee) •ANFIS (R. R. Jang, 1993) •NEFCON (D. Nauck & Kruse, 1998) •EFuNN and dmEFuNN (Kasabov and Song, 1998) •SECoS (Watts, 2002) •MLRul (Piñero, 2002).

Agrupamiento (Clustering)	Generación de árboles de Decisión
•K Means (Mac Queen, 1967). •K-Medoids(L. Kaufman & P.J. Roueseeuw, 1990)	•ID3 (Quinlan, 1986), C4.5 (Quinlan 1996), C5.0. (Quinlan,2000) •GID3 (Xizhao y Hong 1998) •CHAID (Hartigan, 1975). •Naives Bayes (Clark & Nibletts, 1989) •Ramdon Forest(Leo Breiman, 2001)

Las técnicas de inteligencia artificial (IA), se caracterizan por imitar el comportamiento de un ente inteligente.

Las redes neuronales artificiales, pueden aprender de forma supervisada con ejemplos (vectores de entrenamiento) o de forma no supervisada con un nivel aceptable de error (13), para su empleo exitoso, deben valorarse las siguientes ventajas y desventajas (14):

Ventajas	Desventajas
1. Capacidad de aprendizaje 2. Capacidad de generalización 3. Robustez en relación con los posibles disturbios del sistema de información.	1. Imposible interpretación de la funcionalidad, funcionan como una caja negra. 2. Es difícil determinar el número de capas y el número de neuronas.

Por su parte, los sistemas híbridos "… son sistemas difusos que usan algoritmos de aprendizaje basados en gradientes o en inspiraciones de la teoría de las redes neuronales (estrategias de aprendizaje heurísticas) para determinar sus parámetros (particiones y reglas borrosas) a través del procesamiento de patrones (de entrada y salida)" (14) (15).

Los sistemas híbridos neuroborrosos, generan una base de reglas borrosa que permite expresar lo aprendido en un lenguaje afín a cómo piensan los humanos, esto sería una propiedad deseable para la ampliación buscada del modelo de análisis de datos.

Los algoritmos generadores de árboles de decisión, son muy populares, ellos también realizan aprendizaje automático a partir de ejemplos preclasificados (tuplas compuestas por varios atributos y una única clase) como en el resto de los métodos inductivos. En la familia de algoritmos que emplean esta técnica, merece destaque los algoritmos ID3 (Induction Decision Trees), C4.5 y C5.0.

ID3 fue propuesto por J. Ross Quinlan en 1986, toma objetos de una clase conocida y los describe en términos de una colección fija de propiedades o de atributos, y produce un árbol de decisión sobre estos atributos que clasifica correctamente todos los objetos.

C4.5 y C5.0 se basan en el ID3, pero pueden clasificar los atributos continuos al representarlos con intervalos. El Algoritmo C5.0 es una versión comercial del algoritmo C4.5.

La literatura consultada (16) (17), revela la popularidad de las técnicas de generación de árboles en la clasificación de estudiantes atendiendo a su similitud.

El Agrupamiento o Clustering ya fue abordado en los epígrafes anteriores. Los algoritmos K means y K medoids [3]agrupan vectores según criterios de pertenencia dura.

A partir de predeterminar un número K de grupos (Clusters), un valor de tolerancia y un número máximo de iteraciones, se selecciona aleatoriamente un número K de instancias de la población para que sirvan de centroides[4] de los grupos que se desean formar.

Posteriormente se asignará a capa punto de la muestra el centroide de mayor cercanía; que será recalculado, teniendo en cuenta la membrecía de cada cluster y que no existan grandes cambios en las asignaciones. En caso contrario, debe volverse al primer paso.

1.4.1 Conclusiones parciales sobre las técnicas de IA.

- A partir de valorar la posibilidad de generación automática del modelo, las facilidades de interpretación y las posibilidades para la descripción y el pronóstico, el autor considera muy beneficioso el empleo de los sistemas híbridos neuroborrosos para la generación del modelo que resuelva el problema científico de la presente investigación.
- Este tipo de sistemas puede establecer relaciones no lineales entre variables como suele ocurrir en un proceso formativo.
- Este tipo de técnicas pueden procesar variables con métricas y escalas diferentes.
- Los conjuntos de datos pueden contener un elevado número de atributos, su procesamiento mediante árboles de decisión pueden ser de difícil interpretación.

1.4.2 Otros tipos de técnicas:

[3] Un Medoid se define como aquel objeto que posee el menor promedio de disimilitud con todos los elementos de un cluster dado.

[4] Centroides: Representaciones de cada cluster a través de sus vectores centrales, no necesariamente miembros del data set.

Interpretación de la interacción humana sobre sistemas tutoriales.
•Very Predictive NGramas (VNP)(Beal C. & Cohen P, 2008) •algoritmo Baum-Welch (Baum & Welch, 2003) •algoritmo de Proyección(Tompa, M. & Buhler, J., 2001) •Stepwise-HMM-Cluster (Shih,B ; Koedinger, K; Scheines, R, 2010) •Random projection multivariate motif discovery algorithm(Shanabroock,D.;Cooper, D, 2010)

Estas técnicas persiguen la modelación de procesos continuos relacionados con el aprendizaje de los sujetos que utilizan tutores inteligentes. Los datos de entrada son representaciones de las acciones de los estudiantes mediante secuencias de cadenas de caracteres o modelos ocultos de Markov[5] (HMM).

Además de la aplicación de técnicas de análisis de regresión logística y de la generación de modelos de regresión múltiple, se reporta la aplicación de algoritmos para la generación de reglas y el agrupamiento (18), (19).

A partir de identificar una secuencia de acciones estudiantiles A_1, A_2,…, el Algoritmo VNP (20), genera reglas de la forma $A_{n-j} \dots A_n \rightarrow A_{n+1}$ para diferentes valores de j (a diferencia de las cadenas predictivas de Markov que tienen un orden fijo) con el fin de maximizar la predicción dado un número fijo de reglas.

Estas reglas maximizan la probabilidad condicional: $P_r(A_{n+1} \backslash A_{n-j} \dots A_n)$

En comparación con las cadenas predictivas de Markov, VNP encuentra un número menor de reglas con una precisión aceptable.

El algoritmo Stepwise-HMM-Cluster crea poco a poco un conjunto de Modelos ocultos de Markov de gran poder representativo e interpretativo de las acciones estudiantiles (21), para ello se vale del algoritmo de agrupamiento HMM cluster y a su vez del algoritmo Baum-Welch que reactualiza los parámetros de aquel HMM que mejor ajuste los datos observados.

El algoritmo Random projection multivariate motif discovery algorithm permite encontrar secuencias plantadas en una larga cadena de caracteres, estas cadenas pueden estar previamente identificadas y poseer un significado específico.

Para la realización de una minería de datos exitosa se requiere además la aplicación de una buena metodología.

1.4.3 Metodología CRISP-DM en el desarrollo de proyectos de Minería de Datos

[5] Modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos.

CRISP-DM (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, Wirth, 1999) consiste en un conjunto de tareas descritas en cuatro niveles de abstracción: fase, tarea genérica, tarea especializada, e instancia de proceso, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos.

La metodología CRISP-DM puede utilizarse en múltiples herramientas de Business Intelligence y es muy cercana a las metodologías "convencionales" de gestión de proyectos, como RUP, MSF, etc.

En los capítulos Dos y Tres se detallará y aplicará esta metodología.

1.5 Conclusiones parciales de los aspectos 1.3 y 1.4:

Del estudio de cuatro grupos de técnicas de minería de datos, podría concluirse lo siguiente:

- Es amplia la diversidad de modelos que se pueden generar mediante la aplicación de cada técnica.
- Aunque existen varias herramientas dedicadas a la minería de datos, su empleo requiere cierto nivel de preparación de las personas que harán uso de ellas, esto obstaculiza su uso generalizado.
- La combinación de datos multivariados de diferentes formatos, orígenes y momentos puede ser analizada mediante una combinación de técnicas estadísticas exploratorias, de generación de bases de reglas borrosas y de agrupamiento con el fin de describir procesos complejos.
- Técnicas estadísticas como los Análisis de Componentes Principales y Análisis Factorial pueden combinarse con el análisis de Testores en la búsqueda de relaciones entre variables y de las variables más representativas.
- La generación de reglas borrosas facilita la realización de manera automatizada de la clasificación y pronóstico de datos con ruido, de naturaleza multivariada.
- La aplicación de algoritmos de agrupamiento permite combinar las instancias disponibles de manera acorde a su similitud con lo que se facilitaría la descripción y el pronóstico.
- Dado el carácter extremadamente multivariado y ruidoso de los datos, en esta etapa de la investigación no se pretende utilizar algoritmos generadores de árboles de decisión ni de reglas duras.
- Es recomendable seguir la metodología CRISP-DM durante la generación del modelo de un proceso.

1.6 Minería de datos educacionales

La Minería de datos educacionales (Educational Data Mining, EDM) (7) es el proceso de "transformación de los datos en bruto recopilados por los sistemas de enseñanza, en información útil que pueda utilizarse para tomar decisiones informadas y responder preguntas de investigación".

La emergente Minería de Datos Educacionales, combina las técnicas de aprendizaje automatizado mencionadas en el epígrafe anterior con otras que provienen de la psicometría o de otras áreas de las estadísticas y de la visualización de la información (7).

Algunos investigadores observan diferencias entre los métodos de las minerías de datos estándar y de los educacionales en particular debido a "…la necesidad de explicar las causas (y las oportunidades de explotación) que originan la jerarquía multinivel y la no independencia en los datos educacionales" (7)

Por eso es creciente la aplicación de modelos de la literatura Psicométrica en la minería de datos educacional.

El proceso de aplicación de las técnicas de minería de datos a los conjuntos de datos se denomina Modelado, a continuación se describirán algunas experiencias realizadas en diversas partes del mundo.

1.7 Principales tendencias en la generación de modelos de EDM en el mundo

1.7.1 Investigaciones de EDM en Europa

En el continente europeo se generaron numerosas investigaciones de minería de datos educacionales con objetivos que abarcan desde la modelación del trabajo estudiantil sobre plataformas a distancia hasta la predicción del éxito o fracaso docente ya sea en ambientes presenciales o a distancia. A continuación se resume información sobre la obtención de algunos modelos para el análisis de datos que pueden considerarse representativos.

Modelo propuesto por Agathe Merceron y Kalina Yaceff.

Con vistas a facilitar la evaluación de los estudiantes matriculados en la modalidad de enseñanza a distancia, Agathe Merceron y la australiana Kalina Yaceff (6) emplearon en el 2003, la herramienta C.H.I.C.(Logiciel d'analyse de données) (Couturiel, 2001; Couturiel et Al, 2000), para generar modelos del trabajo estudiantil basados en mapas conceptuales mediante la aplicación de técnicas de análisis estadístico implicativo

C.H.I.C. se caracteriza por la facilidad de interpretación de sus resultados para los docentes.

Este modelo aporta como experiencia a la solución del problema a resolver el modo en que implementa el monitoreo del proceso de enseñanza aprendizaje y la inclusión de toda la in-

formación disponible en un formato interpretable; se considera que la variables utilizadas tienen un carácter conductista por lo que pueden limitar la toma de decisiones pedagógicas.

Modelo propuesto por Kotsiantis.

Este estudio busca predecir el rendimiento de los estudiantes matriculados en la carrera de ciencias de la computación en la modalidad a distancia de la universidad helénica abierta de Grecia (22) , antes de iniciar el curso, para ello se analizaron las variables: sexo, edad, estado marital y rendimiento académico de una muestra de 365 estudiantes.

Se compararon los resultados de la aplicación de 5 algoritmos de clasificación relacionados con: Árbol de decisión, Perceptron basado en el aprendizaje, Redes Bayesianas, Aprendizaje basado en Instancias tomándose el resultado más eficiente como modelo .

El modelo de Kotsiantis enseña a aprovechar la información disponible para realizar predicciones del éxito estudiantil. Muestra lo positivo del empleo competitivo de varios algoritmos de clasificación.

El hecho de que las variables empleadas sean indicadores muy indirectos de aprendizaje o que algunas, estén sujetas a cambios temporales, son limitaciones de este modelo ya que hacen riesgosa la predicción.

El modelo provoca análisis muy generales, dirigidos a la población y no a los individuos lo que limita también la toma de decisiones.

Modelo propuesto por Kristjansson, Sigfusdottir and Allegrante

En este estudio del año 2005, se buscó predecir el rendimiento académico de estudiantes de secundaria utilizando variables relacionadas con los hábitos dietéticos, el índice de masa corporal, la autoestima y los resultados académicos de 6346 adolescentes de Iceland. Noruega (23)

Se encontró que un bajo índice de masa corporal, la actividad física y buenos hábitos dietéticos eran buenas predictivas del rendimiento académico.

De este modelo es útil aprender cómo se realizó la integración de variables provenientes de diversas fuentes.

La representación de la información es tan estática como una fotografía. Es un diagnóstico general con fines educativos para padres, estudiantes o docentes que no se actualizará en buen tiempo lo que limitará la toma de decisiones cuando las condiciones cambien.

Modelo propuesto por Cortez y Silva

Cortez y Silva (Portugal, 2007) generaron un modelo para la predicción del fracaso en dos asignaturas claves (Matemática y Portugués) sobre una muestra de 788 estudiantes provenientes de dos escuelas secundarias de la región de Alentejo (24), que participaban en el examen nacional del año 2006.

Se utilizaron 29 variables predictivas y 4 algoritmos de minería de datos: Árbol de decisión, Random forest, Neural Network y Support Vector Machine.

Este modelo sirve de ejemplo porque integró datos de diversas fuentes, empleó de manera competitiva a cuatro algoritmos clasificatorios. Puede considerarse una limitación que los resultados del estudio también tienen un valor general con vistas a prevenir el fracaso escolar.

El modelo no sirvió para el mejoramiento del aprendizaje de los estudiantes muestreados en su etapa secundaria.

Modelo propuesto por Mihaescu, Burdescu, Mocanu e Ioancu

Estos investigadores, dotaron a la plataforma de e-learning "Tesys" de un modelo cognitivo del aprendizaje de los estudiantes, basado en la aplicación de la técnica de agrupamiento K Means sobre los parámetros de las acciones de los aprendices en su interacción con dicha plataforma (25).

Es de utilidad para la presente investigación la retroalimentación con información periódica por el modelo descrito lo que permite tomar decisiones sobre las estrategias de enseñanza a utilizar con cada tipo de estudiante representado en los clusters.

El agrupamiento de los estudiantes en clusters facilita idear estrategias de trabajo específicas para cada tipo.

Los datos empleados en la creación del modelo son en su mayoría de naturaleza conductual lo que limita el alcance de los resultados.

Modelo propuesto por Ceneida Fernández y Salvador Llinares

Esta pareja de investigadores españoles estudiaron la relación entre el pensamiento aditivo y el multiplicativo en estudiantes de educación primaria en el proceso de creación del concepto de razón (26).

Como aspectos positivos para la investigación que se inicia se señala se destaca la creación de nuevas variables categóricas para caracterizar las estrategias usadas por los educandos de primaria cuando resuelven diferentes problemas proporcionales y de estructura aditiva y la posterior realización de un análisis estadístico implicativo con el fin de buscar las relaciones entre las estrategias.

Un punto fuerte en este modelo se relaciona con la generación de modelos con forma de mapas conceptuales por tipos de estudiantes, que permiten seguir los tipos de pensamientos presentes cuando se abordaban diferentes problemas dentro del proceso formativo.

El análisis estadístico implicativo no se utilizará en esta investigación, pues no se crearán modelos de análisis de datos para el desarrollo de las habilidades particulares de una asignatura en específico, pero se tendrá en cuenta para próximos estudios.

Modelo propuesto por Cristóbal Romero, Sebastián Ventura et al.

En un interesante artículo estos autores (18) compararon gran cantidad de algoritmos de minería para la clasificación de estudiantes a partir del registro de los datos de utilización y las evaluaciones alcanzadas en siete cursos de la Universidad de Córdova, montados en la plataforma Moodle.

Se aplicaron algoritmos de clasificación estadística, árboles de decisión, de inducción de reglas mediante aprendizaje automatizado, inducción de reglas borrosas, redes neuronales y algoritmos genéticos.

Un punto fuerte de modelo radica en el hecho de tomar información de siete cursos, la utilización de varios algoritmos de clasificación y el empleo de los registros de la muy extendida plataforma Moodle.

La limitación se encuentra en la naturaleza conductista de las variables y que la eficiencia de los algoritmos de aprendizaje automatizado no supere el 70%.

Comentarios del epígrafe.

Estas investigaciones aportan experiencias para la construcción de un modelo de análisis de datos puesto que:

- Se emplean variadas técnicas de EDM, acordes con los objetivos propuestos, los datos y recursos disponibles y el nivel de conocimiento de los investigadores.
- Se utilizan variables vinculadas con datos de ingreso a centros de estudio, hábitos, rasgos psicológicos, trazas de la interacción con una plataforma web y evaluaciones.
- Interesa la predicción del rendimiento de los aprendices tanto en la modalidad de estudio presencial como a distancia.
- La mayoría de los modelos analizados se elaboran después de la recopilación de los datos y ofrecen información de carácter general sobre la muestra.
- Se observó la aplicación de diversas técnicas y su utilización competitiva.
- La elaboración de estos modelos demanda recursos costosos para la preparación del personal docente, el acopio de información de diversas fuentes, el dominio de la minería de datos, de los medios de cómputo y la percepción por los centros de enseñanza de que con el modelo tradicional de análisis de información es suficiente para gestionar el proceso formativo.

1.7.2 Investigaciones de EDM en Asia

En un balance de las investigaciones de minerías de datos educacionales en La República Popular China hasta el año 2007, en base a los datos del sitio Web "The Chinese National Knowledge Infrastructure (CNKI) website", los señores Ling Jiang, Zongkai Yang, Qingtang Liu, y Haimei Wei reportaron que:

Desde el año 2001 en que se realizó la primera publicación, hasta el 2007 se contabilizaron 156 piezas, incluyendo 37 tesis de maestría y una de doctorado titulada "Study of Data Mining Based Assessment of Distance Learning" y publicada en el 2005 (16).

El 98% de estas publicaciones están relacionadas con la educación superior.

De interés para el modelo de análisis de datos que se quiere construir resultó conocer que el China las técnicas fundamentales aplicadas se relacionaban con la Estadística y Visualización,

Agrupamiento, Clasificación y detección de atípicos, Minería de reglas de asociación, de patrones y por último, Minería de Texto.

Modelo propuesto por May Liu, Wong, Yu y Lee.

El presente estudio presenta interés en la solución del problema científico debido a que busca modelar los diferentes niveles de rendimiento académico y detectar aquellos estudiantes de enseñanza secundaria que necesitan clases remediales en Singapur.

Los investigadores tailandeses (27), utilizaron las variables: Sexo, Región, Rendimiento escolar en años pasados para la predicción para comparar las medidas de asociación: Scoring based on associators (SBA score); C4.5 score y Naives Bayes score, constatándose que la primera superaba en un 20% a la precisión de las otras dos.

