Universidad de las Ciencias Informáticas

Facultad 5

Título.

Análisis para la predicción del éxito o fracaso académico de estudiantes de la Universidad de las Ciencias Informáticas mediante la teoría de conjuntos aproximados.

Trabajo de diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autora: Neyvis Remón González

Tutor: MSc. Roberto Millet Luaces

Cotutor: Ing. Vladir Antonio Parrado Cruz

"Lo que hace crecer al mundo no es el descubrir como está hecho, sino el esfuerzo de cada uno para descubrirlo."

José Martí

DATOS DE CONTACTO:

Tutor: Roberto Millet Luaces

Breve currículo:

• Profesor de Matemática.

• Graduado de Ingeniero Eléctrico en 1986, en Universidad de Camagüey.

• Profesor Auxiliar

• MSc en Ciencias Matemáticas.

• Imparte docencia en universidades desde 1987.

Ubicación: UCI, Cuba.

E-mail: milletp@uci.cu

Agradecimientos:

A mi mamá Roselia y a mi papá Humberto por su apoyo y confianza cuando he estado lejos.

A mi hermana por ser tan linda y tan buena conmigo.

A mis abuelos Celeste, Miguelito, Eugenia y Mongué que aunque no esté, sé que se sentiría muy feliz en este momento.

A mi novio Vladir, por su ayuda diaria para que yo sea mejor y porque lo quiero mucho.

A mi tutor Millet, por su guía y confianza.

A mis tíos, primos y en general a toda la familia que no cabe en esta página.

A mis amigas Yady, Gretchen e Idalmis que son lo máximo.

A Alcides por toda su ayuda.

A todos mis maestros y profesores por enseñarme lo que sé.

A la revolución por esta escuela tan linda.

A Francisca por facilitarme los datos que necesitaba.

Al profesor Villanueva por su ayuda.

A todo aquel que ocupó, ocupa u ocupará un lugar en mi corazón.

Neyvis Remón González.

Dedicatoria:

A mi mamá Roselia y a mi papá Humberto como un regalo.

A Vladir como una gota de la lluvia de regalos que me faltan por darle en la vida.

A mi hermana para que siga como va.

Resumen

La teoría de conjuntos aproximados se muestra al mundo como una parte más de la Inteligencia Artificial a principios de los años ‘80s. En general se utiliza para problemas de Minería de datos. Uno de ellos es la selección de rasgos, que en la teoría de conjuntos aproximados toma el nombre de reducto. Su costo en tiempo de ejecución es alto. El presente trabajo investiga cuán útil es la teoría de conjuntos aproximados en la predicción académica de estudiantes y se utiliza un método para calcular reductos que mejora la calidad de la predicción. Se presentan, además, los resultados obtenidos usando el conjunto de datos correspondiente que contiene las muestras de los datos del primer semestre de primer año de más de 600 estudiantes de la facultad 5 de la Universidad de las Ciencias Informáticas de los cursos 2006-2007, 2007-2008, 2008-2009.

Palabras clave: Inteligencia Artificial, Minería de datos, selección de rasgos, conjuntos aproximados, reducto.

Índice de contenido

Índice de tablas

Índice de Figuras

Introducción

Actualmente se genera cada día una gran cantidad de información, algunas veces conscientes de que se hace y otras veces inconscientes de ello; por el desconocimiento. Se genera información cuando se registra la entrada en el trabajo, cuando se entra en un servidor para ver el correo o cuando se sigue alguna navegación por Internet.

La generación de estas grandes cantidades de datos ha llevado a la realización de estudios que apoyan y contribuyen a los sistemas de predicción. Estos pronósticos pueden ser de gran utilidad en la toma de decisiones en empresas, instituciones y sus procesos.

Ciertamente, esta gran cantidad de datos genera problemas. Un problema es el procesamiento de los datos. Las computadoras han contribuido a su disminución; pero no lo soluciona del todo. Mucha de la información generada no es válida para todo tipo de análisis, esto significa que hay muchos datos redundantes e innecesarios. La disminución de estos datos puede ser, a veces una solución, o al menos reduciría el cómputo y con ello el tiempo en obtener un resultado y que el mismo sea factible. La idea principal es determinar qué datos seleccionar.

En muchos campos se han obtenido soluciones de este tipo, como por ejemplo: en el deporte para selección de talentos y estudio de contrarios, en la salud en el pronóstico de enfermedades, en la economía específicamente para la gestión de la calidad, en la educación en la predicción académica; que garantice el desarrollo futuro y el desempeño de estudiantes ante disímiles tareas.

El análisis de la predicción académica escolar se constituye de extraordinaria importancia dentro del sistema educativo de enseñanza. Aproximarse al éxito o fracaso universitario como objeto de estudio plantea entender su complejidad y su comprensión como un fenómeno multifactorial.

Algunas investigaciones sobre predicción académica se han realizado en varias partes del mundo, como por ejemplo en la universidad de Complutense de Madrid sobre “La

predicción del rendimiento académico: regresión lineal versus regresión logística"de los investigadores Visitación García Jiménez, Alvarado Izquierdo y Amelia Jiménez Blanco (1), en la Universidad de Murcia (España) sobre "Predicción del rendimiento académico en alumnos de ESO y Bachillerato mediante el Inventario Clínico para Adolescentes de Millon" por Miguel Ángel Broc Cavero y Carmen Gil Ciria (2), además "Cinco mitos sobre la inteligencia y el talento" (Colombia, 1986 a 2006) por: Julián De Zubiría Samper (3).

En Cuba también se han desarrollado investigaciones sobre predicción académica; en la Escuela Latinoamericana de Medicina (ELAM), el trabajo" Variables psicosociales y su relación con el desempeño académico de estudiantes de primer año de la Escuela Latinoamericana de Medicina" por los investigadores Carlos Alberto Román y Yenima Hernández Rodríguez (4) y en la Facultad de Ciencias Médicas "Salvador Allende" el trabajo "Resultados diferenciales de la prueba diagnóstica sobre gráficos según procedencia de educación media superior" de la doctora Sonia Damiani Cavero (5).

Investigaciones sobre predicción académica no se pueden generalizar, por la dependencia del lugar donde se realiza, debido a que cada institución tiene sus particularidades, tanto de asignaturas claves en el proceso docente como en la composición del claustro, el estudiantado y otros factores externos que pueden influir en una decisión final.

Como se mencionó anteriormente, en este proceso hay mucha cantidad de datos históricos almacenados y muchas variables que medir.