Un acierto de este modelo radica en utilizar la información del pasado pero desaprovecha la información que aporta el presente, es decir, la que se genera durante la marcha del curso, lo que indicaría de modo más preciso cómo dirigir los remediales en cada momento.

Se considera que aunque las variables utilizadas en la clasificación resultan de fácil adquisición tienen la limitación de no considerar otros parámetros de la personalidad de cada estudiante. Por otra parte, existen cuestionamientos pedagógicos hacia la utilización excesiva de las notas de las evaluaciones como variables clasificatorias.

Modelo propuesto por Hijazi y Naqvi.

Los investigadores paquistaníes (28), estudiaron la relación entre el desempeño estudiantil y las variables relacionadas con las variaciones en la atención a clases, horas dedicadas al estudio después del horario de clases, influencia familiar, edad de la madre y nivel educativo materno.

Se seleccionó una muestra de 300 estudiantes de un grupo de colegios afiliados a la universidad del Punjab en Paquistán y se aplicó la técnica del Análisis de regresión lineal simple.

Entre los aspectos útiles para la investigación que se realiza, puede señalarse el empleo de un modelo que relaciona variables predictoras y una variable clasificadora.

Resultó muy interesante que se registró la observación profesoral para definir las variaciones de la atención a clases. Podría hacerse investigaciones útiles si se utilizara más la observación profesoral como variable en los modelos.

Como aspectos negativos de la investigación obran el pequeño tamaño de la muestra, que se utilizan datos post proceso para ofrecer un diagnóstico general. La generalización del modelo obtenido es riesgosa debido a que a influencia o interacción de las variables estudiadas puede experimentar cambios en dependencia de factores culturales y regionales. No se habla nada sobre si se cumple los supuestos para la aplicación del método de regresión lineal con los datos disponibles.

Modelo propuesto por Khan

Este modelo (29) busca pronosticar el éxito en la educación preuniversitaria a partir del conocimiento de los valores de diversas medidas de la cognición, personalidad y variaciones de-

mográficas. Se tomó una muestra de 400 estudiantes de la escuela secundaria sénior perteneciente a la universidad musulmana de Arligath, India. Se aplicó una técnica de clusterización.

Es de interés como aquí se modelan datos de la enseñanza presencial preuniversitaria que provienen de diferentes fuentes y que incluyen aspectos de la personalidad. También se destaca el propósito perseguido el cual también sería un objetivo a seguir si se amplía el modelo para el análisis de datos de procesos formativos en la Facultad 1.

El modelo de Khan utiliza los datos finales del proceso de enseñanza, es decir, no fue generado dentro del curso escolar y no se utilizó para el monitoreo y corrección de los estudiantes muestreados.

Modelo propuesto por M.Ramaswami.

M.Ramaswami (30), buscó la predicción del desempeño estudiantil, en la educación media superior. Se procesaron datos educacionales y de tipo actitudinal sobre una muestra de 1000 estudiantes con el algoritmo generador de árboles binarios CHAID, que proporcionó una base de reglas duras.

Este modelo aporta a la solución del problema científico trazado, la experiencia de la combinación de datos actitudinales con datos educacionales.

Entre las limitaciones de los árboles binarios encontramos que el proceso de asignación de valores a las variables es en realidad borroso, por lo que se pueden cometer errores y por otra parte, el manejo de un gran número de variables puede generar un árbol no interpretable.

El modelo de M.Ramaswami ofrece información general por lo que puede no coincidir con la clasificación real de cada estudiante.

Comentarios del epígrafe.

Los modelos asiáticos examinados aportan información valiosa en cuanto a:

- La variedad de técnicas de EDM aplicadas en las investigaciones.
- La coexistencia de modelos para la educación presencial y a distancia que buscan facilitar el aprendizaje personalizado, la evaluación y la predicción del rendimiento docente en asignaturas complejas de nivel medio y superior.
- La aplicación competitiva de varios algoritmos de clasificación para seleccionar el mejor modelo.

En los trabajos antes descritos se observa que:

- Pocas investigaciones utilizan variables relacionadas con el desarrollo de la personalidad de los estudiantes o con las observaciones profesorales.
- Son escasos los modelos que se desarrollan a la par del proceso docente educativo y que sirvan para trabajar sobre la misma muestra.

1.7.3 Investigaciones de EDM en Oceanía

En este continente la EDM ha experimentado un fuerte desarrollo. Un grupo importante de trabajos se dedica a la enseñanza a distancia, he aquí una muestra:

Modelo propuesto por Kalina Yaceff y Agathe Merceron.

Las experiencias de trabajo acumuladas por esta pareja de investigadoras con la herramienta C.H.I.C., descrita en el epígrafe 1.3.1.1, sirvieron para el desarrollo, en el año 2005, de la plataforma TADA-Ed (Tool for Advanced Data Analysis for Education) (6) que permite a los profesores visualizar y hacer minería sobre el trabajo del estudiante on line para facilitar el descubrimiento de patrones relevantes.

El modelo se construye mediante la aplicación de algoritmos para la Clasificación y el Agrupamiento (K means, Clusterización Jerárquica y Árboles de Decisión), adaptados de la librería de datos de la herramienta libre Weka.

Entre las cualidades a imitar por el modelo de análisis que se desea elaborar encontramos el monitoreo constante del estudiante dentro del proceso de enseñanza aprendizaje, el propósito evaluativo que se persigue y las facilidades en la interacción de los profesores con el modelo.

Una limitación para el empleo educativo amplio de TADA-Ed radica en que las variables que procesa se refieren a acciones y calificaciones. En el Anexo 2 se muestran dos pantallas de TADA-Ed.

Modelo propuesto por A. K. Domínguez, K. Yaceff y J. Curran

Este trabajo aporta la experiencia de la creación de un modelo para la generación de un sistema de ayudas individualizadas en línea, destinado a los estudiantes participantes en la Competición Anual on Line de Programación del 2009 en Australia (31), que ofrece becas para la educación superior.

Previamente se combinaron los datos relacionados con las identificaciones de cada estudiante (ID), su desempeño en cada una de las 25 preguntas aplicadas atendiendo a los modos (Variable nominal) en que el estudiante intentó contestar o pasar la pregunta y la nota que alcanzó (0 o [5-10]) puntos en las ediciones del 2008 y 2009.

La vista minable incluye las preguntas, las correspondientes etiquetas de las ontologías, los resultados y las propuestas de los estudiantes y fue procesada mediante técnicas de generación de clusters (K means), reglas de asociación y análisis numérico simple quedando definidos tres grupos de estudiantes y conjuntos de ayudas pre intento y post intento (estas últimas, en caso de respuestas erróneas)

Es paradigmático que se genere un modelo del desempeño estudiantil que combina información de ediciones anteriores del concurso con las de la edición actual en el momento en que se genere. Esta es una cualidad que se imitará en esta investigación.

Herramienta Weka

Entre las herramientas de minería de datos ocupa un lugar importante la denominada como WEKA, elaborada por la Universidad de Waikato, Nueva Zelanda. (32)

Esta sofisticada herramienta, elaborada en software libre, tiene una amplia aceptación en todo el mundo y se distingue por la relativa facilidad de uso, por el empleo de diversos modos de acceso a sus numerosos grupos de software.

Una dificultad de WEKA radica en que por su carácter general, no ha sido diseñada expresamente para el tratamiento de datos educacionales y que por lo general, se dificulta la interacción de aquellos profesores que no tienen una preparación específica.

Comentarios del epígrafe

- Todas las investigaciones mencionadas en el epígrafe 1.3.3 se dirigen a la enseñanza o evaluación a distancia, persiguiendo facilitar la labor profesoral y/o proporcionar ayudas más inteligentes a los estudiantes.
- La generación de los modelos se realizan a partir de la recopilación de datos que provienen de la interacción de los estudiantes con las plataformas de enseñanza/evaluación.
- Merece destaque el proyecto destinado a brindar ayudas a los estudiantes para que resuelvan problemas de Anna Domínguez, Kalina Yacef y James Curran, cuestión en la que hay campo de investigación abierto.

1.7.4 Investigaciones de EDM en América del Sur

Modelo propuesto por Elena Durán y Rosanna Costaguta

Este modelo aporta a la presente investigación el hecho de permitir el análisis de un sistema de habilidades primordial para la enseñanza desarrolladora.

Ambas docentes, de la Licenciatura en Sistemas de Información de la Universidad Nacional de Santiago del Estero, Argentina, se apoyaron en el modelo pedagógico para la identificación de estilos de aprendizaje de Felder y Silverman (1988). (33), para detectar los estilos de aprendizaje que prevalecen en los estudiantes de su facultad mediante la aplicación del proceso KDD, siguiendo la metodología CRISP-DM con la utilización del algoritmo FarthestFirst para el análisis de cluster, implementado en WEKA.

Se aplicó un Test de Estilos de Aprendizaje de 44 preguntas con solo dos opciones de respuestas a una muestra del 10% de los estudiantes activos de la licenciatura, posteriormente se parametrizaron las respuestas para valorar el grado de desarrollo de los estilos.

Resulta interesante conocer que la información recabada con el test fue cargada en una planilla de EXCEL, luego convertida en un documento WORD para su posterior conversión a un archivo .arff requerido como entrada para el software WEKA.

Aunque debe utilizarse Software libre, en la presente investigación exploratoria, también se utilizará la plantilla Excel debido a su facilidad de uso

Un aspecto positivo del modelo radica en que proporcionó información general sobre los dos grupos de estudiantes fundamentales para que los profesores perfeccionaran su estilo de enseñanza pero desgraciadamente no fue recogida la pertenencia de cada estudiante al estilo de aprendizaje predominante lo que impidió perfeccionar de manera más exacta la labor educacional.

Comentarios del epígrafe

El trabajo visto aporta experiencias en cuanto a:

- La modelación de estilos de aprendizaje para perfeccionar los estilos de aprendizaje.
- La recolección de datos a través de encuestas basadas en un modelo pedagógico establecido.
- La parametrización de las respuestas de la encuesta para la caracterización del desarrollo de habilidades.
- La aplicación del algoritmo de clusterización y la obtención del modelo.

1.7.5 Investigaciones de EDM en América del Norte

Numerosas investigaciones estadounidenses abordan el monitoreo del aprendizaje estudiantil en su interacción con sistemas de tutoría inteligente, a continuación se describen algunos modelos representativos.

Modelo propuesto por Shanabrook D., Cooper D., Woolf B. e I. Arroyo.

Los autores crearon un modelo para el descubrimiento de motivos en los estudiantes que interactúan con el sistema tutor de aprendizaje matemático “Wayang” (21). Para ello registraron las acciones estudiantiles en la solución de un problema (cantidades de intentos, aciertos o no, ayudas, segundos entre ellas, etc.) para categorizarlas, binarizarlas, simbolizarlas y ordenarlas en un arreglo secuencial y representarlas mediante “palabras” de cuatro caracteres.

En total, se concatenaron 479 Secuencias estudiantiles (15048 caracteres), representando 3762 problemas, que fueron procesadas mediante una variante del algoritmo de proyección de Tompa y Buhler (2001) diseñado para buscar cadenas plantadas en una larga secuencia de caracteres.

Este trabajo aporta como experiencias la utilización de indicadores que el estudiante genera de forma inconsciente como una forma de interpretar los motivos que gobiernan su psiquis, el registro y concatenación en cadenas de palabras de dichos indicadores, la detección de tipos de cadenas especiales que representaban motivos y que una vez identificados, se pudo emitir ayudas personalizadas y perfeccionar la evaluación estudiantil.

Modelo propuesto por Shih,B. Koedinger, K y Scheines, R

Aquí se buscó el descubrimiento de las tácticas de aprendizaje empleadas por los estudiantes que interactúan con tutores inteligentes (20).

Se construyó la vista minable a partir de la representación de las acciones estudiantiles mediante modelos ocultos de Markov (HMM), elaboradas por el algoritmo Baum-Welch.

Mediante el algoritmo de agrupamiento no supervisado Stepwise-HMM-Cluster se seleccionaron modelos ocultos de Markov (HMM) pequeños realizando ajustes sobre las cantidades de modelos HMM a emplear y de los estados representados en cada uno.

Stepwise-HMM-Cluster incorpora iterativamente modelos individuales HMM a la colección final con mejores resultados, mediante una regresión lineal por pasos hacia adelante considerando el número total de secuencias clasificadas por cada modelo HMM por estudiante contra la ganancia de aprendizaje alcanzada entre los test de entrada y salida.

Entre las cualidades relevantes del modelo se encuentran su propósito de diseño: Ayudar a la evaluación precisa del trabajo estudiantil mediante el descubrimiento de sus tácticas de aprendizaje, la representación de procesos mediante modelos ocultos de Markov y el empleo de una técnica de análisis cluster.

El alcance del trabajo no trasciende la realización de una sesión de trabajo con el tutor inteligente a menos que se registren los resultados históricos asociados a cada estudiante.

La utilización de modelos ocultos de Markov pudiera ser incorporada en el futuro al modelo para el análisis de datos en la facultad 1 realizando una adecuación de las variables.

Modelo propuesto por D'Mello, S. Graesser, A.

D'Mello, S. Graesser investigaron las asociaciones entre los estados afectivos y las posturas físicas de los estudiantes en las sesiones de aprendizaje con el sistema de tutoría inteligente "AutoTutor" (19).

Dotaron a los sillones del laboratorio de sensores de presión corporal y de un mecanismo para registrar esos datos. Por otra parte, un grupo de expertos analizó los videos de las caras de los estudiantes mientras interactuaban con el sistema.

Se combinaron los datos y se procesaron con la técnica de regresión logística binaria lo que permitió generar un modelo capaz de discriminar los estados afectivos y en consecuencia, el diseño de mejores sistemas de aprendizaje.

Se modelaron los datos recolectados mediante cadenas ocultas de Markov, cadenas de caracteres con una longitud finita y matrices numéricas.

Se conocía previamente la posible relación entre los tipos de valores en los modelos de datos y grupos de patrones conductuales y/o estados de ánimo

Con el fin de reducir la necesidad por los análisis de los expertos en dominios, se valoró el empleo de técnicas de minería de datos de aprendizaje no supervisado que resultaran robustas y con resultados interpretables, por lo que se seleccionaron los algoritmos Stepwise-HMM-

Cluster, Random Projection, de Regresión múltiple y "Very Predictive Ngrams" (VPN), en la generación de reglas predictivas.

Los resultados alcanzados permitieron realizar modificaciones en la elaboración de sistemas de tutoría inteligente tanto en su estructura como en los materiales didácticos y los sistemas de ayudas.

El estudio del modelo aportó como experiencia que es posible tratar de relacionar comportamientos humanos que escapan de la conciencia humana con los estados afectivos que se experimentan al enfrentar la solución de tareas escolares.

1.7.6 Taxonomía de la EDM propuesta por Ryan S. Baker

El investigador Ryan Baker clasificó los trabajos de minería de datos educacionales a partir de la aplicación de las técnicas de Predicción (mediante Clasificación, Regresión ó Estimaciones de Densidad), Agrupamiento, Minería de Relaciones (incluye Minería de reglas de Asociación, Minería de Correlaciones, Minería de Patrones Secuenciales y Minería de Datos de Causas), Depuración de datos para Jurados humanos y Descubrimiento basado en modelos (17).

La depuración de datos para jurados humanos (como en los tres últimos modelos) está orientada a la toma de decisiones y es una necesidad para cualquier tipo de enseñanza puesto que son los profesores y directivos, los máximos responsables de la calidad del proceso docente educativo.

El desarrollo del modelo de un fenómeno, a través de cualquier proceso que pueda ser validado (usualmente mediante predicción o mediante Ingeniería del conocimiento), permite que ese modelo sea usado como componente en otro análisis (por ejemplo de Predicción o Minería de Relaciones) (17)

Valoración de estos modelos desde el punto de vista de su utilidad para la solución del problema científico trazado en esta investigación:

- Es posible elaborar modelos de análisis de procesos o secuencias de acciones cuando se utilizan sistemas tutoriales a distancia.
- Se observa la utilización de modelos ocultos de Markov y de secuencias de caracteres organizadas en "palabras" en la representación de acciones.
- Se trata de penetrar en la psiquis humana a través de la interpretación de acciones inconscientes del estudiante con un significado predefinido por los expertos en Psicología.
- Estos estudios indican un nivel avanzado con respecto al resto de las áreas geográficas, en la aplicación de las técnicas de minería de datos educacionales en la evaluación del aprendizaje mediante tutores inteligentes.

1.7.7 Investigaciones de EDM en Cuba

Aunque en Cuba, existen numerosas investigaciones relacionadas de alguna manera con la minería de datos educacionales, resultaron relevantes para la presente investigación las siguientes investigaciones:

- El Razonamiento basado en casos en el ámbito de la Enseñanza/Aprendizaje (34) de las autoras N. Martínez, G. Ferreira, M. García, Z. García pertenecientes a la Universidad Central Marta Abreu de Villa Clara.
- Técnicas de Minería de Datos del autor MsC. Ernesto González Díaz (35) profesor de informática del Instituto Politécnico "José Antonio Echevarría" (ISPJAE) de la capital.
- "Minería de datos aplicada a la gestión docente del instituto superior politécnico José Antonio Echevarría" de los autores Raycos Brito Sarasa, Alejandro Rosete Suárez, Rolando Acosta Sánchez profesores de la carrera informática en la CUJAE (36).
- Experiencias con SAED: Sistema informático para la autoevaluación de estudiantes a distancia. (37)

La consulta bibliográfica permitió constatar que algunas investigaciones cubanas se relacionan con la aplicación de agentes inteligentes (38) por lo que se precisa el monitoreo de la actividad estudiantil.

En la publicación relacionada con el razonamiento basado en casos, resulta interesante la propuesta de realizar la modelación del trabajo estudiantil para el almacenamiento y posterior reutilización del caso, se procesan los datos provenientes de las respuestas estudiantiles mediante un análisis de Testores típicos sobre el sistema HESEI.

Los dos trabajos realizados por los profesores del ISPJAE (35) (36), demuestran la posibilidad de realizar minerías de datos sobre la información educacional almacenada en las secretarías del propio ISPJAE y de la UCI.

En ambos casos se construyeron vistas minables que incluían datos personales de matrícula y las evaluaciones de las asignaturas (en el ISPJAE se consideraron todas las asignaturas de las carreras durante varios lustros; En la UCI se seleccionaron las evaluaciones de un grupo de asignaturas significativas en el primer año de la carrera durante cuatro cursos consecutivos)

Los datos del ISPJAE fueron procesados mediante la versión implementada en la herramienta WEKA mientras que los datos de la UCI se procesaron con la herramienta SQL Analysis Server a través de análisis clúster, árboles de decisión y algoritmos de aprendizaje inductivo.

Los objetivos de minería de datos trazados por los investigadores pueden extenderse también al modelo que se desea construir, es decir, estudiar la influencia de la procedencia social en el resultado académico y predecir la nota final en cada asignatura analizada.

Ambos estudios concluyeron con la generación de reglas duras para permitir a los decisores docentes fortalecer aspectos formativos, este formato quizás contribuyó a que los profesores no llegaran a utilizar los modelos para perfeccionar su labor docente.

Dadas las condiciones cambiantes en los programas de estudio, el medio social, los métodos la utilización de información almacenada puede afectar la fiabilidad de los modelos.

En el trabajo relacionado con el empleo del sistema SAED (37), el aprendiz cuenta con un grupo de ejercicios que le permiten autoevaluarse una vez revisados los materiales didácticos del tema. Cada ejercicio provee retroalimentación dándole al estudiante el resultado correcto una vez finalizado. Se reportó una experiencia exitosa con siete estudiantes.

Se aprecia en estas investigaciones, un interés por extraer información relevante que posibilite tomar decisiones que fortalezcan los procesos de enseñanza aprendizaje

No se encontró, entre las investigaciones cubanas, alguna que trabajase la modelación de datos para el análisis de procesos de formación del profesional en régimen presencial dentro del curso escolar, que maneje variables relacionadas con aspectos cognitivos, actitudinales y del desarrollo de la personalidad con el fin de facilitar a los docentes la gestión de la información educacional.

Conclusiones del Capítulo

- La minería de datos abarca un grupo amplio de técnicas, corresponde al investigador utilizar las más apropiadas de acuerdo al objetivo que persiga, los recursos disponibles y la preparación de los encargados de interpretar sus resultados.
- Cada año aumenta el número de aplicaciones de minería de datos educacionales en las modalidades presenciales y a distancia. Los modelos que se generan persiguen la predicción del rendimiento docente, de las estrategias de aprendizaje y características sicológicas, sobre todo a partir del trabajo sobre o en una asignatura.
- No abundan las investigaciones sobre modelos capaces de monitorear procesos de formación de profesionales mediante la integración de datos de diferentes asignaturas es decir de notas, habilidades, valores, competencias, observaciones de los docentes, etc.
- Entre el conjunto de modelos de minería de dato educacional, resultan atractivos aquellos que se generan mediante la integración de técnicas del análisis multivariado, agrupamiento de datos y la generación de reglas borrosas.
- El empleo de la metodología CRISP-DM en la generación del modelo contribuirá a asegurar la calidad en el proceso de MD.

En el siguiente capítulo, se propondrá una ampliación del modelo para el análisis de datos educacionales existente, en el proceso de formación del profesional en la Universidad de las Ciencias Informáticas.

2 Modelo integrado para el análisis de datos de procesos de la formación del profesional

Los procesos de formación del profesional generan todo el tiempo gran cantidad de datos que se distinguen por incluir todos los tipos de escalas, su diversidad de formatos, grado de relevancia variable, por la fuente que generó dichos datos, etc.

Gestionar la información de las evaluaciones sistemáticas, parcial y final por medio de notas unidas con otras informaciones diagnósticas (por ejemplo indicadores del desarrollo de habilidades, competencias y valores, aspectos psicológicos) puede ser muy complicado.

A su vez, cada sujeto que interviene en el proceso formativo llega a formar criterios del mismo a partir de su rol, experiencia, interacción y observación, que sería valioso tener en cuenta por los encargados de tomar las decisiones.

Dada la complejidad inherente a la gestión de los datos educacionales y con vistas a proporcionar una solución práctica a la problemática de encontrar un marco que integre la información educacional disponible a medida que se genere, se decidió adaptar la metodología CRISP-DM al proceso de generación de modelos de análisis de datos educacionales lo que constituye el primer aporte teórico de esta investigación.

2.1 Adaptación de la metodología CRISP-DM para la generación del modelo buscado

A continuación se ilustra mediante un esquema (figura 2.1), la adaptación realizada y posteriormente se describirán sus fases:

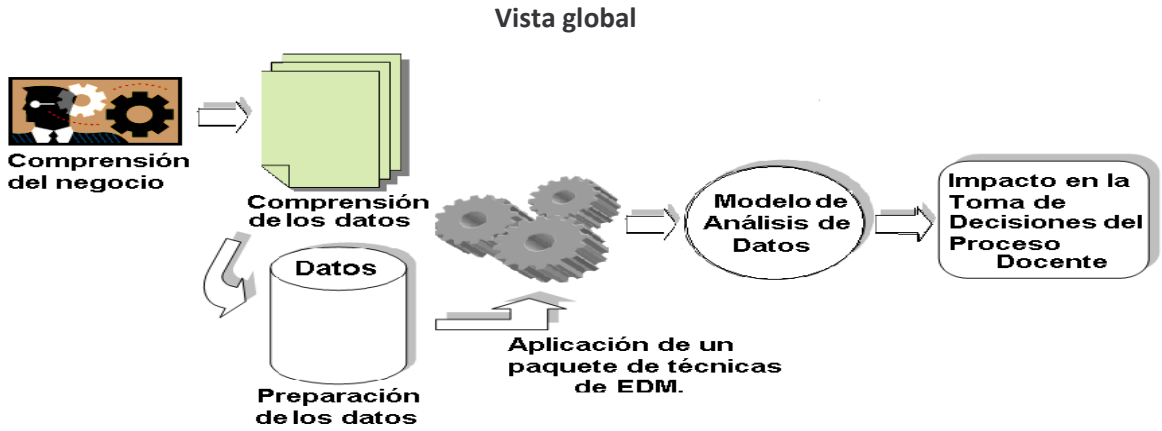


Figura 2.1 Vista Global de las fases de generación del modelo de análisis de los datos para los procesos de la formación profesional (Elaboración propia basada en la metodología CRISP-DM)

Fases de CRISP-DM

1.-Comprensión del negocio (Selección del tipo de proceso de formación profesional a investigar)

2.-Comprensión de los datos (Recopilación de los datos)

3.-Preparación de los datos

4.-Modelado

5.-Evaluación

6.-Despliegue.

El segundo aporte teórico de esta investigación radica en el modelo el cual se encuentra inmerso en el proceso formativo, integrando la información nueva con la vieja, lo que permite analizar la marcha de los procesos modelados "en tiempo real".

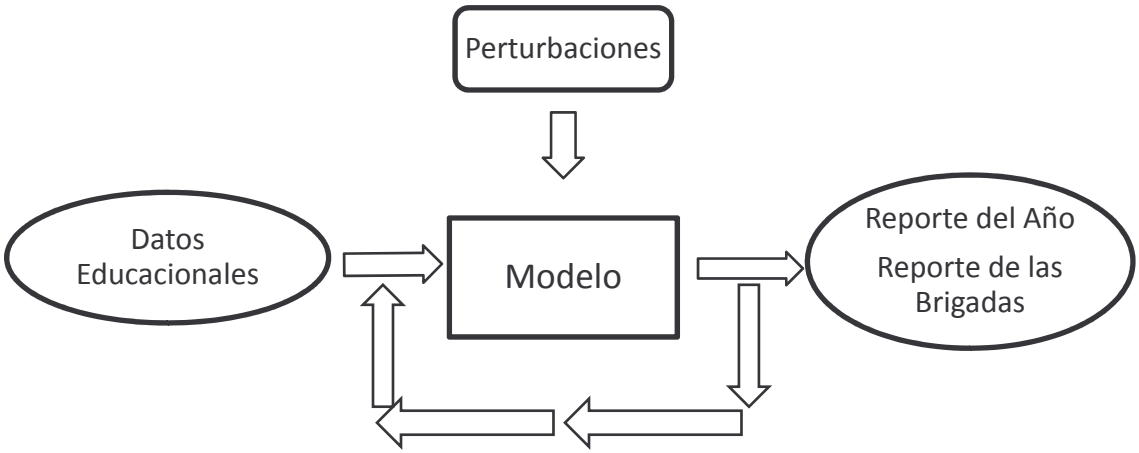


Figura 2.2 Entradas, salidas y perturbaciones en el modelo integrado para el análisis de datos del proceso formativo.

A la entrada del modelo se introducen los datos educacionales disponibles que corresponden generalmente a información diagnóstica de tipo general. La información resultante del modelo también puede ser reintroducida y combinarse con nuevos datos.

Entre los factores que pueden ocasionar perturbaciones al proceso de modelado se puede señalar la existencia de datos incompletos o con errores, excesiva granularidad en los datos fuentes, que los datos ordinales no obedezcan a escalas de Lickert.

Sólo se necesitan dos tipos de reportes como salidas del modelo, debido a que con ellos pueden describirse relaciones entre variables y además las características generales y particulares de los grupos y de los individuos.

Lo anterior se logra, porque ambos reportes contienen las componentes que se ilustran en la figura 2.3 las que permiten dar un tratamiento integral a la información y representarla en un formato más pequeño capaz de transitar sin deformaciones por toda la pirámide de decisores.

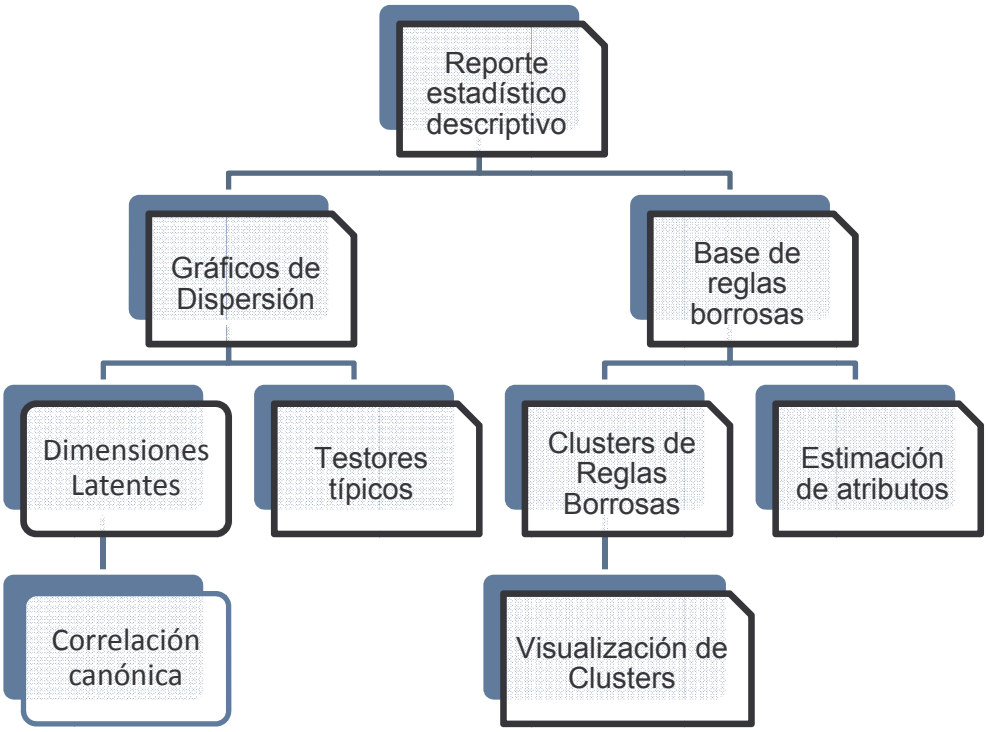


Figura 2.3 Componentes del modelo de la Vista Minable. Fuente: Elaboración propia del autor

El reporte estadístico descriptivo permite explorar a cada una de las variables disponibles de manera separada, devuelve sus medias, desviaciones estándar y su distribución de frecuencias tal y como se hace tradicionalmente.

Mediante gráficos de dispersión multivariados se realiza la exploración de relaciones entre grupos de variables, las dimensiones latentes permiten descubrir aquellos grupos de variables que caracterizan a la muestra, los testores típicos por su parte identifican las variables más relevantes y mediante el análisis de correlación canónica se explora cómo un grupo de variables puede influir sobre el resto de los grupos de variables.

Por la otra rama de componentes, la generación de una base de reglas borrosas permite la descripción de la muestra y la realización automática de clasificaciones con fines predictivos, esto permite encontrar el desarrollo posible de los individuos a partir de la variación de los valores de las variables predictivas.

Los clusters de reglas borrosas constituyen la principal forma de representación de los datos en este modelo pues consiguen visualizar la totalidad de la muestra agrupada en un número relativamente pequeño de individuos muy similares con el fin de permitir la adopción de estrategias educacionales. Se completa el modelo con la inclusión de un grupo de herramientas de visualización.

A continuación se describirán los pasos necesarios para la elaboración del modelo siguiendo la adaptación realizada a la metodología CRISP-DM.

2.2 Fase 1: Comprensión del negocio

Esta fase incluye las siguientes tareas:

Tarea 1: Establecimiento de los objetivos del negocio.

Tarea 2: Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio)

Tarea 3: Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)

Tarea 4: Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

Desarrollo

Para el establecimiento de los objetivos del negocio, debe precisarse el contexto inicial, los fines que se persiguen de manera no técnica y los criterios de éxito.

Como contexto inicial puede señalarse la necesidad de complementar el modelo de análisis de datos del proceso de formación del profesional, en la carrera ingeniería en Ciencias Informáticas.

Este proceso de formación profesional se orienta cada vez más a la formación de competencias y valores, por lo que su desarrollo y evaluación requiere un seguimiento general y particularizado.

En la UCI se trabaja por potenciar el aprendizaje de los estudiantes por lo que se requiere fortalecer la gestión del conocimiento relacionada con la evaluación y el control para una correcta toma de decisiones.

Como objetivos del negocio se propone:

Elaborar un modelo para el análisis de los datos de los procesos formativos del profesional de la carrera informática basado en las herramientas y metodologías de la minería de datos educacionales.

Se adoptaran como criterios de éxito los elementos de la primera columna de la tabla que aparecen en el Anexo 3; Puede verse que en las columnas 2 y 3 de dicha tabla, se establecen los indicadores y criterios de medida correspondientes.

Evaluación de la situación:

Es necesario monitorear de forma cerrada las trayectorias individuales de los estudiantes.

Las encuestas aplicadas en el claustro de profesores revelan una opinión favorable hacia la búsqueda de mejoras en los procesos de análisis de los datos de los procesos formativos.

Se dispone en la universidad de un sistema de gestión académica y de herramientas para la aplicación de diagnósticos que pueden servir de soporte a investigaciones para la modelación de datos de análisis de procesos.

Objetivos de la minería de datos

A partir de los objetivos trazados anteriormente y de los criterios de éxito esbozados en la tabla del anexo 2, se proponen los siguientes objetivos de minería de datos:

1) Realizar la descripción de muestras mediante la aplicación de estadísticas descriptivas.
2) Búsqueda de relaciones entre las variables.
3) Búsqueda de relaciones entre los individuos mediante la realización de una clasificación múltiple de los estudiantes con el fin de describir la muestra.
4) Proporcionar un marco fácil de socializar e interpretable, que permita incluir y contrastar la nueva información con la vieja ya almacenada, de forma que se pueda valorar la marcha del proceso formativo en el tiempo.
5) Estimación de Consecuentes.

Criterios de éxito

Desde el punto de vista de la aplicabilidad de la propuesta, se considerará el criterio de un conjunto de 100 directivos, profesores y especialistas en minería de datos acerca de la validez del modelo propuesto.

Desde el punto de vista técnico, el cumplimiento por objetivos se garantiza de la siguiente forma:

Objetivo 1: Las estadísticas descriptivas tienen probado éxito en la exploración de conjuntos de datos.

Objetivo 2: La búsqueda de relaciones entre variables permite detectar las más relevantes logrando una mejor interpretación con la reducción de la dimensión de los datos. Se determinarán aquellas variables que logren una explicación de la variabilidad de los datos superior al 60%, ya que la literatura en el campo de los estudios sociales, acepta como bueno o permisible el 60%.

Objetivo 3: La búsqueda de relaciones entre los individuos permitirá orientar la intervención pedagógica hacia los diferentes tipos de estudiantes.

Primero, se realizará una clasificación automatizada múltiple de los estudiantes mediante la generación de bases de reglas borrosas. Dada la cantidad de premisas y consecuentes que puede tener una regla borrosa, será considerado satisfactorio:

- Que se alcance una tasa de Aprendizaje (asociación entre premisas y consecuentes) superior al 90% sobre los datos de entrenamiento y sobre 70% en las estimaciones sobre los datos de prueba (datos reservados no para el entrenamiento sino para evaluar el aprendizaje con fines de prueba), considerando que en la base de datos existan ca-

sos ambiguos o dudosos que no se les dio tratamiento. Se aplicará la técnica de validación cruzada para la selección de la mejor base de reglas borrosas.

Posteriormente, se agruparán aquellas reglas por su similitud realizando una validación cruzada que permita seleccionar un número de centroides no muy grande y que mejor representen a los datos muestreados.

Objetivo 4: El empleo de reglas borrosas y de clusters de las mismas permite:

- La utilización de variables lingüísticas en cada atributo lo que permite la representación numérica y simbólica de sus valores inherente a los conjuntos borrosos.
- Ubicar grupos de variables predictivas y clasificadoras en una misma regla o clúster con lo que se puede:
    - Representar la evolución de los procesos en el tiempo al permitir la incorporación de nuevos grupos de variables a los ya existentes.
    - Enriquecer la representación del proceso al incluir mayor cantidad de variables.
    - Visualizar las fortalezas y debilidades de cada grupo de individuos clusterizados.
- Ahorro de espacio pues el modelo a socializar solo incluye los clusters de reglas que resumen información, esto facilita que a todos los niveles de decisión se pueda observar y analizar la misma información.

Objetivo 5: La generación de una base de reglas neuro borrosas otorga al modelo capacidad de pronóstico, al evaluar de manera automatizada dicha base mediante juegos de datos de las variables predictivas para calcular consecuentes estimados.

O sea, es posible clasificar de manera automatizada a los individuos siempre que se conozcan los valores de sus variables predictivas.

Los expertos humanos podrán interpretar los resultados del modelo y tomar las decisiones apropiadas.

Diseño de Pruebas

Se realizarán los diseños de casos de prueba según la técnica de Validación Cruzada; la cual consiste en dividir los datos en 10 grupos y realizar 10 corridas o iteraciones donde en cada una se combinan 9 muestras para formar un conjunto de entrenamiento y se deja la restante como muestra de prueba.

De esta forma todas las muestras son utilizadas como experimento y como prueba. Al final se selecciona el experimento sobre el cual se realicen mejores predicciones, o sea donde el error sea menor.

Plan de Procesamiento: Está previsto realizar el análisis de las posibles combinaciones entre los atributos con el fin de generar reglas borrosas.

2.3 Fase 2: Comprensión de los datos

Aquí se persigue una familiarización con los datos teniendo presente los objetivos del negocio.

Las tareas de la fase son:

- Recopilación inicial de datos.
- Descripción de los datos.
- Exploración de los datos.
- Verificación de calidad de datos.

Desarrollo

Los objetivos de la investigación dependen de lo que se desee investigar (los procesos de aprendizajes de contenidos, de desarrollo de competencias, productivos, capacidades investigativas, liderazgo, fracaso escolar, motivación, etc.)

La recopilación de los datos deberá hacerse desde todas las fuentes disponibles que contengan información de interés de acuerdo a los objetivos del negocio.