A la Universidad de las Ciencias Informáticas arriban estudiantes de todas las provincias y municipios del país, desde todo tipo de centros de procedencia. Estos tienen padres con nivel escolar, actividad laboral y ocupación laboral heterogéneas. La cantidad de hembras y varones son proporcionales.

En la Universidad de las Ciencias Informáticas el plan de estudio está concebido de tal forma que los proyectos productivos e investigativos juegan un papel importante en el desarrollo de habilidades cognoscitivas de los estudiantes. Por las características de esta universidad se hace necesario seleccionar estudiantes capaces de enfrentar tareas productivas de gran envergadura. Hasta el momento no se ha realizado en el

centro ningún estudio que garantice la predicción acedémica ni investigaciones que tributen a esta.

Una posible solución sería una correcta selección de datos que realmente aporten información relevante sobre los estudiantes.

Precisamente estas son cuestiones de las que se encarga la IA, preparación de los datos y procesamiento “inteligente” de los mismos.

Por lo anteriormente planteado el **problema a resolver** sería:
¿Cómo obtener los rasgos fundamentales que contribuyan al diagnóstico del rendimiento académico de estudiantes de la Universidad de las Ciencias Informáticas, aplicando la teoría de los conjuntos aproximados?

Objetivo general:
Identificar los rasgos fundamentales que contribuyan al pronóstico del éxito o fracaso de estudiantes de la Universidad de las Ciencias Informáticas, aplicando la teoría de los conjuntos aproximados.

Objeto de estudio:
El proceso de obtención de los principales rasgos para el pronóstico del éxito o fracaso mediante la teoría de los conjuntos aproximados.

Campo de acción:
Aplicación de la teoría de conjuntos aproximados en la obtención de los principales rasgos para el pronóstico del éxito o fracaso de estudiantes de la Universidad de las Ciencias Informáticas.

Tareas investigativas

1. Revisión de la bibliografía relacionada con las aplicaciones de la teoría de los conjuntos aproximados, para la exploración del estado del arte.
2. Estudio de la teoría de los conjuntos aproximados, para su aplicación en la investigación.

3. Estudio del proceso de realización del pronóstico del rendimiento académico, para su aplicación en la investigación.
4. Aplicación del método Delphi de la consulta de expertos, para la validación de las variables a tener en cuenta.
5. Estudio del software WEKA, para la aplicación a la base de casos de algoritmos implementados en este software.
6. Estudio del asistente estadístico STATGRAPHICS, para la realización del análisis estadístico de la información obtenida mediante el método Delphi de la consulta de expertos.
7. Implementación de un método de búsqueda utlilizando la teoría de los conjuntos aproximados para obtener los rasgos que influyen en el pronóstico del rendimiento académico de los estudiantes en la Universidad de Ciencias Informáticas.

Para la realización de este documento se utilizan los métodos científicos que se enuncian a continuación:

Métodos Teóricos:

***Analítico-sintético*:** se utilizó para el desglose de la información en las diferentes áreas de importancia según los objetivos de la investigación, para facilitar su estudio, tales como la inteligencia artificial, la reducción de atributos y otras; para después integrar esa información e ir construyendo el estado del arte y los otros aspectos básicos de esta investigación.

Análisis histórico lógico: se utilizó para el estudio de la evolución y desarrollo histórico de la teoría de conjuntos aproximados y las áreas que la rodean: de forma general, minería de datos e Inteligencia artificial. Además para caracterizar la evolución de la funcionalidad de los métodos existentes.

Inductivo deductivo: se utilizó para obtener reglas de inferencia a partir de datos y regularidades iniciales para llegar a conclusiones.

Métodos Empíricos:

Consulta de expertos: se utilizó para obtener el criterio de los especialistas para guiar la investigación por el camino correcto.

Encuesta: se utilizó para aplicar el método Delphi de la consulta de expertos para obtener información relevante para los objetivos de la investigación.

Experimento*:* se utilizó para verificar la utilidad de la teoría de conjuntos aproximados en el pronóstico del rendimiento académico.

1 Fundamentación teórica

Introducción

El contenido de este capítulo constituye la base teórica del presente trabajo. En él se describen los principales conceptos y lineamientos, como resultado de la investigación realizada para la aplicación de la teoría de conjuntos aproximados en la predicción académica de estudiantes, tomando como muestra los datos del primer semestre de primer año de los estudiantes de los cursos académicos 2006-2007, 2007-2008, 2008-2009, pertenecientes a la facultad 5 de la Universidad de las Ciencias Informáticas.

Se brinda una vista global de los temas relacionados con la predicción académica de estudiantes, así como los principales conceptos asociados al dominio del problema. Además se realiza una comparación de las herramientas existentes y se determina cuáles van a ser las utilizadas.

1.1 Inteligencia artificial

La inteligencia artificial (IA) es la parte de las ciencias de la computación que intenta aplicar rasgos del pensamiento humano a la solución automatizada de problemas. Se utiliza cuando los métodos de búsqueda directos de soluciones son inaplicables debido a la estructura del espacio de búsqueda (frecuentemente, su alto número de dimensiones). Los resultados que ofrece la IA no son necesariamente óptimos, pero pueden satisfacer los criterios de calidad necesarios.

Existen tres componentes principales de estudio de la IA que son: el conocimiento, los algoritmos heurísticos y la incertidumbre. Luego aparece otro campo denominado Aprendizaje Automático y más tarde la teoría de conjuntos aproximados la cual crea una mezcla entre el aprendizaje automático y la incertidumbre. Figura 1: La teoría de conjuntos aproximados en el contexto de la Inteligancia Artificial.