En la UCI se dispone del gestor de datos Akademos (39) el cual posee datos del proceso docente educativo por lo que debe ser una de las fuentes de información más importante.

Otras fuentes de datos pueden encontrarse en los diagnósticos de inicio y fin de cada ciclo pedagógico y en investigaciones de carácter científico general que pueden realizar colectivos de profesores.

Tarea: Descripción de los datos:

Cada tipo de dato será descrito teniendo en cuenta su recorrido, naturaleza numérica o nominal, continua o discreta, ordinal o categórica.

Tarea: Exploración de los datos

Deben detectarse datos con omisiones en algunas celdas, para darle el tratamiento debido.

Tarea: Verificación de la calidad de datos.

Es muy importante el trabajo con datos fiables que cuenten con la aprobación de especialistas competentes y si es posible con el aval de haber sido utilizados en diferentes investigaciones de forma exitosa.

2.4 Fase 3: Preparación de los datos

Se busca aquí obtener la vista minable o dataset, por lo que se deben realizar las siguientes tareas:

- Selección de los datos.
- Limpieza de datos.
- Construcción de datos.

- Integración de datos.
- Formateo de datos.

Tarea: Selección de los datos.

La selección de los datos obedece al cumplimiento de los objetivos trazados en la fase 2 de la metodología CRISP-DM, de la disponibilidad existente y del seguimiento de los pasos descritos previamente.

Con el fin de aplicar la versión del algoritmo SECoS (40) que genera reglas de aprendizaje tipo Sugeno grado 0, se deberán definir variables predictivas y clasificadoras.

Es importante garantizar el cumplimiento de las premisas previas a la aplicación de cada técnica de minería de datos.

Tarea: Limpieza de datos

Se sugiere que se descarten aquellas instancias con datos incompletos y que posteriormente se estimen los mismos por medio de algoritmos de inteligencia artificial en lugar de sustituirlos por medias o medianas.

Tarea: Construcción, integración y formateo de datos

El proceso de integración de grupos de datos que provienen de diferentes fuentes genera una tabla como la que aparece en el Anexo 4.

Los grupos de variables contienen información diagnóstica, resultados de evaluaciones, de investigaciones u otra información educacional disponible.

Los atributos de la tabla en tanto representan datos educacionales con escalas continuas, discretas, ordinales o nominales con sus correspondientes dominios de definición.

Mediante numerización o variables artificiales dicotómicas, se llevaran las variables no métricas a métricas, esto se hace con el fin de facilitar el cálculo de distancias en los algoritmos generadores de reglas y agrupamiento.

Para el caso de variables numéricas continuas o discretas de diferentes recorridos, se recomienda su estandarización.

En esta investigación se asumirá la pertenencia a escalas de Lickert de los datos categóricos ordinales para numerizarlos y facilitar la aplicación de algoritmos que generen reglas borrosas.

En caso de necesidad, se crearán nuevas variables a partir de otras.

Para la generación de reglas esta vista minable será redistribuida de manera que una parte de los atributos sean predictores y la otra parte clasificadores (en este último caso sus atributos deben tomar valores constantes) Ver Anexo 5.

2.5 Fase 4: Modelado

Con vistas a aplicar las técnicas de minería de datos a los dataset se debe:

- Seleccionar las técnicas del modelado.

- Diseño de la evaluación del resultado.
- Construir el modelo de la minería de datos.
- Evaluar dicho modelo.

Desarrollo

Tarea: Selección de la técnica de modelado.

En la tabla 2.2 se muestra la selección de las técnicas de modelado.

Mediante la aplicación de técnicas de exploración podrá investigarse la relación entre las diferentes variables lo que se tendrá en cuenta a la hora de interpretar los resultados de las técnicas de agrupamiento.

A partir de la separación de las variables en predictivas y clasificadoras se pueden generar reglas borrosas del tipo Sugeno grado 0, que podrán ser evaluadas con las premisas de otras instancias con el fin de realizar estimaciones de consecuentes.

2.5.1 Paquete prototípico SystDiagnosPrognosxHQ

Para materializar y evaluar la propuesta de esta investigación, se elaboró en Matlab 7.0 un software prototípico denominado **SystDiagnosPrognosxHQ** que incluyó las siguientes técnicas de minería de datos:

- Análisis exploratorio para buscar relaciones entre las diversas variables de la vista minable.
- Análisis exploratorio factorial y de testores típicos, para reducir las dimensiones del modelo, buscar las dimensiones latentes entre variables.
- Extracción de reglas borrosas tipo Sugeno grado 0, mediante los algoritmos neuro borrosos MLRUL (Piñero Pérez et al, 2006) (40) y SECoS (Watts., 2005) (39 págs. 11-24). Se conformará una base de reglas para la clasificación y el pronóstico.
- Agrupamiento de la Base de Reglas mediante los algoritmos K Means y Fuzzy C-Means.

2.5.1.1 Módulos y Flujo de datos en el paquete SystDiagnosPrognosxHQ

El paquete SystDiagnosPrognosxHQ incluye 4 módulos para el análisis exploratorio de los datos, el aprendizaje, la evaluación y visualización, cada módulo a su vez, agrupa diversas técnicas, a continuación se muestra un esquema:

Módulos del paquete SystDiagnosPrognosxHQ

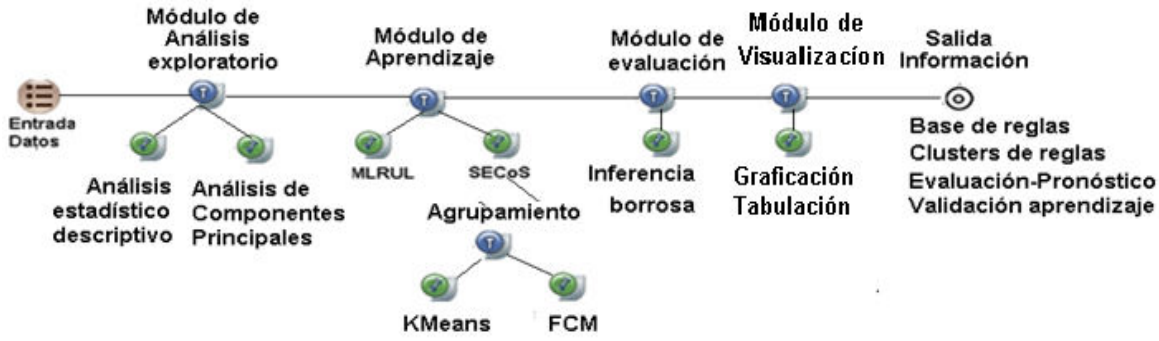


Figura 2.4 Relación entre los módulos del paquete SystDiagnosPrognosxHQ(fuente: elaborado por el autor)

A continuación se describirán con detalle las características de los algoritmos de los módulos de Análisis exploratorio, Aprendizaje y de Evaluación:

2.5.1.2 Módulo de análisis exploratorio

En este módulo se realiza una exploración de la vista minable, mediante análisis de tipo estadístico descriptivo se determinan medidas de tendencia central y de dispersión de la muestra.

Con el fin de explorar la existencia de relaciones entre variables, se realizan análisis componentes principales y de búsqueda de comunalidades.

Para realizar el análisis clásico factorial, SystDiagnosPrognosxHQ utilizó la función Matlab "**factoran**"

Factoran devuelve la matriz λ de comunalidades, la varianza explicada por cada factor, los parámetros de λ rotados convenientemente y un conjunto de datos estadísticos relacionados con la prueba de hipótesis sobre las dimensiones de la matriz mencionada.

Este módulo incluye el algoritmo de análisis de testores típicos BT (Lias, Pons, 2009) que busca reducir la dimensión de la vista minable al detectar, en muchos casos, las variables relevantes.

Existe la posibilidad de incluir otras técnicas estadísticas en este módulo.

2.5.1.3 Módulo de aprendizaje de SystDiagnosPrognos

Este módulo realiza aprendizaje de tipo supervisado, para ello se busca relacionar la matriz con los datos de las variables predictoras I con la matriz O de salidas deseadas para las variables de clasificación.

I (de dimensiones n x m) y O (de dimensiones k x m) son matrices transpuestas que provienen de la vista minable representada en la tabla 2.3.

Entre la variedad de software de aprendizaje automatizado se seleccionaron, en esta etapa de la investigación, los algoritmos SECoS y MLRuL basados en la generación de reglas borrosas.

2.5.1.3.1 Algoritmo SECoS (Watts., 2004)

SECoS (ó Simplified Evolving Connectionist System) (Watts., 2004) (40) es una versión simplificada de EFuNN (Evolving Fuzzy Neural Networks) (Kasabov, N. 1998), por lo que aprende de manera idéntica, pero sin las limitaciones de poseer funciones de membrecía fijas que restringen a esta última Red.

SECoS es un algoritmo basado en instancias, que según la literatura consultada genera bases de reglas con tamaño razonable, aunque esto depende lógicamente de la estructura de los datos a procesar. De cualquier forma, el entrenamiento se realiza en una sola iteración, por lo que son muy útiles para aplicaciones en tiempo real.

Componentes relacionadas con el trabajo del algoritmo SECoS:

- Capas del Sistema SECoS.
- Generación de reglas borrosas.
- Módulo de Aprendizaje.
- Algoritmo de aprendizaje.
- Algoritmo de extracción de una base de reglas inicial.
- Algoritmo de optimización de la base de reglas.
- Módulo de evaluación.

Desarrollo:

Capas del Sistema SECoS:

SECoS posee 3 capas de neuronas:

1.- Capa de entrada.

2.-Capa evolutiva o constructiva, en la que se agregan neuronas y tiene lugar el aprendizaje. La activación de neuronas evolutivas se basa en la distancia entre los vectores de pesos de la capa evolutiva de entrada y el vector de entrada actual.

3.-Capa de salida.

Dentro del conjunto de entrenamiento, cada vector de entrada I posee n componentes (I_1, I_2,…,I_n) y cada vector de salida O, posee k componentes (O_1,O_2,…,O_k).

Generación de Reglas Borrosas en SECoS.

Se pueden generar reglas tipo MIMO (múltiples entradas - múltiples salidas) o MISO (múltiples entradas - única salidas) y contempla formatos de reglas tipo Tagaki –Sugeno grado 0 y tipo Mamdani.

Cada regla r posee dos neuronas, una en la capa evolutiva y la otra en la capa de salida.

Como las redes neuronales SECOS están totalmente conectadas, el número de conexiones dirigidas hacia una neurona de la capa evolutiva es igual al número n de neuronas de entrada,

entonces el correspondiente vector de pesos WI, tiene la misma dimensión que el vector de entrada hacia la capa evolutiva.

A su vez, el vector de pesos WO de las neuronas de la capa de salida posee la dimensión k de los vectores de salida.

Cada componente de los vectores de pesos WI, se interpreta como una premisa y cada componente de WO representa una consecuente de la regla dada.

Todas las premisas se representan mediante variables lingüísticas (Ver Anexo 1), definidas en el conjunto universo U_1,…, U_n (ó $U \equiv U_1 x U_2 x \ldots x U_n \subset R^n$. De ahí que las etiquetas (o términos borrosos) que nombran los valores numéricos de las variables I_i equivalen a conjuntos borrosos F_i en U con sus correspondientes funciones de membrecía $\mu_{F(Ii)}$ [6].

El conjunto de variables consecuentes está definido en el conjunto universo V_1,…,V_k (ó $V \equiv V_1 x V_2 x \ldots x V_k \subset R^k$). Para las reglas de tipo Sugeno grado 0, estas variables toman valores constantes específicos por cada una (y_1^r, y_2^r,…, y_k^r) y que provienen en este caso del entrenamiento de la red neuroborrosa.

Así, una regla borrosa r del tipo Sugeno grado 0 tipo MIMO, tendrá la estructura:

- IF x_1 is F^r_1 and…and x_n is F^r_n THEN $y_1=c_1$ and ... and $y_k=c_k$ (3)

Para un valor numérico x_i, la expresión "x_i is F^r_i" tendrá el grado de veracidad igual al grado de pertenencia de x_i al conjunto borroso F^r_i con i=1,2,…,n

Por tanto: "x_1 is F^r_1 and…and x_n is F^r_n" expresa el grado de pertenencia de la tupla $x\equiv(x_1,x_2,\ldots,x_n)$ a la composición de particiones borrosas F^r_1 ,…,F^r_n (expresa como la tupla x se caracteriza en la composición de variables lingüísticas definidas en los universos U_1, U_2,…, U_n) y se calcula en la implementación realizada, mediante la T-norm:

- $\mu_{F_1^r x \ldots x F_n^r}(x) = min\{\mu_{F_1^r}(x), \ldots, \mu_{F_n^r}(x)\} = \mu_{A^r}(x)$ (4)

Aunque existen otras variantes de cálculo incluso más sensibles.

Cómo ya se dijo, la salida de la regla r será la tupla $y^r \equiv(y_1^r, y_2^r, \ldots, y_k^r)$

En el proceso de Desfuzzificación, los valores fuzzy resultantes de las Reglas son transformados de nuevo en valores numéricos reales (13)

Sean y^1, y^2,…, y^M las salidas de una base de M reglas generadas a partir de la entrada de la tupla x, la salida final tendrá la forma:

- $y_j = \frac{\sum_{r=1}^{M} y_j^r \cdot (\mu_{A^r}(x))}{\sum_{r=1}^{M} (\mu_{A^r}(x))}$ (5)

Donde y_j *e* y_j^r son las j-ésimas componentes de las tuplas $y\equiv(y_1, y_2, \ldots, y_k)$ e $y^r \equiv(y_1^r, y_2^r, \ldots, y_k^r)$ respectivamente. La primera tupla ofrece las salidas desfuzzificadas por

[6] Se refiere a: $\mathrm{GAUSSMF}(X, [\sigma_i, x_{oi}]) = e^{-\frac{1}{2}\left(\frac{x-x_{oi}}{\sigma_i}\right)^2}$

componentes de toda la base de M reglas, la segunda tupla devuelve la salida borrosa de cada regla r en sus k componentes.

Descripción del algoritmo de aprendizaje, (Watts., 2004):

El algoritmo de aprendizaje SECoS se basa en acomodar dentro de la capa evolutiva nuevos ejemplos entrenados ya sea modificando a la vez todas las conexiones de la neurona correspondiente de la capa evolutiva o adicionando una nueva neurona a la capa.

Pasos del algoritmo:

- Paso1 Propagar el vector de entrada I a través de la red.
- Paso 2 Encontrar la neurona j más activada y su activación Aj. (La activación A de una neurona n de la capa evolutiva está determinada por la ecuación: A_n=1-D_n)
- Paso 3 Si $A_j < S_{Umbral}$, entonces
    - Agregar una neurona

  Sino
    - Evaluar el error entre el vector de salida calculada (Oc) y el vector de salida deseada (Od)
- Paso 4 Si $|Oc - Od| > E_{Umbral}$, entonces
    - Agregar una neurona

  Sino
    - Actualizar las conexiones de la neurona ganadora de la capa evolutiva
- Paso 5 Repetir para cada vector de entrenamiento.

Cuando se adiciona una neurona, su correspondiente vector de pesos de conexión es el conjunto del vector de entrada I, y su vector de salida es el conjunto del vector de salida deseada Od.

Los pesos de las conexiones desde cada entrada i hacia la neurona ganadora j se modifican de acuerdo a la ecuación

$$w_{ij}(t+1) = w_{ij}(t) + \mu_l . I_i - w_{ij}(t-1) \quad (6)$$

Donde:

- $w_{ij}(t)$ es el peso de la conexión desde la entrada i hasta j al momento t.
- I_i es el i-ésimo componente del vector de entrada I.

Los pesos desde la neurona j hasta la salida O se modifican por la ecuación:

$$w_{jo}(t+1) = w_{jo}(t) + \mu_l . A_j . Eo \quad (7)$$

Donde: Eo= Od -Ao

Algoritmo para la extracción de una base de reglas borrosas inicial.

Mediante un algoritmo que busca la mejor función de membrecía para cada etiqueta de las variables lingüísticas, se extrae una base de reglas inicial desde SECoS, posteriormente se debe

refinar esta base mediante un algoritmo evolutivo que realice modificaciones de los parámetros de las funciones de pertenencia de modo que se minimice el error en el aprendizaje.

Las funciones de membrecía pueden ser de cualquier tipo.

El software contempla la inclusión de reglas a partir del criterio de los expertos, una solución inicial para la extracción de reglas difusas es la siguiente (40):

- **Paso 1**: Por cada neurona de la capa evolutiva h:
    - Crear una nueva regla.
- **Paso 2**: Para cada neurona de entrada i
    - Encontrar la función de membrecía (mf) asociada con i activada más fuertemente por el peso $w_{i\,h}$.
- **Paso 3**: Adicionar esa función de membrecía al antecedente de la regla para esa entrada (Esa es la función de membrecía para esa entrada, para esa regla)
- **Paso 4**: Insertar el grado de membrecía de los pesos de la función de membrecía ganadora como el grado de importancia para esa condición.
- **Paso 5**: Para cada neurona de salida O
    - Encontrar la función de membrecía asociada a O activada más fuertemente por el peso $w_{h\,o}$.
- **Paso 6**: Adicionar esa función de membrecía a la consecuente de la regla para esa salida (Esa es la función de membrecía para esa salida en esa regla)
- **Paso 7**: Inserte el grado de membrecía de los pesos de la función ganadora como el grado de importancia para ese consecuente.

En esta investigación y debido al empleo de reglas del tipo Sugeno grado 0, se sustituyen los pasos 5, 6 y 7 por la asignación a la consecuente de la regla el valor de la constante clasificatoria almacenado en las neuronas de la capa de salida, esto provoca una reducción del número de operaciones del algoritmo.

Algoritmo de optimización de la base de reglas borrosas

La literatura consultada (40) reporta que en ocasiones, las reglas extraídas pueden resultar imprecisas debido al no ajuste de las particiones del espacio de entrada con las regiones de Voronoi definida para cada neurona (Ver figura en el anexo 6), en ese caso, es preciso elevar la precisión de las reglas mediante algún procedimiento (Watts., 2004).

Si se toma la base de regla extraída anteriormente como una solución inicial, puede realizarse el posterior ajuste de las reglas difusas empleado un algoritmo genético diseñado para optimizar los parámetros de las funciones de membrecía(Cromosomas) regla por regla o de toda la base de regla en conjunto(enfoque de Pittsburg) (40).

En iteraciones alternas, se ajustan las extensiones y los centros de las funciones de membrecía, esto debe considerarse si de realizar interpretaciones se trata.

Número de reglas, Agrupamiento o Combinación de reglas borrosas

El número de reglas que se genere depende estrechamente de la cantidad de premisas y consecuentes que se consideren y de la entropía de la información en sí.

Las reglas que poseen similares consecuentes podrían combinarse utilizando disyunciones y conjunciones lógicas, de esa manera se reduciría el número de reglas de manera significativa, Como se trata de procesamiento de información del proceso docente educativo, es importante preservar lo más que se pueda a las características particulares de cada sujeto del diagnóstico.

Para facilitar la descripción y posterior intervención pedagógica, se privilegiará la opción de agrupar las reglas (Los individuos) según el grado de similitud.

Con el agrupamiento de la base de reglas, se podrá visualizar a todos los individuos con la mayor cantidad de cualidades conocidas.

2.5.1.3.2 Otros algoritmos

Está prevista la incorporación al modelado del algoritmo generador de reglas borrosas MLRuL, "que construye particiones sucesivas por medio de reordenamiento del conjunto de entrenamiento hasta concluir con una partición donde queden agrupados los casos con propiedades similares" (41) y que son representados mediante una regla borrosa.

Aunque para la selección del atributo más relevante puede utilizarse cualquiera de los criterios de selección de rasgos, Pedro Yovanis sugiere el uso del criterio MLRelevance (41).

Este algoritmo puede trabajar con datos numéricos y/o simbólicos y posee muy buenos indicadores de aprendizaje.

2.5.1.4 Módulo para la evaluación de reglas de SystDiagnosPrognos

El módulo de evaluación permite validar la eficiencia del aprendizaje y la realización de predicciones, lo primero se alcanza evaluando las premisas de las reglas creadas y comparando la proporción de aciertos entre los consecuentes calculados y los consecuentes de la regla correspondiente a dichas premisas.

La evaluación de reglas fue descrita en el sub epígrafe de la generación de reglas del epígrafe 2.4.1.3.1

2.5.2 Diseño de la Evaluación

Se aplicará Validación Cruzada cuando se utilicen técnicas de inteligencia artificial. Será seleccionado el experimento sobre el cual se realicen mejores predicciones en relación con el porciento de aciertos.

2.5.3 Construcción del modelo y su evaluación

El procesamiento de la vista minable representada por la tabla 2.3 se ilustra mediante el siguiente diagrama de actividad:

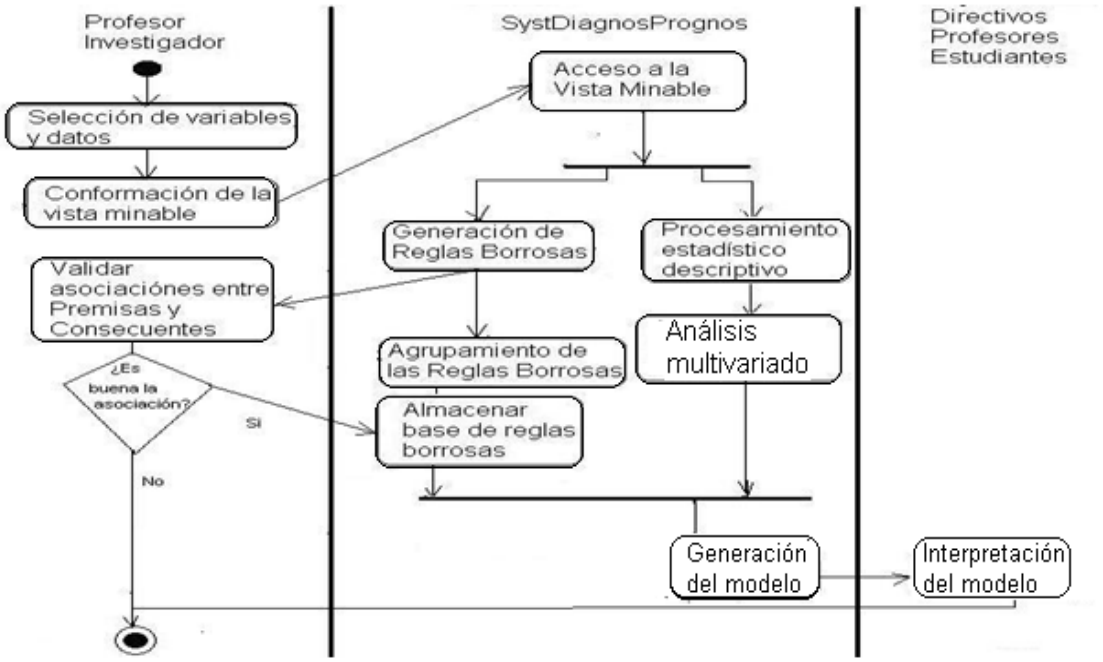


Figura 2.5 Diagrama de actividad de SystDiagnosPrognos propuesto por el autor.

El siguiente esquema representa los diferentes flujos de trabajo de SystDiagnosPrognos y que se relacionan con la naturaleza de los datos a procesar.

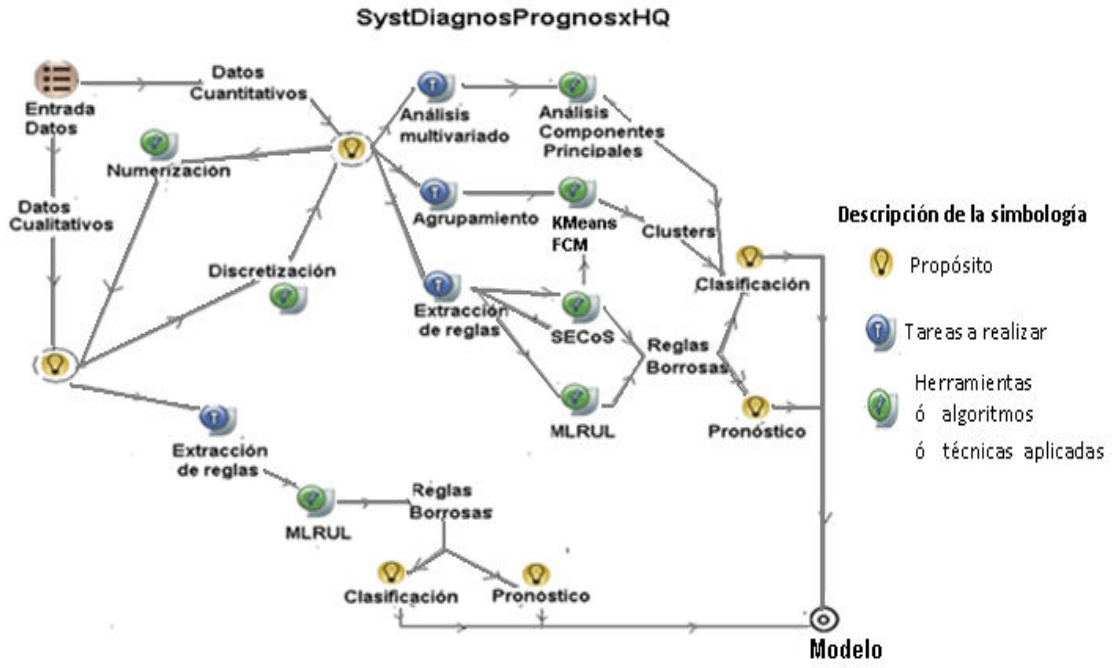


Figura 2.6: Flujo de trabajo con los datos en el paquete SystDiagnosPrognosxHQ (fuente: elaborado autor)

En los anexos 7 y 8 se muestran los diagramas de casos de uso y de clases del sistema.

2.5.3.1 Modelo de la vista minable

En este epígrafe se presentan los modelos que socializarán la información procesada en la corrida de cada uno de los algoritmos que integran el paquete SystDiagnosPrognos y que integran una estructura socializada en el epígrafe 2.0.

Descripción de las componentes del modelo

La aplicación de estas componentes tiene por fin facilitar el cumplimiento de los objetivos trazados en la fase 2.1 de la aplicación de la metodología CRISP-DM

El reporte estadístico descriptivo permite obtener información general sobre la población estudiada de manera igual a la práctica generalizada.

La unión de gráficos de dispersión permite la visualización de relaciones entre parejas de variables.

Las tablas de componentes principales categóricos, factorial y de testores típicos dan cumplimiento al objetivo 2.

La estructura de los C clúster de reglas borrosas se muestra en la tabla 2.5:

Sean n y k las cantidades de variables predictoras y clasificadoras respectivamente.

Para N grupos del primer tipo de atributos y Q grupos del segundo resulta:

$n=n_1+n_2+\ldots+n_N=n$ y $k= n_{N+1}+\ldots+n_Q$;

Los términos r_i, representan el radio de agrupamiento entre las reglas que conforman un clúster

($r_i \in R: r_i \geq 0; 1 \leq i \leq c$)

En el Anexo 9 se ilustra una representación de los C clusters de reglas borrosas generados por el algoritmo de agrupamiento K-Means.

Comentarios

- Las etiquetas $\boldsymbol{etiq_{G,i,j}}$ representan términos borrosos de la variable lingüística correspondiente al atributo predictor j del grupo de variables G ubicado en el clúster i;
- Los términos denotados por $\boldsymbol{clase_{p,q,s}}$ representan al valor clasificador de la variable q, perteneciente al grupo de variables p en el clúster s.
- Analizando por columnas, cada clúster representa un promedio de reglas de la forma:
    - IF x_1 is etiq $_1$ and…and x_n is etiq $_n$ THEN y_1=clase 1 and ... and y_k =clase k (8)
- Cada clúster tendrá características únicas de ahí que se pueda acompañar con comentarios específicos.
- El radio de agrupamiento indica el grado de compactación de los elementos integrantes de un clúster en torno a su centroide.

Los clusters de reglas borrosas tienen muchas ventajas:

Provocan una reducción de la dimensión al sustituir la totalidad de instancias por una cantidad mucho menor de centroides de las mismas, esto contribuye a la socialización e interpretabilidad de la información.

Debido a que las reglas borrosas relacionan grupos de atributos diferentes y el resto como variables clasificadoras puede asociarse información abundante y variada

2.6 Fase 5: Evaluación del modelo

En este momento se desea determinar si los modelos creados en la fase anterior son útiles a las necesidades del negocio por ello se proponen las siguientes tareas:

- Evaluación de resultados sobre todo, desde el punto de vista de la eficiencia, con respecto a los procesamientos tradicionales.
- Revisar el proceso
- Establecimiento de los siguientes pasos o acciones.

Desarrollo

Tarea: Evaluación de resultados.

El modelo propuesto integra armónicamente un grupo de técnicas de minería de datos que permiten el reconocimiento de patrones en cuanto a:

- La utilización de un modelo multivariado más apto para la descripción de procesos complejos*.
- La inclusión de las estadísticas descriptivas tradicionales con vistas a la obtención de información a una escala general.
- Buscar relaciones entre grupos de variables así como entre grupos de instancias todo ello provoca una reducción de la dimensión*
- El empleo de reglas borrosas cosa que además de facilitar la interpretación por su similitud con el razonamiento humano permiten preservar e identificar a los atípicos debido al empleo de algoritmos de agrupamiento*.
- El empleo de herramientas de visualización*.
- Ahorro de tiempo y recursos de computo*
- Formación de una cultura investigativa de los problemas educacionales al dotar a los profesores de una herramienta de cálculo que facilita el procesamiento de datos.
- Impacto social del uso de este tipo de modelo que puede incluir cualquier cantidad de datos y amplios tipos de investigaciones.*

Si se calcula la eficiencia del procesamiento y de su modelo resultante en el sentido de la ecuación:

$$E=P/R \text{ (Chiavenato, 2004),} \qquad (9)$$

Donde P representa a los productos resultantes y R a los recursos utilizados y teniendo en cuenta que los productos nuevos o con mejoras se encuentran marcados con asterisco en el párrafo superior y que los recursos utilizados comprenden el uso de una PC, que una parte de los datos utilizados provienen de un gestor académico, que otros provienen de encuestas o de otras vías, puede notarse un incremento en el valor de E en comparación a la práctica tradicional vigente.

2.7 Fase 6: Despliegue

En caso de satisfacerse los objetivos del negocio se propone desplegar el modelo a las instancias docentes encargadas de la toma de decisiones en la Universidad de las Ciencias Informáticas, con vistas a mejorar el proceso docente educativo.

Con tal fin deben cumplirse las siguientes tareas:

- Planificación de despliegue mediante la inserción de el sistema SystDiagnosPrognos como herramienta del gestor de datos Akademos.
- Planificación de la monitorización y del mantenimiento
- Generación de informes finales por parte de profesores simples, jefes de colectivos de asignatura y de jefes de año.
- Socialización y análisis por los profesores y dirigentes docentes.

2.8 Conclusiones del Capítulo

En este capítulo se describió la propuesta general del modelo para el análisis de datos del proceso de formación del profesional.

- Se adaptó la metodología CRISP-DM para la elaboración del modelo de análisis de datos.
- En la generación del modelo se aplican técnicas estadísticas descriptivas exploratorias multivariadas para la búsqueda de relaciones entre variables y entre grupos de instancias y técnicas de inteligencia artificial.
- Las técnicas de inteligencia artificial utilizadas en el modelo, se basan en la generación de reglas, por un sistema híbrido neuro borroso, que potencia la descriptibilidad, la interpretación y el pronóstico.
- En comparación con otros tipos de modelos del proceso docente, este es más amplio y efectivo por las características de los datos que procesa, por los pocos recursos que demanda, por los patrones que devela y por las posibilidades de socialización e interpretación.

En el siguiente capítulo se aplicará este modelo al análisis de datos reales de los estudiantes de primer año de una facultad.

3 Validación del modelo integrado para el análisis de datos educacionales

En este capítulo se compara cual modelo de análisis de datos de procesos formativos permite sustentar mejor la toma de decisiones: el tradicional o el modelo creado con técnicas de minería de datos educacionales.

La construcción del modelo para el análisis de datos se realizará mediante el procesamiento de la información real que se logró recoger en el primer semestre del curso 2009-2010

3.1 Lugar de aplicación

La Universidad de las Ciencias Informáticas posee siete facultades en su sede central más tres facultades regionales. En cada facultad tienen lugar la formación del estudiante mediante la docencia, producción e investigación dosificada en los ciclos básico y profesional.

El presente estudio tomará como muestra un segmento de la población de la Facultad 1, integrado por estudiantes del ciclo básico, del primer año de la carrera y se pretende realizar una ampliación del modelo para el análisis de los datos educacionales con vistas a fortalecer la gestión de la información educacional y sustentar mejor la toma de decisiones.

En la siguiente tabla se describe la composición de la muestra:

Tabla 1: Sexo y Centro de procedencia de los estudiantes que componen la muestra

Sexo		Centro de procedencia					
F	M	IPVCE	Técnico Medio Informática	EMCC	DEPORTE	IPUEC	IPU
58	82	36	55	3	4	32	8

Puede observarse el predominio de la cantidad de hombres sobre las mujeres, que la cantidad de técnicos medios en informáticas y de graduados en IPUEC duplica a la de los graduados de IPVCE, esta desproporción puede indicar dificultades en el aprendizaje (lo que realmente sucedió) lo que exige de un trabajo serio del colectivo pedagógico.

3.2 Análisis de la toma de decisiones con el modelo de análisis de datos elaborado en el curso 2009-2010

Constatada la necesidad de manejar información educacional relevante, a inicios del curso 2009-2010, se elaboró un modelo para el análisis de datos que incluía tres tipos de reportes digitales para primer y cuarto año:

- La Caracterización del estudiante.

- Estadísticas de una facultad.
- Estadística de la UCI.

Comentarios sobre el modelo a través de sus componentes

El primer tipo de reporte contiene los datos generales de cada estudiante y los resultados de diversos cuestionarios relacionados con el nivel sociocultural y motivacional, su cultura general e ideo política, el dominio de la lengua materna y el inglés, la capacidad intelectual, dominio técnico integral, la motivación por roles, los conocimientos técnicos generales, la matriz de fortalezas y debilidades así como los test de auto percepción de Belbin (Belbin, 1981) y de inventario de estilos de decisión para un total de 170 variables.

Este documento resulta ser extenso, se observó que un ejemplar ocupaba trece páginas.

En cuarto año se elaboraron 326 reportes de este tipo

Por su parte, los reportes con las estadísticas de la Facultad 1 y de los estudiantes del primer año en la Universidad, resumían en 13 y 14 páginas, las distribuciones de frecuencias absolutas y porcientos de las caracterizaciones individuales de los estudiantes de dicha facultad y la cohorte de la UCI, respectivamente.

Del análisis de dichos documentos puede concluirse lo siguiente:

- La información aparece dispersa en documentos diversos y pasa bruscamente de un nivel de generalización mínimo a uno máximo.
- El empleo de estadísticas descriptivas en los reportes destinados a las facultades y la universidad ocultan la diversidad de tipos de individuos, hacen parecer que existiese un único tipo de estudiante.
- Los reportes se vuelven más generales a medida que se sube por la escala jerárquica de mando, esta limitación de la información obstaculiza la toma de decisiones en correspondencia con los paradigmas de la educación cubana.
- No se dividen las variables en predictivas o clasificadoras, es difícil encontrar relaciones entre variables o entre individuos que hagan rico el análisis y que permitan tomar decisiones educacionales.
- Una vez comenzado el semestre, no resulta fácil integrar la información diagnóstica disponible con la información tácita de cada profesor o directivo. Se necesita crear esa posibilidad.
- Es una necesidad vincular la información disponible actual y pasada con la futura para ver la variación de los parámetros en el tiempo.
- La cantidad de reportes no integrados y la forma de procesamiento de la información, no puede facilitar la colaboración entre los especialistas.

A continuación y teniendo en cuenta la propuesta del capítulo Dos, se realizará la elaboración de un modelo para el análisis de los datos reales de los procesos formativos de los estudiantes del primer año de la Facultad 1 que permita superar las limitaciones expresadas arriba.

3.3 Aplicación del modelo

Para elaborar el modelo, se tomaran como bases de datos las siguientes fuentes:

- Diagnóstico de Perfiles de Inteligencia (García,E.,2009).
- Diagnóstico de los Estilos de Aprendizaje (García,E.,2009).
- Diagnóstico de las estrategias de aprendizaje (García,E.,2009).
- Primer y Segundo Cortes Evaluativos (Akademos, UCI, 2009).
- Resultados de los exámenes finales del semestre (Akademos, UCI, 2010).

Y se adoptó el siguiente objetivo:

Caracterización del avance académico de los estudiantes de primer año de la facultad 1 de la UCI teniendo en cuenta la información recogida en las fuentes citadas arriba.

Con el fin de garantizar la calidad del proceso de minería de datos sobre la información diagnóstica, se decidió seguir las fases adaptadas de CRISP-DM, detalladas en el Capítulo 2.

3.3.1 Fase 1: Comprensión del negocio en el procesamiento de los diagnósticos realizados a los estudiantes de primer año

Contexto Inicial

En la introducción de este trabajo, se hace mención de la necesidad en la universidad de trabajar con información educacional relevante que permita planificar adecuadas estrategias interventoras.

Durante el semestre se hizo evidente la falta de un modelo socializable que integrara la información diagnóstica inicial con los resultados que en cada etapa aportaba el proceso docente educativo y que propiciara la toma de decisiones pertinentes según la evolución de los resultados.

Objetivos del negocio

Se desea investigar la posible relación entre las características y habilidades de la personalidad con el avance académico estudiantil.