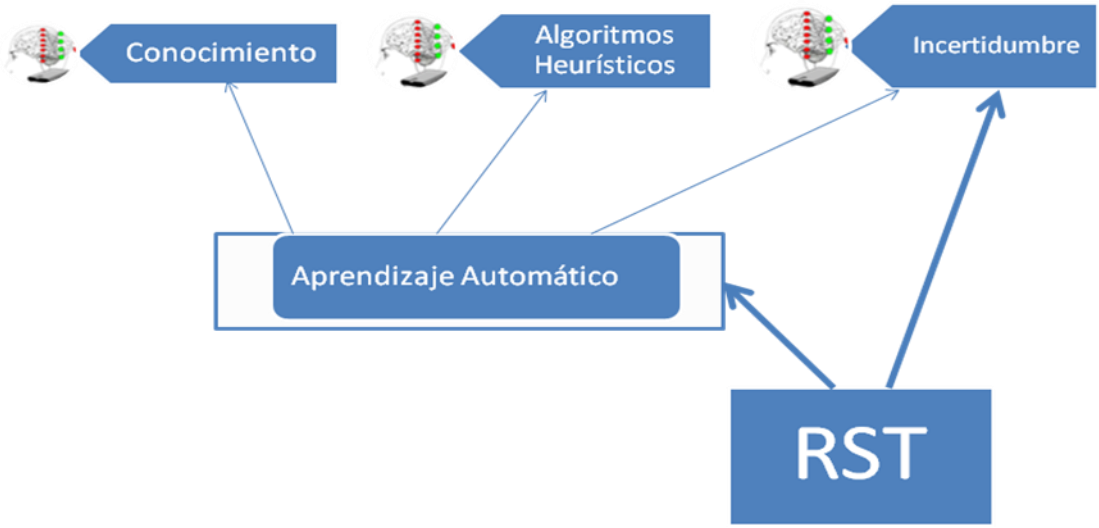


Figura 1: La teoría de conjuntos aproximados en el contexto de la Inteligancia Artificial.

Fuente: Conferencia del doctor Rafael Bello en el curso de verano para profesores de la Universidad de las Ciencias Informáticas.

La teoría de los conjuntos aproximados fue introducida por Z. Pawlak en 1982. Se basa en aproximar cualquier concepto por un par de conjuntos exactos llamados aproximación inferior y aproximación superior del concepto (6).

Con esta teoría es posible tratar tanto datos cuantitativos como cualitativos, y no se requiere eliminar las inconsistencias previas al análisis como los datos continuos y la información incompleta.

Actualmente, el mundo se encuentra camino a lo que se le llama *lógica disposicional.* En la lógica disposicional, las proposiciones se supone que están calificadas por frecuencia del uso, por ejemplo, «usualmente hace calor y llueve en La Habana durante el verano» y «usualmente lo que es escaso resulta caro». Desde esta perspectiva, la lógica disposicional puede contemplarse como la lógica del conocimiento y el razonamiento de sentido común. Otro de los temas que aún están en desarrollo es el concepto de *soft computing* como una especie de consorcio o sociedad entre la lógica borrosa, la neurocomputación y el razonamiento probabilístico, en el que la última incluiría los algoritmos genéticos, el razonamiento basado en evidencias y los sistemas caóticos.

El empleo creciente del *soft computing* ha supuesto una contribución importante para la concepción, el diseño y el desarrollo de sistemas inteligentes. Se aproxima el tiempo en que se necesitará una forma de medir la inteligencia de los sistemas hechos por el hombre. *Computación con Palabras* es el nombre que LOTFI A. ZADEH le da a esta nueva ciencia, esta proporciona un marco conceptual para calcular y razonar con palabras en lugar de con números. La idea básica que subyace a esta ciencia es que, en general, la información se transmite restringiendo los valores que puede tomar una variable. El punto de partida en la *Computación con Palabras* es la suposición de que la información dada se representa como una colección de proposiciones expresadas en un lenguaje natural. Cada proposición se contempla como una restricción implícita de una variable implícita.

El propio Zadeh expresó:

"C*reo que en unos años la computación con palabras llegará a ser una metodología por derecho propio, con un impacto de amplio rango tanto a nivel básico como a nivel aplicado. En el análisis final, el modelo de papel para la computación con palabras es la mente humana. Estamos entrando en una era de sistemas inteligentes que tendrán un impacto profundo –y esperemos que positivo – en las formas en que nos comunicamos, tomamos decisiones y utilizamos las máquinas. Creo que la lógica borrosa –junto con sus socios en el soft computing – jugará un papel importante en conseguir que la era de los sistemas inteligentes sea una realidad* (7).

1.2 Minería de datos

La minería de datos surge como una tecnología que intenta ayudar a comprender el contenido de un conjunto de datos. De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye, a estos datos, algún significado especial, pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación conjunta entre la información y ese modelo represente un valor agregado, entonces se refieren al conocimiento. En la Figura 2**:** Proceso de la Minería de datos. se ilustra la jerarquía que existe en una base de datos entre: dato, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en

esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento. La minería de datos trabaja en el nivel superior buscando patrones, comportamientos, agrupaciones, secuencias, tendencias o asociaciones que puedan generar algún modelo que permita comprender mejor el dominio para ayudar en una posible toma de decisión (8).

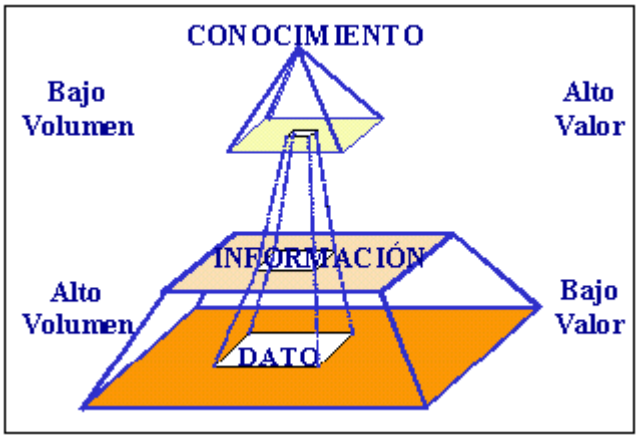


Figura 2: **Proceso de la Minería de datos.**

Fuente: tomado de la publicación “Minería de datos y aplicaciones” de los autores Fernando Virseda Benito y Javier Román Carrillo de la Universidad Carlos III de España.

Formalmente, minería de datos es el proceso de descubrir patrones de información interesante y potencialmente útil, inmersa en un gran conjunto de datos (8).

La Minería de datos es una combinación de procesos como:

• Extracción de datos
• Limpieza de datos.
• Selección de características.
• Análisis de resultados.

La idea de la minería de datos no es nueva. Ya desde los años sesenta los estadísticos manejaban términos pesca de datos (del Inglés *data fishing)*, minería de datos o arqueología de datos (del Inglés *data archaeology)* con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con información incompleta y datos continuos. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold,

Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de minería de datos. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

La minería de datos es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software (9).

Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones de software en cada etapa que pueden ser: estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de minería de datos muy poderosas que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

1.3 Preprocesamiento de los datos.

Las series históricas de datos pueden ser aprovechadas para la generación de nueva información. A costos relativamente bajos se pueden originar los nuevos usos de la información que no habían sido identificados al momento que fue creada la fuente original de datos. Estos nuevos usos están orientados fundamentalmente a la preparación de modelos de predicción, modelos de clasificación de patrones mediante su reconocimiento, etc. Uno de los medios para alcanzar estos objetivos es el análisis exploratorio de datos. Esta exploración es requerida en ciertos procesos que conducen a la construcción de modelos de pronóstico o clasificación y se le conoce también con el nombre de preprocesamiento.

La selección de estos datos a partir de las fuentes históricas debe ser representativa de su fuente original y a su vez, consistente al ser conformada, en algunos casos, por un conjunto reducido de datos (*reductos*) que caractericen la fuente histórica, mediante las variables originales o las transformadas.

Los datos reales pueden ser impuros, pueden conducir a la extracción de reglas poco útiles. Esto se debe a:

- Datos Incompletos: falta de valores de atributos.
- Datos con Ruido: datos inconsistentes.

La preparación de datos puede generar un conjunto de datos más pequeño que el original, esto puede mejorar la eficiencia del proceso de minería de datos. Esta actuación incluye:

Selección relevante de datos:

- Eliminando registros duplicados.
- Eliminando anomalías.

Reducción de Datos:

- Selección de características.
- Muestreo o selección de instancias.
- Discretización.

La preparación de datos genera “datos de calidad”, los cuales pueden conducir a reglas de calidad.

Por ejemplo, se puede:

- Recuperar información incompleta.
- Eliminar valores atípicos.
- Resolver conflictos.

El preprocesamiento de datos engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de minería de datos puedan obtener mayor y mejor información (mejor porcentaje de clasificación, reglas más completas, etc.) (10)

Los aspectos principales donde se desarrollan las técnicas de preprocesamiento de datos son:

- Recopilación e integración de datos.
- Limpieza de datos.
- Transformación de datos.
- Reducción de datos (selección de funciones, Instancia de selección, discretización)

El preprocesamiento de datos suele ser una necesidad cuando se trabaja con una aplicación real, con datos obtenidos directamente del problema.

Una ventaja: El preprocesamiento de datos permite aplicar los modelos de Minería de datos de forma más rápida y sencilla, obteniendo patrones de más calidad: precisión e interpretación

Un inconveniente: El preprocesamiento de datos no es un área totalmente estructurada con una metodología concreta de actuación para todos los problemas. Cada problema puede requerir una actuación diferente, utilizando diferentes herramientas de preprocesamiento.

1.3.1 Reducción de datos

La reducción de datos no es más que la selección de datos relevantes para la tarea de la Minería de datos. Figura 3: Representación de la reducción de los datos.

Diferentes tipos de la Reducción de Datos:

- Selección de características
- Selección de instancias
- Discretización

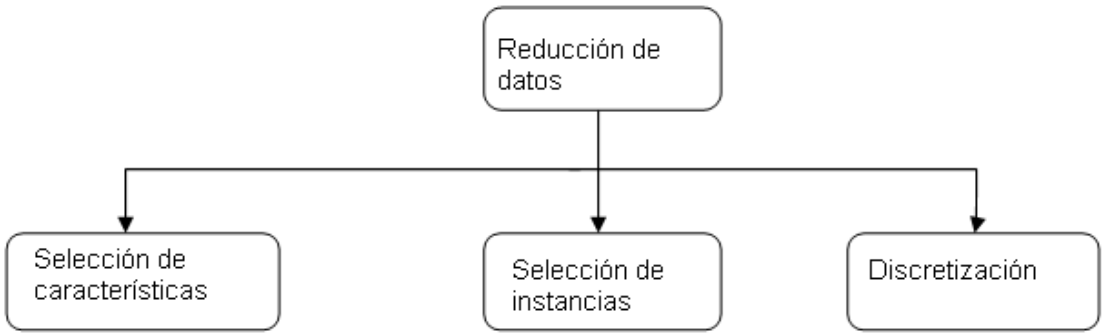


Figura 3: Representación de la reducción de los datos.

Fuente: Elaboración Propia.

1.3.1.1 Selección de atributos

La selección de atributos o características, como su nombre lo indica, selecciona un subconjunto de los atributos originales. Se puede considerar como un problema de búsqueda. (Ver Figura4**:** Proceso de la selección de atributos.). Pretende elegir atributos que sean relevantes para una aplicación y lograr el máximo rendimiento con el mínimo esfuerzo. Permite mejorar la precisión e interpretación de los métodos de aprendizaje automático, además de reducir el tamaño de la base de datos y el tiempo de los algoritmos de aprendizaje. El resultado de la selección de atributos sería:

- Menos datos: los algoritmos pueden aprender más rápido.
- Mayor exactitud: el clasificador generaliza mejor.
- Resultados más simples: más fácil de entender.
- Menos atributos: evitar obtenerlos posteriormente.

Las ventajas esperadas de este proceso son:

- Mejorar el desempeño predictivo.
- Construir modelos eficientemente.
- Mejorar el entendimiento sobre los modelos generados.

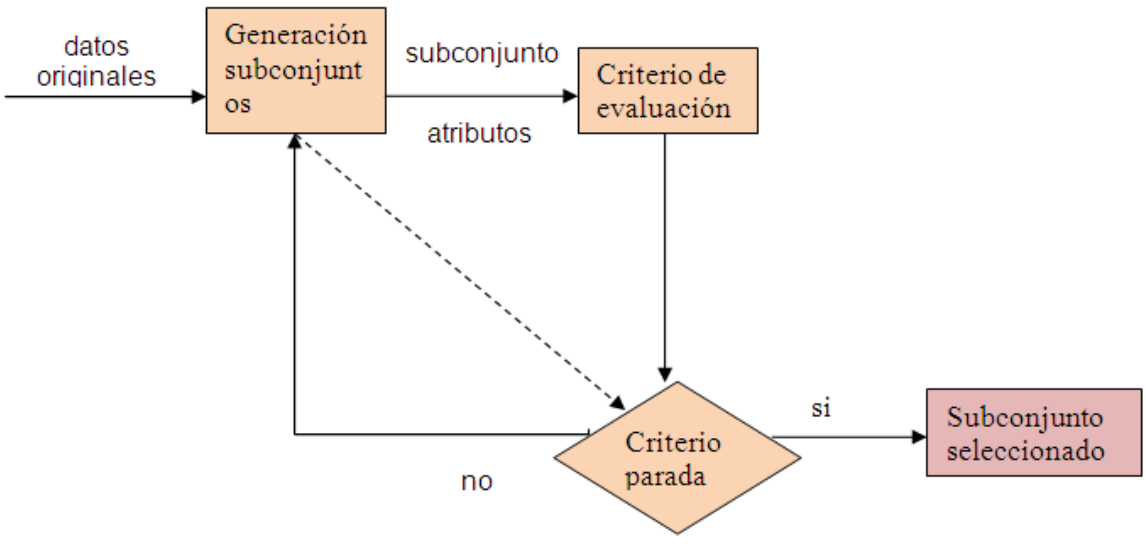


Figura4: **Proceso de la selección de atributos.**

Fuente: Elaboración Propia.

Aplicaciones recientes en categorización de textos o en selección de genes, ilustran claramente la necesidad de reducir el número de atributos donde existe una gran cantidad de estos y relativamente pocos datos.

Sea *S* un subconjunto de atributos de *F*. La meta es seleccionar el subconjunto más pequeño de atributos *S* de *F*, tal que la calidad obtenida con el subconjunto *S sea* semejante a la calidad obtenida con el conjunto original *F*. Sin embargo, encontrar el subconjunto de atributos óptimo es en general intratable.

El proceso de selección de atributos involucra cuatro pasos:

- Generación de candidatos, que involucra una estrategia de búsqueda.
- Evaluación de candidatos, que requiere un criterio de evaluación.
- Criterio de parada: Este puede darse por la estrategia de búsqueda, el número de iteraciones realizadas, el número de atributos seleccionados, que no se mejore el criterio de evaluación al añadir o quitar otro atributo, que el error de clasificación está por debajo a un valor.
- Validación de resultados: Si se conoce de entrada cuáles son los atributos relevantes, se puede comparar el resultado del algoritmo contra esos atributos conocidos. Como normalmente no se sabe, se puede comparar el error en la clasificación con y sin la selección de atributos.

En general, los algoritmos de selección de atributos se distinguen por su forma de evaluar atributos y se pueden clasificar en tres tipos:

- Filtros (del inglés filters): seleccionan los atributos en forma independiente del algoritmo de aprendizaje Figura 5: Esquema genérico de algoritmo tipo filtros. (11).
- Envolturas (del inglés wrappers): usan el desempeño de algún clasificador para determinar lo deseable de un subconjunto.
- Híbridos: usan una combinación de los dos criterios de evaluación en diferentes etapas del proceso de búsqueda.

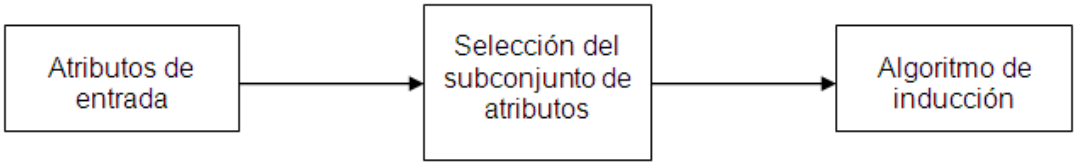


Figura 5: Esquema genérico de algoritmo tipo filtros.

Fuente: elaboración propia.

Cambiando la forma de generar nuevos candidatos y el criterio de parada, se pueden generar diferentes versiones. Diferentes criterios de evaluación generan diferentes algoritmos de tipo filtros. Diferentes algoritmos de aprendizaje generan diferentes algoritmos de tipo envoltura.

El otro tipo de algoritmo es el híbrido que combina el criterio de evaluación de ambos. La idea general del algoritmo es la siguiente:

- Empieza con un subconjunto S_0.
- Aumenta la el tamaño del subconjunto inicial en un atributo (cardinalidad) y evalúa todos los elementos de esa cardinalidad con una medida independiente del algoritmo de aprendizaje.
- El mejor subconjunto se evalúa con un algoritmo de aprendizaje y se queda con el mejor (de todas las cardinalidades vistas hasta el momento).
- Continúa con la siguiente cardinalidad.

En caso que no mejore la calidad del algoritmo de aprendizaje, esta condición se puede tomar como criterio de parada. (12).

La selección de atributos tiene distintas clasificaciones, estas pueden ser:

1. Según la evaluación: filtros, envolturas, híbridos.
2. Disponibilidad de la clase: supervisados, no supervisado.
3. Según la búsqueda: completa, heurística, aleatoria.

4. Según la salida del algoritmo: lista ordenada de todos los atributos según su importancia, subconjunto de atributos.

1.4 Predicción

La definición de predicción es simple; sin embargo envuelve muchas situaciones que influyen en su resultado o en el cumplimiento del mismo. Una predicción es la estimación anticipada del valor de una variable. Pueden ser utilizadas en muchos campos de la sociedad dentro de ellos: la salud, el deporte, la economía y la educación.

"Una predicción es un inicio o una señal por donde se puede saber un dato futura mediante indicios" (13).

La predicción juega un papel muy importante en la planificación de materiales, se pueden encontrar predicciones de abastecimiento, de condiciones, comerciales, de tecnología, precios y otros muchos y en cualquiera de estas ramas la predicción es necesaria para la toma de decisiones.

A pesar de que hoy en día el manejo de predicciones es muy cotidiano en las pequeñas y medianas empresas, existen problemáticas en cuanto a la planeación de necesidades futuras. La problemática principal es la poca confiabilidad.

Es común encontrar diversas técnicas para el planteamiento de predicciones y estas se van desarrollando rápidamente, estas técnicas pueden ser cuantitativas y cualitativas, además se pueden emplear separadas o en conjunto.

1.4.1 Predicción académica

La predicción académica se encarga de reflejar por adelantado cuáles son los estudiantes con posibilidades de éxito y cuáles son los factores que influyen en este.

Sin duda, las características socioeconómicas y culturales de la familia afectan de manera importante el desempeño de los estudiantes, puesto que son determinantes en la preparación del estudiante desde antes de su entrada al sistema educativo y durante toda su trayectoria académica. Además, la influencia del entorno educacional también

es importante, debido a que el rendimiento se ve afectado por lo que sucede dentro del aula y por las características de los compañeros y profesores con que se relaciona el estudiante.

Por lo tanto, identificar estos factores y analizar conjuntamente su influencia en los resultados de los estudiantes como una forma de potenciar sus bondades y disminuir sus debilidades, parece ser una estrategia interesante de probar, dado que la identificación temprana de elementos de riesgo puede permitir la realización oportuna de acciones correctivas en el proceso educativo.