De esa manera, se podrá:

- Analizar las variables que mejor explican la información disponible para determinar cuáles estudiantes, desde el punto de vista de sus características, modifican su rendimiento académico.
- Conocer las fortalezas y debilidades del desarrollo estudiantil, que deben considerarse, para orientar el proceso de intervención educacional.
- Elaborar una base de reglas que describa las características fundamentales de la población muestreada y que pueda ser utilizada para el pronóstico y la evaluación durante el semestre y otros cursos.

- Contribuir a la socialización de la información a través de la elaboración de documentos en pequeños formatos.

Evaluación de la situación:

Se consideró que se disponen de los datos suficientes para la realización del procesamiento, se utilizará el software prototípico **SystDiagnosPrognos** descrito en el capítulo 2 que trabaja con buenos niveles de aprendizaje y predicción sobre datos de prueba.

Objetivos de la minería de datos

- Buscar relaciones entre variables y entre individuos.
- Generar una base de reglas borrosas y representarla de manera sintetizada mediante clusters de reglas para lograr describir de forma compacta las características de las posibles asociaciones entre datos pertenecientes a las variables predictivas y los datos ubicados en las variables consecuentes.
- Clasificar los tipos fundamentales de las instancias y pronosticar la evolución futura.
- Posibilitar el pronóstico a partir de la evaluación de premisas en la base de reglas.

Criterios de éxito

Dada la cantidad de premisas y consecuentes que puede tener una regla borrosa, serán considerados los siguientes criterios de éxito:

- Que se logren detectar correlaciones fuertes entre las diferentes variables que componen el proceso siempre y cuando se cumplan las premisas estadísticas.
- Alcanzar una adecuada reducción de la información.
- Que se alcance una tasa de Aprendizaje (asociación entre premisas y consecuentes) superior al 90% sobre los datos de entrenamiento y sobre 65% en las estimaciones sobre los datos de prueba (datos reservados no para el entrenamiento sino para evaluar el aprendizaje con fines de prueba), considerando que en la base de datos existan casos ambiguos o dudosos que no se les dio tratamiento.

Plan de Procesamiento: Está previsto realizar el análisis de las posibles combinaciones entre los atributos con el fin de generar reglas borrosas.

3.3.2 Fase 2: Comprensión de los datos relacionados con la información recabada

El proceso de familiarización con los datos se inició con la recopilación de los mismos. En el Anexo 10 se describen las ubicaciones y formatos de las bases de datos fuente y destino.

En esta investigación se optó por el empleo de un libro Excel como base de datos receptora de la información que recibirá el procesamiento, por los siguientes motivos:

- Por realizarse una investigación exploratoria de las potencialidades de los algoritmos de procesamiento, no se elabora en este momento una herramienta.
- La amplia difusión y empleo del Excel en el medio docente universitario.

- Los gestores de datos contemplan la generación de la información en tablas Excel.
- Sus bondades para el tratamiento y representación de datos.
- La capacidad de guardar la información a partir de Excel en 24 formato diferentes.
- Su comunicación con software en Matlab y con herramientas como Weka y SPSS.
- La unión de las diversas bases de datos se realizó mediante una Macro de Visual Basic.
- No se crearán bases de datos con más de 264 atributos.

Tarea.- Descripción de los datos:

Se conformó una matriz con la información organizada por: número de lista, nombre y apellidos del estudiante, los indicadores sobre sus perfiles de inteligencia, estilos y estrategias de aprendizaje junto a los resultados del primer y segundo corte evaluativos del primer semestre.

Tarea.- Exploración de los datos

En la nueva base de datos, se encontraron omisiones en algunas celdas, pero en un porciento pequeño, estos errores provenían de las fuentes originales.

Los perfiles de inteligencia, las estrategias y estilos de aprendizaje, se clasificaban como cuantitativos en escalas de intervalo, las evaluaciones del primer corte por asignaturas de primer año y para la diferencia entre dos cortes consecutivos eran del tipo cualitativo ordinal.

También se diferenciaban los recorridos de las variables numéricas.

Tarea.- Verificación de la calidad de datos.

Los diagnósticos que se aplicaron contaron con la aprobación de especialistas competentes, y con el aval de haber sido utilizados en diferentes investigaciones de forma exitosa.

El resto de la información tiene carácter oficial, es empleada en los análisis educacionales y proviene del gestor de datos Akademos de la Universidad de las Ciencias Informáticas.

La aplicación de los diagnósticos se efectuó a través de la herramienta LymeSurvey en condiciones apropiadas de tiempo y vigilancia con el fin de no contaminar la muestra.

Por todo lo anterior se considera que la calidad de los datos es buena.

3.3.3 Fase 3: Preparación de los datos

Tarea.- Selección de los datos

Para cumplir los objetivos trazados en la fase 1 de la metodología, se construyó una vista minable cuyos atributos generales agrupados en Perfiles de Inteligencia, Estilos y Estrategias de aprendizaje, resultados del primer corte evaluativo del semestre se utilizaron como predictores del avance académico resultante de la comparación de los dos cortes evaluativos consecutivos.

Tarea.- Limpieza de datos

En algunos casos se descartaron aquellas instancias con datos incompletos, en otros casos, se completaron las celdas vacías con los valores medios promedios alcanzados por los atributos.

Quedó constituida una matriz de 141 instancias y 36 atributos, de los cuales los últimos 7 constituían clasificadores del avance estudiantil al considerarse el posible avance académico entre dos cortes evaluativos consecutivos.

Tarea.- Construcción, integración y formateo de datos

En virtud de que los atributos correspondientes a los grupos de los perfiles de inteligencia, estilos y estrategias de aprendizaje recogen puntuaciones reales con diferentes recorridos, se decidió realizar una estandarización de las mismas, dividiéndolas por la máxima puntuación posible en cada atributo.

De esta manera, el producto por 100 de estas razones, expresaría el porciento de madurez en estos atributos con que arribaron los estudiantes tras su incorporación a la universidad.

Las evaluaciones de ambos cortes docentes fueron numerizadas según el criterio: B=3, R=2, M=1 y NE=0.

Para representar el avance académico del estudiante entre dos cortes evaluativos, se creó un nuevo grupo de variables linealmente dependiente: La matriz de diferencias entre ambas evaluaciones parciales, los nuevos atributos tomaron valores representados por las etiquetas: Retrocede mucho= 0, Retrocede=1, Nulo=2, Avance=3 y Avanza Mucho=4.

Esta información se incorporó en el libro Excel mencionado en el epígrafe precedente para ser procesado por un algoritmo prototípico de minería de datos elaborado en Matlab.

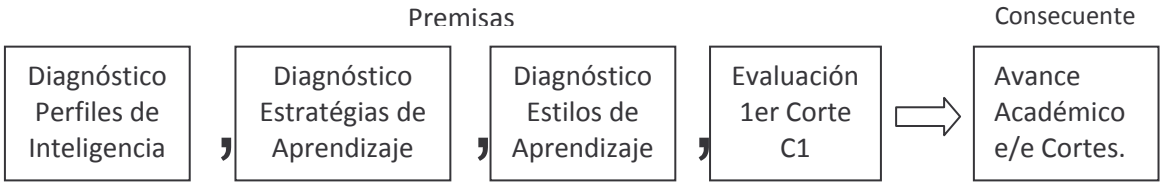
3.3.4 Fase 4: Modelado

3.3.4.1 Tarea: Selección de la técnica de modelado

Aunque las técnicas que se aplicarán en el modelado se describen en el epígrafe 2.2.4 es conveniente comentar la estructura general de las reglas borrosas que se generarán en el procesamiento de la información ya que son el corazón del modelo propuesto.

La relación entre los resultados de los diagnósticos aplicados al primer año y el avance estudiantil entre dos cortes evaluativos consecutivos mediante reglas borrosas del tipo Sugeno grado 0 se representará de la siguiente forma:

Figura 3.1 Estructura semántica de las reglas vista por sus grupos de variables.



A cada atributo premisa le corresponde una variable lingüística representada mediante etiquetas.

En correspondencia con la última tarea del epígrafe 3.3 los resultados de los diagnósticos iniciales serán representados mediante las etiquetas (aunque sean números) 0; 0,25; 0,50; 0,75 y 1.

Las evaluaciones de los cortes conservan la misma denominación nominal ordinal usual.

Comentarios:

•Los valores numéricos 0, 0.25, 0.50, 0.75 Y 1.0, corresponden a las medias de los conjuntos borrosos Suma: Baja, Media-Baja, Media, Media–Alta y Alta.

•Las puntuaciones por atributo se presentan como una razón entre 0 y 1 (cociente de la puntuación alcanzada por el de estudiante y el máximo de puntos posible asignado al atributo) para dar una noción del grado de desarrollo.

•Los valores numéricos inferiores en alguna habilidad, reflejan características del desarrollo en un momento dado, no se pueden tomar como inamovibles.

•Las etiquetas B, R, M y NE expresan la evaluación alcanzada por el estudiante en el primer corte evaluativo por cada asignatura, para el procesamiento, se realizó una numeración asignándoles los valores numéricos 3,2,1 y 0 respectivamente, suponiendo su pertenencia a una escala de Lickert.

Atributos clasificatorios o consecuentes

Para clasificar el avance académico estudiantil en siete asignaturas entre dos cortes evaluativos consecutivos, se definieron cinco clases o etiquetas posibles: Retrocede mucho (- -) = 0, Retrocede (-) =1 , Nulo (Nul) = 2, Avanza (+) = 3 y Avanza mucho (+ +) = 4

Diseño de la Evaluación

Para evaluación de los resultados de la aplicación de la minería de datos se debe valorar la capacidad del modelo obtenido en la descripción, clasificación, predicción, el uso de la información educacional y la socialización de la información, en el epígrafe 3.5, se le asignarán valores al resultado del procesamiento que se realiza en este capítulo.

3.3.4.2 Construcción del modelo y su evaluación

A partir de la construcción de la vista minable detallada en el epígrafe 3.3.3, se elaboró un modelo de representación de esa información mediante la aplicación de las técnicas identificadas en el epígrafe 2.4.4.1, agrupadas en el algoritmo SystDiagnosPrognos ya descrito.

Este modelo tiene las siguientes componentes:

- Componente analítico descriptivo
- Gráficos de dispersión multivariados
- Análisis tras la generación de testores típicos.
- Análisis de componentes principales
- Análisis clásico factorial
- Componente basado en la generación de reglas borrosas y agrupamiento.

Desarrollo:

3.3.4.2.1 Componente analítico descriptivo:

En el anexo 11, se muestran tablas que resumen los reportes estadísticos-descriptivos que pertenecen a esta componente.

Comentarios:

Puede observarse que la población estudiada presenta predominio de puntuaciones altas y medias en casi todos los atributos excepto en algunos que se relacionan con las estrategias de aprendizaje y con los estilos de planificación, cooperación e independencia.

El análisis de la información de los dos cortes evaluativos refleja los mayores problemas en las asignaturas Matemática 1 y Matemática discreta.

Componentes del análisis multivariado:

3.3.4.2.2 Gráficos de dispersión multivariados

En el Anexo 12 se ilustra mediante un ejemplo, los tipos de gráficos de dispersión 2D generados mediante la comparación de diversos grupos de variables y que contribuyen a garantizar visualmente la posible existencia de relaciones entre sus variables. Esta facilidad permite aventurar supuestos que sustenten la aplicación de otras técnicas estadísticas así como la detección de atípicos.

3.3.4.2.3 Análisis tras la generación de testores típicos

Para la aplicación del algoritmo BR (Lías, Pons, 2009) para la generación de testores típicos, se realizaron transformaciones sobre la vista minable (de 140 filas y 38 columnas), para convertir todos los datos numéricos en cualitativos nominales.

El promedio de los avances académicos en un grupo de grupo de asignaturas básicas permitió clasificar el avance en tres clases diferentes. Esto generó una matriz de diferencias de 3255 filas y 38 columnas cuando se consideraron los promedios de avances entre los dos cortes como variables clasificatorias.

En consecuencia, la matriz básica alcanzó dimensiones de 3248 filas y 38 columnas.

El resultado del procesamiento fue un testor típico (Ver anexo 13) compuesto por 31 variables o atributos y que se identifican en el anexo electrónico e7 por su numeración y denominación:

Se considera que por las características de los datos no hubo una fuerte reducción de la dimensión por lo que no se utilizará este resultado en el modelo.

3.3.4.2.4 Componente del análisis clásico factorial

A partir de la aplicación **Factoran** sobre la vista minable se describen 10 factores ó dimensiones latentes (ver Anexo 15) a partir del procesamiento de 140 estudiantes evaluados con 54 variables.

En la tabla 3 se ilustran las variables que más contribuyen a la varianza común explicada por los cuatro primeros factores (ó dimensiones latentes) de la vista minable

Tabla 3 Variables más representativas de los cuatro primeros factores en explicar la varianza común

Factores	Variables	Cargas fact.	Variables	Cargas fact.	Variables	Cargas fact.	Variables	Cargas fact.	Variables	Cargas fact.
1	De Ensayo**	0,7	De Reflexión Activa**	0,7	Muy Dependiente de ayuda**	0,6	Control Motivacional**	0,6	Comprensión Analítica**	0,8
2	Musical*	0,6	Corporal Quinestésica*	0,6						
3	1erC1_EF1	0,7	1erC1_IP.	0,6	1erC1_M1	0,7	1erC1_PHCC	0,7	1erC1_PP1	0,74
4	1erC1_SN	0,7	Avances_SN	-1						

Comentarios

- El primer factor está compuesto por cinco de variables relacionadas con las ocho estrategias de aprendizaje identificadas, esto significa que debe adoptarse una estrategia de enseñanza especial con estos estudiantes, que tienen limitaciones en su control motivacional y en la consulta a documentos escritos.
- El segundo factor habla de las actitudes artísticas de la muestra, hay un potencial para la realización exitosa de actividades culturales.
- El tercer factor habla del impacto de la dimensión “Primer corte evaluativo” sobre los estudiantes, la no aparición de la asignatura Matemática discreta en este factor, puede significar una evaluación no ajustada de los profesores a la realidad.
- El cuarto factor habla de que debe revisarse la motivación en la impartición de la asignatura Seguridad Nacional pues se acusa un descenso académico en ella.

3.3.4.2.5 Componente basado en la generación de reglas borrosas y agrupamiento

Para permitir la valoración de las potencialidades de esta componente se generaron tres sub modelos basados en el agrupamiento de reglas borrosas realizados por el algoritmo SECoSxHQ, del software SystDiagnosPrognos en los que se varió la composición de los grupos de variables premisas y consecuentes de la siguiente forma:

Sub modelo 1 **Variables predictivas**: Perfiles de inteligencia, Estilos y Estrategias de aprendizaje y las Notas del primer corte evaluativo; **Variables que clasifican**: Avance académico entre cortes.

Sub modelo 2 **Variables predictivas**: Perfiles de inteligencia, Estilos y Estrategias de aprendizaje, Notas del primer y segundo corte evaluativo; **Variables Clasificadoras** Resultados en los exámenes finales del semestre.

Sub modelo 3: **Variables predictivas**: Notas del primer y segundo corte evaluativo; **Variables Clasificadoras**: Avance académico entre cortes evaluativos.

Cada submodelo contiene reportes del año completo y de las brigadas que los componen (Ver Anexo Libro de Reportes).

Con el primer submodelo, SECoSxHQ realizó una asociación de los grupos de variables descritos, del 99,9% para el conjunto de instancias de entrenamiento y del 83,7% para el conjunto de instancias utilizadas en la prueba de aprendizaje en la primera corrida y del 100% para el conjunto de entrenamiento y del 97.14 para los casos de pruebas para el segundo submodelo.

Se generaron y almacenaron sendas bases de reglas borrosas para propósitos descriptivos y predictivos y se construyeron sendas tablas con los agrupamientos de las reglas y que constituyen los sub reportes a socializar.

Submodelo 1

Los centroides de los nueve clusters (Tabla 4) de estudiantes clasificados en base a la similitud de sus características permiten diferenciar de manera clara a las personas de acuerdo a sus fortalezas y debilidades. En la tabla 5 aparecen los identificadores de los estudiantes que integran cada uno de los doce clusters con vista a facilitar intervención docente desarrolladora.

Este submodelo se recoge en la hoja Corr1 del Anexo Libro de Reportes, allí se incluyen estimaciones a nivel de clusters que permiten decidir cual variable predictiva modificar, mediante acciones planificadas de los profesores y la junta de año, de modo que se experimenten resultados positivos en las variables clasificadoras.

3.3.4.2.6 Descripción del avance estudiantil del año completo, entre cortes evaluativos

Tabla 4 Caracterización del aprendizaje entre cortes evaluativos

Atributos \ Clusters de reglas:	1	2	3	4	5	6	7	8	9
Perfil Intelig.LingVerb	0,64	0,68	0,64	0,82	0,61	0,74	0,71	0,61	0,84
Perfil Intelig. Lógica-Matemática	0,66	0,72	0,66	0,68	0,59	0,44	0,72	0,73	0,96
Perfil Intelig.Visual Espacial	0,69	0,84	0,68	0,73	0,69	0,84	0,76	0,75	0,66
Perfil Intelig.Musical	0,69	0,72	0,73	0,88	0,66	0,84	0,75	0,73	0,42
Perfil Intelig.Corporal-Quinestésica	0,66	0,66	0,69	0,83	0,61	0,66	0,72	0,62	0,72
Perfil Intelig.Intra Personal	0,67	0,56	0,69	0,84	0,67	0,7	0,72	0,67	0,8
Perfil Intelig.InterPersonal	0,73	0,82	0,66	0,84	0,76	0,88	0,78	0,81	0,7
Perfil Intelig. Naturalista	0,49	0,3	0,46	0,63	0,52	0,7	0,52	0,54	0,16
Estrategia Ensayo	0,47	0,8	0,58	0,75	0,85	0,9	0,64	0,6	1
Estrategia Reflexión Activa	0,5	0,85	0,6	0,69	0,75	0,8	0,68	0,78	0,95
Estrategia Búsqueda Ayuda	0,53	0,6	0,58	0,79	0,8	0,6	0,75	0,67	0,9
Estrategia Consulta Documentos	0,46	0,9	0,59	0,7	0,9	0,5	0,64	0,7	1
Estrategia Aplicación en la Práctica	0,16	0,3	0,21	0,24	0,3	0,2	0,22	0,27	0,4
Estrategia Control Emocional	0,38	0,6	0,49	0,45	0,9	0,7	0,51	0,6	0,6
Estrategia Control Motivacional	0,47	1	0,53	0,7	0,95	0,8	0,64	0,73	0,6
Estrategia Control Comprensión	0,49	1	0,61	0,72	0,9	0,8	0,69	0,53	0,9
Estilo Aprend. Percepción Visual	0,7	0,8	0,74	0,73	0,65	0,9	0,82	0,83	1
Estilo Aprend. Percepción VerbAud	0,43	0,3	0,49	0,48	0,45	0,3	0,49	0,37	1
Estilo Aprendizaje Global	0,59	0,9	0,66	0,62	0,8	0,7	0,67	0,6	0,6
Estilo Aprendizaje. Analítico	0,65	0,8	0,71	0,55	0,5	0,7	0,74	0,73	0,9
Estilo Aprendizaje. Planificado	0,77	1	0,87	0,87	0,9	1	0,88	0,83	0,9
Estilo Aprendizaje Espontáneo	0,38	0,4	0,45	0,41	0,65	0,4	0,45	0,3	0,6
Estilo Aprendizaje Cooperativo	0,59	0,3	0,69	0,82	0,7	0,6	0,67	0,6	0,9
Estilo Aprendizaje Independiente	0,4	0,9	0,43	0,42	0,45	0,6	0,45	0,37	0,5
1er C1 EF.I.	M	B	B	M	B	B	B	B	M
1er C1 IP.	M	M	R	B	M	B	B	B	B
1er C1 M.I.	M	M	M	M	M	M	R	B	B
1er C1 MD.I.	B	B	M	B	B	B	B	B	B
1er C1 PHCCU.	M	B	B	B	B	B	B	B	B
1er C1 PP.I.	M	B	R	B	B	M	B	B	B
1er C1 SN.	B	B	R	B	R	R	R	B	B
	↓	↓	↓	↓	↓	↓	↓	↓	↓
AvanceEFI	--	++	++	Nul	++	++	++	+	++
AvanceIP	--	--	-	--	-	++	+	++	++
AvanceMI	--	Nul	-	Nul	-	Nul	Nul	++	++
AvanceMDI	--	--	--	++	Nul	--	Nul	+	++
AvancePHCCI	--	++	++	Nul	++	++	++	++	++
AvancePPI	--	-	Nul	-	Nul	Nul	+	+	++
AvanceSN	Nul	-	-	++	Nul	+	+	+	++
Radio agrupamiento	**0,77**	**1,41**	**1,68**	**1,13**	**1,59**	**0**	**1,61**	**1,28**	**0**
Cant. Instancias	**4**	**4**	**19**	**2**	**8**	**1**	**64**	**37**	**1**

Leyenda sobre la clasificación del avance académico:

Retrocede mucho (- -); Retrocede (-); Nulo (Nul); Avanza (+); Avanza mucho (+ +)

La figura 3.3 permite comparar visualmente la evolución de los clusters estudiantiles a través de las distintas evaluaciones o test aplicados.