Numerosos investigadores han propuesto procedimientos para la predicción del éxito académico, con la intención de detectar tanto a los estudiantes con alto riesgo de fracaso académico, como a los estudiantes con altas probabilidades de éxito, todo ello con el fin de mejorar la calidad de la gestión docente.

Mediante la utilización de modelos de pronósticos se ha podido anticipar de cierta forma la realidad esperada.

El problema de la adecuada determinación del criterio en un estudio predictivo es de suma importancia y, a la vez, se ha de reconocer la dificultad que entraña conseguir un criterio que, al reunir una serie de características técnicas, haga posible determinar con adecuada precisión el valor real de los predictores.

Los trabajos de Schneider, Sorensen, Hallinan, McParüand y Levin, ponen de manifiesto una serie de consideraciones (14):

- El uso de una sola medida del rendimiento (tests de rendimiento o calificaciones) en un momento determinado (final de curso, etapa preescolaridad) está subestimando los efectos educativos de la escuela.

- Debería tenerse presente no sólo el aprendizaje inmediato, sino también el aprendizaje relevante al futuro del alumno (tanto educativo como ocupacional o vital).

- En los resultados de la enseñanza debe incluirse tanto los logros alcanzados en los objetivos académicos como en los de formación y desarrollo de la personalidad.

- Las actitudes ante el fenómeno de aprender, así como las conductas sociales de clara incidencia en la convivencia de la comunidad, deberían tener un mayor peso en la determinación de la calidad del producto educativo (14).

García Llamas (1986) mantiene: *"en estos momentos las calificaciones escolares, a pesar de las múltiples críticas de las que son objeto, constituyen el mejor indicador del rendimiento académico".*

1.5 Árbol de decisión

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta que es una decisión tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser valores discretos o continuos. Se utilizan más los valores discretos por simplicidad, cuando se utilizan valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utilizan los continuos se denomina regresión (15).

Un árbol de decisión lleva a cabo un test a medida que este se recorre hacia las hojas para alcanzar así una decisión. El árbol de decisión suele contener nodos internos, nodos de probabilidad, nodos hojas y arcos. Un nodo interno contiene un test sobre algún valor de una de las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, este tipo de nodos es redondo, los demás son cuadrados. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada (16).

Los árboles de decisión son diagramas de decisiones secuenciales que muestran sus posibles resultados. Éstos ayudan a determinar cuáles son las opciones al mostrarles las distintas decisiones y sus resultados.

1.6 Teoría de los conjuntos aproximados

La teoría de conjuntos aproximados (del inglés rough sets theory RST) se ha venido desarrollando desde la década de los 80, cuyo autor es el investigador polaco Zdzislaw Pawlak. A través de todo este tiempo, como toda teoría, se ha ido enriqueciendo con nuevos aportes, derivados de una mayor investigación sobre sus alcances y bondades, tanto para aplicaciones teóricas como prácticas.

En 1982, el profesor Zdzislaw Pawlak publicó un artículo en el cual presentó las primeras referencias sobre conjuntos aproximados, con el cual abrió una nueva dirección en el desarrollo de teorías sobre la información incompleta.

Según la universidad de Kentucky de los EE.UU. RST “es una importante herramienta en el análisis de datos con importantes aplicaciones en la Minería de datos y el descubrimiento del conocimiento” (20).

Para el Instituto de Estadísticas de la India “el uso de la teoría para el descubrimiento y minar reglas es ampliamente reconocido” (20).

La RST parece ser un modelo matemático natural para la vaguedad y la incertidumbre. La vaguedad es una propiedad de los conjuntos (conceptos) y puede ser atribuida a los límites del conjunto, mientras que la incertidumbre es una propiedad de los elementos de un conjunto y tiene que ver con pertenencia o no de costos (17).

Una de las principales ventajas de la teoría de conjuntos aproximados es que no necesita de ninguna información adicional o preliminar, como las distribuciones de probabilidad en los enfoques estadísticos, la asignación de probabilidad básica en la *Teoría de la evidencia*, o el valor de posibilidad en la *teoría de conjuntos borrosos* (18).

En ese sentido, la RST se basa en el concepto de "indiscernibilidad". Considerando que indiscernir significa no conseguir distinguir una cosa de otra, por medio de los

sentidos o de la inteligencia humana, lo que busca la RST es encontrar todos aquellos objetos (acciones, alternativas, candidatos, pacientes, etc.) que producen un mismo tipo de información, es decir, aquellos objetos que son "indiscernibles" (19).

1.6.1 Representación de la RST.

RST= *Sistema de información* + *Relación de separabilidad*.
Dónde *Sistema de información* no es más que los datos del dominio, y la *Relación de separabilidad* está constituida por las componentes de la teoría.

1.6.1.1 Sistema de información

Si se tiene en cuenta que a todo objeto se le puede asociar algún tipo de información basados en sus atributos o características, entonces es factible representar estos atributos por medio de una tabla o *sistema de información*. Dentro de éste sistema de información, las filas representan los objetos, en cuanto que las columnas representan los atributos. Las entradas de la tabla, que no son otra cosa más que pares (objeto, atributo), vienen a ser los valores de cada objeto para cada atributo. La tabla de información debe ser entendida, en términos prácticos, como una matriz finita, tal como se observa en la siguiente tabla 1.

Tabla 1: Representación de un sistema de información.

Objeto m	**Valor m, 1**	**Valor m, 2**	**…….**	**Valor m, n**
	Atributo 1	**Atributo 2**	**…….**	**Atributo n**
Objeto 1	Valor 1, 1	Valor 1, 2	…….	Valor 1, n
Objeto 2	Valor 2, 1	Valor 2, 2	…….	Valor 2, n
……	…….	…….	…….	…….

Sea un conjunto de atributos $A = \{A_1, A_2, ..., A_n\}$ y un conjunto *U* de ejemplos no vacío llamado universo (objetos, entidades, etc.) descritos usando los atributos A_i.

Un sistema de información se convierte en un sistema de decisión cuando a cada elemento de *U* se le agrega un nuevo atributo ¨*d* ¨ llamado decisión. Entonces:

Sistema de decisión sería: $(U, A \cup \{d\}) : d \notin A$ donde *d* es el rasgo de decisión.

La tabla 2, como se puede observar debajo, es un ejemplo de sistema de decisión:

Tabla 2: Representación de un sistema de decisión.