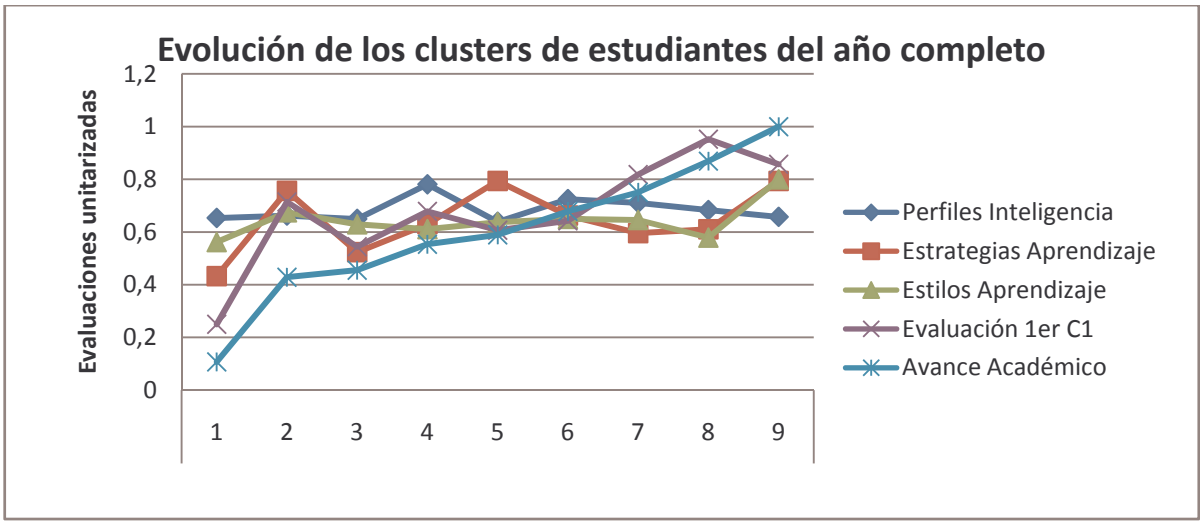


Figura 3.3 Comportamiento de los grupos de estudiantes según las medias de los grupos de variables procesadas

Tabla 5. Identificadores (números de lista) de los integrantes de cada cluster.

Clusters de reglas:	1	2	3		4	5	6	7					8			9
Cantidades	4	4	19		2	8	1	64					37			1
Número de identificación en el listado general (Id.), del año completo.	10	31	3	108	118	32	21	4	28	68	95	128	1	56	113	131
	52	38	7	112	134	35		5	30	69	98	135	2	60	119	
	62	72	11	133		61		9	34	70	102	136	6	67	120	
	100	81	22	139		121		13	37	77	104	138	8	73	127	
			39			124		14	42	78	107		12	75	129	
			40			125		15	43	79	109		25	76	137	
			45			130		16	46	80	110		29	88	140	
			48			132		17	47	82	111		33	89		
			53					18	51	84	114		36	93		
			58					19	57	85	115		41	96		
			71					20	59	86	116		44	97		
			74					23	63	90	117		49	99		
			83					24	64	91	122		50	101		
			87					26	65	92	123		54	103		
			106					27	66	94	126		55	105		

En el Anexo 16 se muestran indicadores que permiten comparar los grupos de estudiantes entre sí

Comentarios

- Los clusters de la tabla 4 se encuentran ordenados de menor a mayor atendiendo en primer lugar a los valores de las variables clasificadoras y en segundo lugar a los valores de las variables predictivas con el fin de eliminar empates y facilitar la interpretación de los resultados.
- El ordenamiento descrito permite identificar instancias atípicas en los extremos de la tabla 4 y también en su interior.
- El ordenamiento explica la tendencia creciente en cada uno de los grupos de variables representados.
- Los nueve centroides de la tabla 4, expresan diferentes grados de aprendizaje, los clusters del uno al cinco representan a 27 estudiantes con situación crítica en varias asignaturas1 y que poseen los menores valores, de manera general, en las puntuaciones de los diagnósticos realizados
- Los clusters seis y siete representan a 67 estudiantes con buen avance académico, pues su evaluación mínima es regular en una asignatura, no obstante a ello pueden verse calificaciones insuficientes en algunas características diagnosticadas al inicio del curso por lo que se pudieran programar acciones en su mejora.
- Los clusters ocho y nueve representan a estudiantes con gran avance académico y que no serían objeto de análisis en una junta de año (al igual que los de los clusters seis y siete), sin embargo, pueden ser programárseles actividades que ayuden a elevar aspectos diagnosticados como insuficientes.
- Mediante la evaluación de los atributos de las reglas borrosas (Ver Anexo Libro de Reportes), se predijo que:
    - Si se eleva el nivel de la Inteligencia Lingüístico Verbal, se elevará el avance académico en los clusters uno, dos, cuatro, cinco y seis.
    - Si se eleva el nivel de la inteligencia Lógica Matemática, se elevará el avance académico en los clusters cuatro, cinco, seis, siete y ocho
    - Si se eleva el nivel de la inteligencia Corporal Quinestésica, se elevará el avance académico en el cluster siete.

A continuación se comentará los resultados del sub modelo 2.

3.3.4.3 Descripción de la evolución de grupos de estudiantes durante el primer semestre (Submodelo2)

En el Anexo 17, se muestra una tabla con once clusters de reglas borrosas, generadas en la segunda corrida de SystDiagnosPrognos, en la figura 3.6 se muestra una síntesis de dicha tabla, que ofrece también información sobre cómo se comportaron en el semestre, los indicadores de los grupos de variables y permite hacer valoraciones holísticas.

Pueden ser analizados fenómenos como la deserción escolar

El grado de validez de las evaluaciones de los cortes.

El trabajo de las asignaturas

Las características que llevan a que los estudiantes alcancen determinados resultados.

Comentarios:

- Los estudiantes con mejores perfiles de inteligencia no alcanzaron los más altos resultados (clúster 11).
- Los segundos grupos de estudiantes en perfiles de inteligencia si llegan a alcanzar la segunda posición en los exámenes finales (clúster uno) precedidos por los últimos estudiantes en perfiles de inteligencia, estrategias y estilos de aprendizaje pero que en ambos cortes se mantuvieron con buenas evaluaciones (clúster cuatro).
- Los clusters Tres, Cinco, Ocho y Diez representan a estudiantes que posiblemente causaron baja docente algunos de ellos fueron aparecen ubicados en los clusters con peor avance docente en las tablas 3.8 y 3.10, en otros casos, aparecen sorpresas en las que sería interesante investigar las causas.

En el Anexo 18 se describe el sub modelo 3

3.3.5 Fases 5 y 6: Evaluación y Despliegue

Tarea: Evaluación de resultados.

A continuación se evaluará el modelo que se elaboró cuando se combinó la información diagnóstica de entrada con los resultados del primer corte evaluativo en la predicción del avance académico de los estudiantes. En la tabla 6 se muestran las evaluaciones alcanzadas por la parametrización de las cualidades del modelo y en la tabla 7 se describen los recursos utilizados en la generación del modelo.

Tabla 6 Evaluación del modelo elaborado a través de los resultados de sus componentes.

Tipo de Evaluación	Componentes	Indicadores	Medida
Reducción de dimensiones, estratificación y Visualización.	Estadística descriptiva	Recorrido, medidas de tendencia Central y de Variación, representación Gráfica.	Una por cada tipo de indicador (36)
Reducción de la dimensión	Análisis Factorial	Cantidad de Factores Latentes	De 2 a 3 Factores por cada uno de los 5 grupos de variables
		Por ciento de la variabilidad explicada con dichas componentes	69%
	Agrupamiento	Cantidad de Grupos	12-11 de 36 posibles
		Grado de similitud entre los Ele-	0-4.5 unidades de

		mentos de un Clúster	39 posibles.
	Sistemas Neuroborrosos	Cantidad de Reglas	De 118 a 140 sobre 140 puntos
		Por ciento de casos bien relacionados a partir de las premisas y los consecuentes.	100%
		Capacidad Pronóstica	71-100%
Estratificación	Análisis ACP	Identificación de casos comunes y no comunes	141 (100%) Muestra
	Agrupamiento	Cantidad de casos cubiertos	94-95%
	Sistemas neuroborrosos	Cantidad de casos cubiertos	100%
Visualización	Análisis Multivariado de Componentes Principales	Calidad de la representación gráfica	100%
		Representación Tabular	Si
	Agrupamiento	Calidad de la representación gráfica	100%
		Representación Tabular	Por desarrollar
	Sistemas neuroborrosos	Calidad de la representación gráfica	100%
		Representación Tabular	Si-No
	Formato del Modelo	Capacidad de Generalización	Si
		Capacidad de Particularización	Si
		Interpretabilidad de la información del Procesamiento	Aceptable
		Capacidad de socialización	Buena, el resultado puede circular entre todos los decisores de forma íntegra.
		Apoyo a la toma de decisiones	Superior al modelo tradicional.
Procesamiento	Modelo	Reutilización de la Información	Si
		Combinación diversas Evaluaciones.	Si
		Búsqueda de asociaciones no lineales entre variables	Si
		Cumplimiento de los objetivos del procesamiento.	Si

Tras el análisis de la tabla 6, puede concluirse lo siguiente:

•Se logró la caracterización de la muestra de estudio en sus cualidades generales y particulares puesto que se aplicaron las siguientes técnicas::

- Se conjugaron las técnicas estadísticas descriptivas habituales con otros métodos de `procesamiento.
- Se generó una base de reglas borrosas, que refleja la información particular con un grado alto de condensación.
- Las reglas permiten tratar información multivariada.
- Se generaron clusters de Reglas Borrosas (agrupan la información por tipos de estudiantes).

•Es posible aprovechar la experiencia docente previa y reutilizar los resultados en próximos procesamientos

- Puede traducirse al formato de reglas borrosas.
- Asociar resultados de investigaciones diferentes sobre los mismos sujetos y poblaciones.
- Las reglas del tipo MIMO permiten realizar las asociaciones.
- Realizar pronósticos.
- La evaluación de las reglas permite estimar resultados a partir de las características muestrales.

•Se pueden socializar los resultados en un formato minimalista que condensa mucha información, ya sea en forma tabular o gráfica.

Tabla 7.- Recursos utilizados en la generación del modelo que incluye información diagnóstica, información del corte evaluativo y que permite predecir el avance académico.

Recursos empleados en la generación del modelo		
Proceso	Tipo de Recurso	Cantidad
Procesamiento de la información	PC	1
	Personal	1-2 personas (Un docente y técnico informático)
	Tiempo	Media jornada laboral

Conclusiones del análisis de la tabla 7:

Son racionales lo recursos utilizados en la generación del modelo y que permiten concluir que se alcanza un ahorro aceptable en correspondencia con el beneficio que se puede obtener.

En la etapa actual de la investigación, no se tiene planificada la realización de despliegue de la información procesada.

3.4 Comparación del antes y el después

A continuación se realizará la comparación del modelo propuesto para el análisis de datos educacionales en esta investigación, con respecto a los cuatro tipos de caracterización vigentes, se considerarán las siguientes dimensiones:

1) Número de técnicas implicadas en la elaboración de cada modelo.
2) Tipos de estudiantes representados en cada modelo.
3) Nitidez de la información descrita en cada modelo.
4) Búsqueda de relaciones entre variables.
5) Búsqueda de relaciones entre individuos.
6) Seguimiento a los procesos evolutivos.

Desarrollo:

1)Número de técnicas implicadas en la elaboración de cada modelo

Producto de la interacción de siete técnicas diferentes en la elaboración del modelo ampliado, ver figura 3.5, se logra enriquecer la calidad de la información resultante en el sentido de alcanzar la estratificación de los datos, la búsqueda de relaciones entre variables y entre individuos, así como la predicción automatizada.

Mientras el modelo tradicional solo incluye distribuciones de frecuencias, en la elaboración del modelo ampliado intervienen las distribuciones de frecuencias, el análisis Factorial, los Testores típicos, el Análisis de correlación canónica, la generación de reglas a través de una red híbrida neuroborrosa, el agrupamiento de reglas borrosas, y la inferencias borrosa.

2)Tipos de estudiantes representados en cada modelo

El modelo tradicional describe a la población través de sus estratos, esto proporciona tendencias generales, esto equivale a describir a un solo tipo de individuo medio o a tres tipos de estudiantes atendiendo a las clases definidas.

El modelo ampliado a través del agrupamiento de clusters borrosos llegó a describir a 9 tipos de estudiantes, ver figura 3.6.

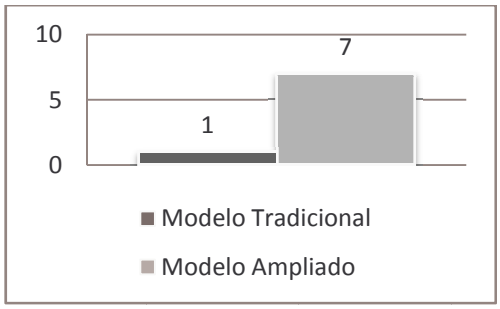


Figura 3.5 Comparación de las cantidades de técnicas aplicadas en la elaboración de cada modelo.

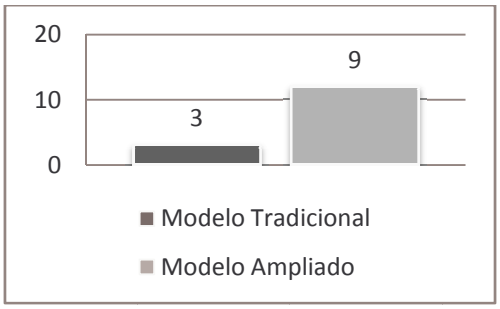


Figura 3.6 Comparación de los números de tipos de estudiantes representados en cada modelo

3)Nitidez de la información descrita en cada modelo.

El modelo ampliado permite caracterizar varios tipos de estudiantes dentro de las categorías con dificultades, con desarrollo medio y alto, esto ofrece una riqueza mayor en la descripción de la población. Ver en la figura 3.7 la comparación con el modelo tradicional.

El nuevo modelo, incluye al mismo tiempo información general y particular simultáneamente y eso se logra a través de la técnica de análisis clúster

El modelo ampliado además, ubica a todos los individuos de la población en el subgrupo de individuos que más se le parece, lo que permite la identificación y facilita la elaboración de estrategias de trabajo personalizadas.

4)Búsqueda de relaciones entre variables.

Si del análisis empírico de las tablas de frecuencia pudieran establecerse algunas relaciones entre variables, el modelo ampliado permite relacionar variables a través del análisis factorial, los coeficientes de correlación canónica y del formato de las reglas borrosas que además de describir a la población, permiten realizar predicciones automatizadas. En la figura 3.8 se comparan ambos modelos.

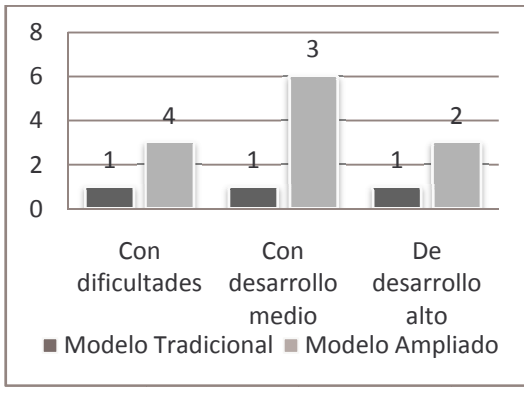


Figura 3.7 Comparación de la nitidez de la información de cada modelo.

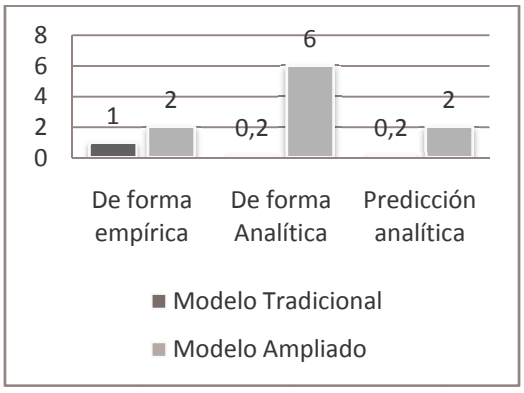


Figura 3.8 Vías para buscar relaciones entre variables.

7) Búsqueda de relaciones entre individuos.

Si en el modelo tradicional es prácticamente imposible relacionar individuos, en el modelo ampliado pueden identificarse, compararse y realizar predicciones sobre individuos no muestreados. Ver figura 3.9

8) Seguimiento a los procesos evolutivos

Debido a que la vista minable tiene formato de una matriz ampliable y a la aplicación de técnicas multivariadas robustas, es posible incorporar nueva información paulatinamente, esto permite la elaboración y actualización constante del modelo a crear dentro de un proceso evolutivo, manejado datos con diferentes escalas.

En la figura 3.10 se compara en este sentido, los modelos tradicional y ampliado.