Paciente	**Dolor de cabeza**	**Dolor muscular**	**Temperatura**	**Gripe** ***(Atributo de decisión)***
P1	No	Si	Alta	Si
P2	Si	No	Alta	Si
P3	Si	Si	Muy alta	Si
P4	No	Si	Normal	No

1.6.1.2 Relación de separabilidad

Un atributo $A_i \notin A$, siendo *A* el conjunto de atributos inicial, separa un objeto ´*X*´ de un objeto ´*Y*´ si y sólo si los valores para el mismo atributo en elementos diferentes sean totalmente diferentes.

$$(R_1) : Separa(A_i, x, y) \Leftrightarrow f(x, A_i) \neq f(y, A_i)$$

1.6.1.2.1 Relación de separabilidad para información incompleta

$V_i = V_i \cup \{*\}$

$$(R_2): Separa(A_i, x, y) \Leftrightarrow f(x, A_i) = f(y, A_i)$$

y

$f(x, A_i) \neq *$

y

$f(y, A_i) \neq *$

1.6.1.2.2 Relación de separabilidad para rasgos continuos.

$$(R_3): Separa(A_i, x, y) \Leftrightarrow \left| f(x, A_i) - f(y, A_i) \right| > \varepsilon$$

1.6.2 Matriz de separabilidad

Una matriz de separación (*MA)* es una matriz $|U|x|U|$ donde cada entrada *MA* (x, y)$\subseteq$A contiene el conjunto de atributos que distinguen los elementos ´*X*´ e ´*Y*´ de U.

$$MA(x, y) = \{A_i \in A: Separa(A_i, x, y)\}$$

Matriz de separabilidad para el ejemplo del sistema de decisión anterior.

	P_1	P_2	P_3	P_4
P_1		{1,2}	{1,2,3}	{3,4}
P_2			{2,3}	{1,2,3,4}
P_3				{1,3,4}
P_4				

En este ejemplo se puede observar que los objetos P1 y P2 se distinguen en los rasgos 1 y 2.

1.6.3 Relación de inseparabilidad

A cada subconjunto de atributos *B* de *A* $(B \subseteq A)$ está asociada a una relación binaria de inseparabilidad denotada por I_B la cual es el conjunto de pares de objetos que son inseparables unos de otros por esta relación:

$I_B = \{(x, y) \in UxU : f(x, A_i) = f(y, A_i) \forall A_i \in B\}$

I_B es una relación simétrica reflexiva y transitiva (Relación de equivalencia).

Un sistema de este tipo es consistente si:

$Sea \delta B:$

$\delta B(x) = \{V \in Vd : \exists y \in IB(x) : d(y) = V\}$

o sea: si y sólo si δ B es unitario.

Cuando un sistema es consistente todos los elementos inseparables entre ellos tienen el mismo valor de decisión.

1.6.4 Conceptos básicos de la RST

Aproximación inferior de *X* es:

$B_i(X) = \{X \in U | B(x) \in X\}$

Aproximación superior de *X* es:

$B_s(X) = \{X \in U | B(x) \cap X \neq \{\}\}$

Frontera de *X* es:

$BNB(X) = B_s(X) - B_i(X)$

En caso de múltiples decisiones dado el sistema de decisión $S = (U, A \cup D)$, sean $I \subseteq A$ y $I \subseteq D$, dos relaciones de equivalencia, las cuales incluyen sobre *U* las particiones $A^* = \{x_1, x_2, ..., x_n\} y D^* = \{x_1, x_2, ..., x_n\}$

$Entonces:$

$POS(D^*) = \cup B_i(Y_j) \forall Y_j \in D^*$

$BND(D^*) = \cup B_s(Y_j) - B_i(Y_j) \forall Y_j \in D^*$

$NEG(D^*) = U - \cup B_s(Y_j) \forall Y_j \in D^*$

1.6.5 Propiedades de la RST

La teoría de conjuntos aproximados tiene un conjunto de propiedades que se muestran a continuación:

a) La aproximación inferior del conjunto *X* $B_i(X)$ es un subconjunto del propio conjunto *X* y, a su vez, este es un subconjunto de la aproximación superior de *X* $B_s(X)$.

$$B_i(X) \subseteq X \subseteq B_s(X)$$

b) La aproximación inferior del conjunto vacío $B_i(\{\})$ es igual a la aproximación superior del conjunto vacío $B_s(\{\})$, siendo estos igual al propio conjunto vacío.

$$B_i(\{\}) = B_s(\{\}) = \{\}$$

c) La aproximación inferior del universo $B_i(U)$ es igual a la aproximación superior del universo $B_s(U)$, siendo estos igual al propio universo.

$$B_i(U) = B_s(U) = U$$

d) La aproximación superior de la unión de dos subconjuntos X e Y del universo $B_s(X \cup Y)$ es igual a la aproximación superior del subconjunto X unido con la aproximación superior del subconjunto Y $B_s(X) \cup B_s(Y)$.

$$B_s(X \cup Y) = B_s(X) \cup B_s(Y)$$

e) La aproximación superior de la intersección de dos subconjuntos X e Y del universo universo $B_s(X \cap Y)$ es igual a la aproximación inferior del subconjunto X interceptado con la aproximación inferior del subconjunto Y $B_i(X) \cap B_i(Y)$.

$$B_s(X \cap Y) = B_i(X) \cap B_i(Y)$$

f) Un subconjunto X es subconjunto de otro subconjunto Y, siendo X e Y subconjuntos del universo, si y sólo si la aproximación inferior de X $B_i(X)$ es subconjunto de la aproximación inferior de Y $B_i(Y)$ y la aproximación superior de X $B_s(X)$ es subconjunto de la aproximación superior de Y $B_s(Y)$.

$$X \subseteq Y \Rightarrow B_i(X) \subseteq B_i(Y) y B_s(X) \subseteq B_s(Y)$$

g) La aproximación inferior de un subconjunto X del universo U $B_i(X)$ es igual al universo U menos la aproximación superior del universo menos el conjunto X $B_s(U-X)$.

$$B_i(X) = U - B_s(U-X)$$

1.6.6 Mediciones asociadas a la RST

Precisión de la aproximación: $$\alpha_B(X) = \frac{|B_\bullet(X)|}{|B^\bullet(X)|}$$

$|X|$: Cardinalidad del conjunto *X*.