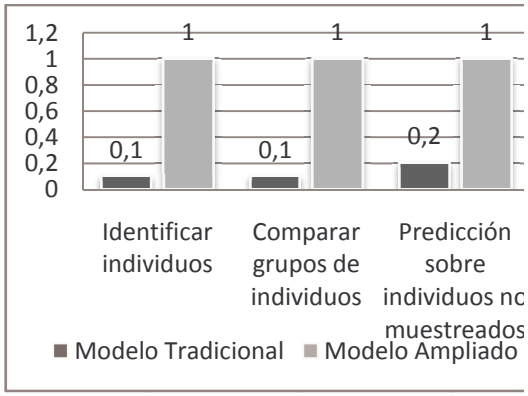


Figura 3.9 Vías para relacionar individuos.

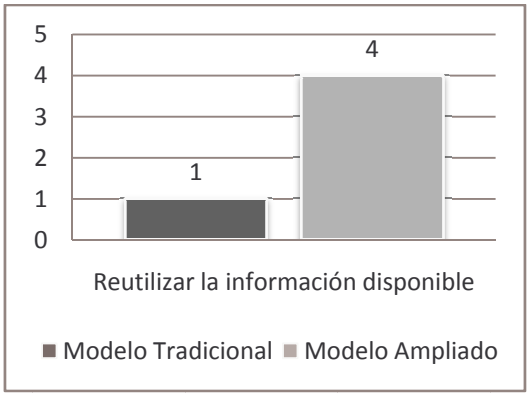


Figura 3.10 Vías para dar seguimiento a procesos evolutivos.

Las bondades descritas arriba contribuyen de manera positiva a la toma de decisiones educacionales.

Este modelo puede ser socializado debido a su formato relativamente breve, de acuerdo a la información que resume, es interpretable por los profesores debido a que muestra los resultados del procesamiento en las variables originales por lo que contribuye a la generación de análisis colectivos en la búsqueda de estrategias de trabajo.

Conclusiones del epígrafe

Si en el modelo existente coexisten tres tipos de reportes (la caracterización del estudiante y los reportes estadísticos de la facultad y la universidad) con numerosa cantidad de ejemplares, ahora se pueden sustituir por un solo reporte que contiene dos sub componentes generadas con técnicas de minería de datos.

La primera subcomponente contiene datos estadísticos multivariados para explorar la relación entre variables e incluye gráficos de dispersión y técnicas analíticas de análisis de componentes principales y análisis clásico factorial, que serán eficientes en la medida en que se cumplan los supuestos de normalidad y homocedasticidad.

La segunda subcomponente explora la relación entre individuos a través de la presentación de clusters de reglas borrosas que relacionan grupos de variables con informaciones diferentes ya sea por su naturaleza o por los valores que alcanzaron en determinado período.

Lo anterior presenta las siguientes ventajas:

- Representar en un formato abreviado a toda la muestra con un grado pequeño de distorsión, de manera que se destaquen los tipos fundamentales de individuos, con sus fortalezas y debilidades de manera que se pueda facilitar el análisis de los datos.
- Acompañar dichos clusters con gráficos y tablas que los comparen, así como de los listados de estudiantes que los forman. De esta manera se pueda dirigir el trabajo hacia los individuos como demanda el enfoque pedagógico sociocultural.
- Si antes para poder transitar a los niveles superiores de decisión, los reportes estadísticos se hacían cada vez más generales, por motivos de espacio. Ahora, el reporte basado en el agrupamiento de reglas borrosas que resume la información de toda la muestra, puede transitar por toda la escala de mando sin reducirse, con su información general y particular al miso tiempo. Esto brinda una plataforma para el trabajo colaborativo de los decisores.

Paralelamente a ello, el hecho de crear y almacenar bases de reglas borrosas por el sistema automatizado SystDiagnosPrognos permitió:

- Realizar inferencias clasificatorias mediante la evaluación de las variables predictivas en las reglas borrosas.
- Facilitar la interpretación de las reglas debido a la robustez que genera la evaluación de variables predictivas lingüísticas.
- Como cada regla borrosa puede representar a todos los sujetos con características bastante similares, la base de reglas representa a toda la muestra y su evaluación realiza una estimación de los parámetros poblacionales.
- El análisis de la estructura de las reglas facilita la descripción de la muestra de manera general y a cada individuo en particular al visualizarse la relación que se establece entre todos los datos utilizados en el modelo, destacándose sus debilidades y fortalezas.
- Con ello se pueden construir reportes predictivos de utilidad cuando no se cuente con información diagnóstica actualizada.

Adicionalmente, la utilización de procedimientos automatizados para la generación de los modelos mediante la metodología CRISP-DM permite manejar grandes cantidades de datos con eficiencia en el procesamiento y una estandarización en los reportes.

Por todas estas cuestiones el modelo ampliado se convierte en una plataforma colaborativa de trabajo.

Este tipo de modelo contribuirá al incremento de las investigaciones pedagógicas por parte de los profesores al facilitarle el análisis de datos.

Conclusiones del capítulo

En este capítulo se generó un modelo del proceso docente educativo mediante la aplicación del paquete de algoritmos SystDiagnosPrognos en el procesamiento de una combinación de datos reales provenientes de diagnósticos pedagógicos y de los cortes evaluativos del primer semestre del curso 2009-2010, de los estudiantes de primer año de la facultad 1.

Con el objetivo de demostrarlas cualidades del modelo ampliado para caracterizar y predecir el aprendizaje de la población estudiada, en cuanto a las tendencias generales y a sus particularidades.

Se demostró en este caso, que las técnicas de minería de datos permitieron la generación de un

Modelo para la gestión del conocimiento docente con las características siguientes:

- Incluye los indicadores tradicionales de la estadística descriptiva e inferencial.
- Incorpora técnicas de generación de reglas borrosas, agrupamiento y visualización provenientes del Data Mining, por lo que se pueden establecer relaciones no lineales entre variables.
- Se conservó la información global y la particular de la muestra lo que facilita el seguimiento de las trayectorias estudiantiles en correspondencia con el enfoque histórico cultural, de aquí que es posible brindar a cada tipo de estudiante el tratamiento específico que merece
- Se realizaron pronósticos a partir de la evaluación de la base de reglas neuro borrosas, lo que ofrece un soporte para que las estimaciones no sean únicamente empíricas.
- Se reutilizó convenientemente la información y su cruzamiento.
- El diseño de las reglas neuro borrosas empleadas permitió representar información con múltiples entradas y múltiples salida lo que hace más rico el poder de procesamiento.
- Con respecto a la información original, la salida del procesamiento presenta una dimensión más reducida, una estratificación por grupos de instancias similares y posibilidades de visualización mejores, todo ello, ayuda a la socialización de la información y brinda más elementos para la toma de decisiones, con respecto a la práctica actual.

Estas técnicas no sustituyen el análisis humano, sus resultados indican tendencias, no son categóricas, pero pueden contribuir a elevar el número de investigaciones científicas y el control del proceso docente educativo en un modelo de formación centrado en los aprendizajes.

En Resumen, con la aplicación de las herramientas de minería de datos al procesamiento de la información diagnóstica, se logró elevar la eficiencia del procesamiento del diagnóstico pedagógico con respecto a los procedimientos habituales de la práctica docente.

Conclusiones de la tesis

1. El modelo tradicional de análisis de los datos, basado en la estadística descriptiva, empleado en la Universidad de las ciencias Informáticas tiene que ser ampliado para que pueda sustentar una mejor toma de decisiones.
2. Se diseñó una ampliación del modelo tradicional de análisis de datos educacionales que incorpora técnicas de minería de datos o la estadística inferencial, lo que permitió fortalecer la descripción y la predicción, detectar relaciones entre variables y entre individuos. contar con tablas de formato pequeño, fáciles de socializar y que no sufren deformación al transitar por todos los niveles de la escala de mando. Se identificó y adaptó, la metodología necesaria para la realizar la ampliación del modelo. Fue puesto a punto un software prototípico integrador del conjunto de técnicas.
3. Si el modelo tradicional permitió representar tres tipos de estudiantes en todo el año, atendiendo a los estratos establecidos, el modelo ampliado logró representar nueve tipos de estudiantes, de ellos cuatro tipos de estudiantes con resultados bajos, dos tipos de estudiantes con resultados medios y tres tipos de estudiantes con resultados altos dejándose ver las fortalezas y debilidades de cada clase de estudiante lo que facilita el diseño de acciones educativas apropiadas y personalizadas. Todos los estudiantes se clasifican en los doce tipos obtenidos.
4. El modelo tradicional, por su naturaleza generalizadora tiende a ocultar las individualidades. El modelo ampliado, ofrece una vía para identificar individuos, una vía para poder compararlos y una vía para realizar predicciones sobre individuos no muestreados a partir de la información de la muestra.
5. El modelo ampliado ofrece cuatro vías para reutilizar la información disponible en lugar de la única vía del modelo tradicional, esto permite la inserción del modelo en un proceso formativo y su constante actualización.

RECOMENDACIONES

Se propone la continuación de la presente investigación en las siguientes direcciones:

1. La implementación del algoritmo MLRul en el paquete de algoritmos SystDiagnosPrognos debido a sus buenos resultados en el aprendizaje automatizado.
2. La incorporación al paquete de algoritmos SystDiagnosPrognos de una herramienta visual que permita dar seguimiento a la evolución en el tiempo de una población siguiendo las trayectorias de sus instancias individuales.
3. La realización de más experimentos con los algoritmos y la búsqueda de nuevas funcionalidades o aplicaciones.
4. La transformación del prototipo SystDiagnosPrognos en una herramienta profesional que implemente los algoritmos aquí descritos y que se aplique al procesamiento sistemático de diagnósticos pedagógicos.

REFERENCIAS BIBLIOGRÁFICAS

[1] VALERA SIERRA, René, *"El proceso de formación del profesional en la educación superior basado en competencias: el desafío de su calidad, en busca de una mayor integralidad de los egresados".* Civilizar, Jun 2010, Vol 10, No 18, p.117-134.ISSN 1657-8953.www.scielo.una.edu.co/cgi-bin/wxis/iah?Isis. 2010. Fecha de consulta 5 de enero del 2011.

[2] MES, *Resolución para el trabajo docente metodológico en la educación superior cubana número 210 del 2007*. 2007. Cuba. 2007.

[3] COMISION DE CARRERA DE LA UCI, Vicerrectoría de formación. *Modelo del profesional y Objetivos generales*. Documentos rectores, 2010.

[4] COMISION DE CARRERA DE LA UCI, Vicerrectoría de formación. *Estrategia educativa*. Documentos rectores, 2010.

[5] MARTÍNEZ LEYET Olga Lidia. *Estrategia de caracterización. Libro blanco*. Informe del C.I.C.E, presentado a la Vicerrectoría de formación de la Universidad de las Ciencias Informáticas, 2009

[6] MERCERON Agathe, YACEFF Kalina. *Tada-ed for educational data mining*. Interactive Multimedia Electronic Journal of Computer-Enhanced. Educational data mining: a case study.

http://chai.it.usyd.edu.au/Publications/year?opts=view...]. Fecha de consulta: 12 de abril del 2011.

[8] ARMATTE, Michel. *Sociología e História de la modelización estadística*. EMPÍRIA. Revista de Metodología en Ciencias Sociales. No 3, pp. 11-34, 2000.

[9] FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMITH, P., UTHURUSAMY R; *Advances in Knowledge Discovery and Data-Mining*, AAAI Press / The MIT Press, 1996.

[10] HAIR, J.F; ANDERSON, Jr. R. E.; TATHAN, *"Análisis Multivariante"*, Madrid, Prentice Hal, 1999.

[11] GRAS,R.; SUZUKI,E.;GUILLET,F.;SPAGNOLO,F,Statistical Implicative analysis. Theory and Applications. London:Springer, 2008.

[12] Shulcloper, J.R.; ALBA CABRERA,E.;LAZO CORTÉS,M."Introducción a la teoría de testores típicos".Serie verde 50,CINESTAV-IPN. 1995.

[13] DEL BRIO, Bonifacio Martín; SANZ MOLINA. Alfredo. *"Redes neuronales y sistemas difusos"* (Español) Martín del Brío, B. / México, Alfaomega, 2da. ed., 399 p. **Registro:** MON-001313; **Formato:** Impreso; En línea; Electrónico; **ISBN** 970-15-0733-9. 2005.

[14] VIEIRA, José.; MORGADO Fernando.; MOTA Alexandre. " *Neuro-Fuzzy Systems: A SurveyWSEAS* Transactions on Systems, Issue 2, Vol. 3, pp. 414-419, 2004.

[15] NAUCK, D.; KLAWON, F.; KRUSE, R. *"Foundations of Neuro-Fuzzy Systems"*, J. Wiley and Sons. 1997.

[16] ROMERO, C.; VENTURA, S. Educational Data Mining: *A Survey from 1995 to 2005*. *Expert Systems with Applications 33*, 125-146.), 2007.

[17] BAKER, Ryan; YACEF, Kalina; *The State of Educational Data Mining in 2009: A Review and Future Visions Vol 1*. Issue 1. EDM 2010. 2010.

[18] ROMERO, Cristobal; VENTURA, Sebastián; ESPEJO, Pedro G.; HERVÁS, Cesar. *Data Mining algorithms to clasiffy students.* educationaldatamining.org. 2008.

http://sci2s.ugr.es/keel/pdf/specific/congreso/Data%20Mining%20Algorithms%20to%20Classify%20Students.pdf. Fecha de consulta 20 de Abril del 2009.

[19] D'MELLO, Sidney.; GRAESSER, Art.. *Mining Bodily Patterns of Affective Experience during Learning*. EDM 2010 Submission 26. 2010.

[20] SHIH, B.; KOEDINGER, K.; SCHEINES, R. Discovery of Student Strategies using Hidden Markov. http://www.ml.cmu.edu/current_students/DAP_shih.pdf, 2010.

[21] SHANABROOK, David. H.; COOPER, David G.; PARK WOOLF, Beverly.; ARROYO, Ivon. *Identifying High-Level Student Behavior Using Sequence-based Motif Discovery* . EDM 2010 Submission 15. 2010.

[22] KOTSIANTIS, S.B.; PINTELAS, P.E. *Predicting students marks in Hellenic Open University*. Conference on advanced Learning Technologies. IEEE, 2005. pp. 664-668. 2005.

[23] KRISTJANSSON, A. L; SIGFUSDOTTIR, I. G.; ALLEGRANTE, J. P. "*Health Behavior and Academic Achievement Among Adolescents: The Relative Contribution of Dietary Habits, Physical Activity, Body Mass Index, and Self-Esteem*", Health Education & Behavior, (In Press) *Health Education & Behavior* Vol. XX (X) - xx-xx. DOI: 10.1177/1090198107313481. 2010.

[24] CORTEZ, P.; SILVA, A. "*Using Data Mining To Predict Secondary School Student Performance*", In EUROSIS, A. Brito and J. Teixeira (Eds.), pp.5-12. 2008.

[25] MIHAESCU, Cristian; DAN BURDESCU, Dumitru;MOCANU, Mihai; IONASCU, Costel. *Obtaining knowledge using educational data Mining*, Intelligent distributed computing III,SCI 237, pp. 257-262. Springerlink@Springer-Verlag berlin Heidelberg, 2009

[26] FERNÁNDEZ VERDÚ, Ceneida; LLINARES CISCA, Salvador. *Relaciones entre el pensamiento aditivo y multiplicativo en estudiantes de educación primaria. El caso de la construcción de la idea de razón.* Horizontes Educacionales, Vol. 15, Nro. 1: 11-22,2010.

[27] MAY.; LIU, B.; WONG, C.K.; YU, P.S.; LEE, S.M. "*Targeting the Right Students Using Data Mining*", Proceedings of KDD, International Conference on Knowledge discovery and Data Mining, Boston, USA, 2000, pp. 457-464. 2000.

[28] HIZAJI, S. T. and NAQVI, R. S. M. M. "*Factors Affecting Student's Performance: A Case of Private Colleges*", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[29] KHAN, Z. N. ; *"Scholastic Achievement of Higher Secondary Students in Science Stream"*, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[30] RAMASWAMI, M. A-CHAID-Based-Performance-Prediction-Model-in-Educational-Data-Mining. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010. www.IJCSI.org. http://www.scribd.com/doc/26416533/15 .2010.

[31] DOMINGUEZ, Anna Katrina.; YACEF, Kalina.; CURRAN, James R. *Data Mining for Individualised Hints in eLearning.* EDM 2010 Submission 34, 2010.

[32] BOUCKAERT, Remco R., FRANK, Eibe et al, *WEKA Manual for version* 3-6-0. University of Waikato, Hamilton. New Zealand. Diciembre 18, 2008.

[33] DURÁN, E.;COSTAGUTA, R. *Minería de datos para descubrir estilos de* aprendizaje, Revista Iberoamericana de Educación. OEI. ISSN: 1681-5653, Número 42/2. 10- 03 -07. http://www.rieoei.org/1674.htm, 2008.

[34] MARTÍNEZ N., FERREIRA G., GARCÍA Z. *El Razonamiento Basado en Casos en el ámbito de la Enseñanza/Aprendizaje*. Revista de Informática Educativa y Medios Audiovisuales Vol. 5(10), págs. 25-33. ISSN 1667-8338 © LIE-FI-UBA. liema@fi.uba.ar, 2008.

[35] GONZÁLEZ, Ernesto. *Técnicas de Minería de Datos*, Monografías.com. Accesible en http://www.monografias.com/trabajos55/mineria-de-datos/mineria-de-datos.shtml. Fecha de consulta Enero 2009, 2008.

[36] BRITO SARASA, Raycos., ROSETE SUÁREZ, Alejandro, ACOSTA SÁNCHEZ, Rolando. *"MINERÍA DE DATOS APLICADA A LA GESTIÓN DOCENTE DEL INSTITUTO SUPERIOR POLITÉCNICO JOSÉ ANTONIO HECHEVARRIA"*. Reporte de investigación. Evento Informática 2009.

[37] TOLEDO RIVERO, Viviana.; PÉREZ PÉREZ, Daisy M.; PÉREZ GONZÁLEZ, Albert.; PÉREZ GONZÁLEZ, Jalbert. *Experiencias con SAED: Sistema informático para la autoevaluación de estudiantes a distancia*. Informática 2010. La Habana. 2010.

[38] RIOS, Lydia Rosa, LEZCANO BRITO, Mateo. *Ambiente Inteligente de enseñanza aprendizaje para Prolog.* VIR 171. III Taller Internacional la Virtualización en la Educación Superior. Universidad 2010. Memorias del Evento. ISBN 978-959-16-1164-2. 2010.

[39] FACULTAD 1, CENIA. UCI. *Gestor de datos: Akademos*. UCI, 2005.

[40] WATTS, M.J. *ANN Rule Extraction using Evolutionary Programmed Fuzzy Membership Functions*. In: International Journal of Information Technology: Special Issue on Evolutionary Algorithm and Advanced Learning System (2005) 11(10).2005

[41] PIÑERO, Pedro Yovanis. Un modelo para la ayuda a la toma de decisiones basado en técnicas de SOFTCOMPUTING. BFMC 2006; 7(1): 11-24, 2006