$0 \le \alpha_B \le 1$

Si $\alpha_B = 1$ entonces *X* es un conjunto duro o exacto con respecto a *B*.

Si $\alpha_B < 1$ entonces *X* es aproximado o vago con respecto a *B.*

Calidad de la aproximación: $\gamma_B(X) = \frac{|B_\bullet(X)|}{|X|}$

Esta medida no es más que la frecuencia relativa de los objetos correctamente calificados por medio de tributos en *B* además: $0 \le \alpha_B(X) \le \gamma_B(X) \le 1$

Calidad de la clasificación del sistema: $\gamma_B(Y) = \frac{\sum_{i=1}^{n} |B_\bullet(Y_i)|}{|U|}$

Si $\gamma_B(Y) = 1$ el sistema es consistente.

Grado de pertenencia: $$\mu_x^B(x) = \frac{|X \cap B(x)|}{|B(x)|}$$

Propiedades

a) $\mu_x^B(X) = 1 \rightarrow x \in B_i(X)$

b) $\mu_x^B(X) = 0 \rightarrow x \in U - B_s(X)$

c) $0 < \mu_x^B(X) < 1 \rightarrow x \in BNB(X)$

d) $\mu_{U-x}^B(X) = 1 - \mu_x^B(X) \forall x \in U$

e) $\mu_{x \cup y}^B B(X) = \max(\mu_x^B(X), \mu_y^B(X)) \forall x \in U$

f) $\mu_{x \cap y}^B B(X) = \min(\mu_x^B(X), \mu_y^B(X)) \forall x \in U$

1.6.7 Reducto

Dado un sistema de información *S= (U, A)*, donde *U* es el universo y *A* es el conjunto de atributos, un reducto de éste es un conjunto mínimo de atributos $B \subseteq A$ tal que $I_A \approx I_B$.

Un reducto es un subconjunto de atributos que mantiene la separabilidad de los objetos del universo, o sea, la mayor partición, la más fina del universo, se logra si se tienen en cuenta todos los rasgos, lo que no significa que esta misma separabilidad no se pueda lograr con menos atributos.

1.6.8 Complejidad temporal de procesos

- Encontrar $B_i(X)$: $O(\mathrm{Im}^2)$
- Encontrar $B_s(X)$: $O(\mathrm{Im}^2)$

I: número de atributos que describen los objetos.

m: número de objetos del universo.

- Encontrar un reducto: $O(I^2 m^2)$
- Encontrar todos los reductos: $O(2^I J)$.

J: Costo de encontrar un reducto

1.6.9 Extensiones de la RST

Las extensiones de la RST están definidas para los casos de objetos incompletos (datos omitidos) y con rasgos continuos.

1.6.9.1 Objetos incompletos

Sean $x, y \in U$ y $P \subseteq A$, e Ipx $\forall q \in P$ se cumple:

$f(x,q) = f(y,q), f(x,q) = * ó f(y,q) = *$

La relación I_p es reflexiva y simétrica pero no transitiva.

2da relación de separabilidad.

Para cada $P \subseteq A$ se define el conjunto U*P de objetos sin valores desconocidos para algún atributo en P.

$U * P = \{x \in U : f(x,q) \neq * \text{ para algún } q \in P\}$

Y $II_p x$: todo $q \in P$ se cumple:

$f(x,q) = f(y,q), f(x,q) = * ó f(y,q) = *$

$II_p(x) = \{y \in U : y II_p x\}$

conjunto de objetos completos (no tienen ningún valor omitido).

Aproximación superior e inferior de II_p:

$II_{pi}(x) = \{x \in U * P : II_p(x) \subseteq X\}$

$II_{ps}(x) = \{x \in U * P : II_p(x) \cap X \neq \{\}\}$

1.6.9.2 Rasgos continuos

Cuando una relación $R \subseteq UxU$ es reflexiva (x R x) para cualquier $x \in U$, es simétrica ((x R y) si (y R x)) para cualquier par de elementos $x, y \in U$.

El par (U, R) se denomina espacio de tolerancia.

Aproximaciones superior e inferior para rasgos continuos.

Dada la relación binaria reflexiva R:

$R(x) = \{y \in U : yRx\}$ objetos similares a *x*.

$R^{-1}(x) = \{y \in U : xRy\}$ objetos a los que *x* es similar.

$R\inf(x) = \{x \in U : R^{-1}(x) \subseteq X\}$

$R\sup(x) = \{x \in U : R^{-1}(x) \cup X\} = \cup R(x), x \in X$

$R\inf(x) \leq x \leq R\sup(x).$

Teniendo en cuenta que la RST es una herramienta de Minería de datos, actualmente tiene aplicaciones en diferentes campos, principalmente en sistemas de apoyo a la decisión y sistemas gerenciales de información. Las herramientas de minería de datos, son el conjunto de procedimientos y técnicas que buscan extraer patrones dentro de un conjunto de datos (20).

Además la RST se puede utilizar para modelar la incertidumbre de la siguiente manera:

- Medida de la consistencia del sistema de decisión: calidad de la clasificación.
- Cálculo de la incertidumbre del conocimiento descubierto.
- Integración con otros modelos de incertidumbre como la Teoría de Conjuntos Difusos (del Inglés Fuzzy Sets Theory: FST) para crear modelos más robustos

Vinculación de las teorías de conjuntos borrosos y difusos.

- Fuzzy-Rough Sets.
  Cálculo de la aproximación inferior y superior de un conjunto borroso *X.*
- Rough-Fuzzy Sets.
  Discretización de los datos utilizando conjuntos borrosos.
- Relaciones entre las funciones de pertenencia borrosa y aproximada.

1.6.10 Definición formal de un conjunto borroso

Un conjunto borroso *A* en *x* es expresado como un conjunto de pares ordenados:

$A = \{(x, \mu_A(x)) | x \in X\}$

X: Universo.

μ_A: Función de pertenencia.

A: Conjunto borroso.

x: Elemento del conjunto.

Aplicación de la RST en la Minería de datos.

1. Análisis de los atributos a considerar.

- Selección de los atributos.

- Análisis de la dependencia entre los atributos.
- Reducción de atributos.
- Cálculo de la importancia de un atributo.
- Cálculo de la calidad de un conjunto de entrenamiento.

2. Formulación del conocimiento descubierto

- Descubrimiento de reglas causales.
- Cálculo de la certidumbre de las reglas causales.

En la Figura 6: Mapa conceptual sobre RST.se representa un mapa conceptual sobre la RST.

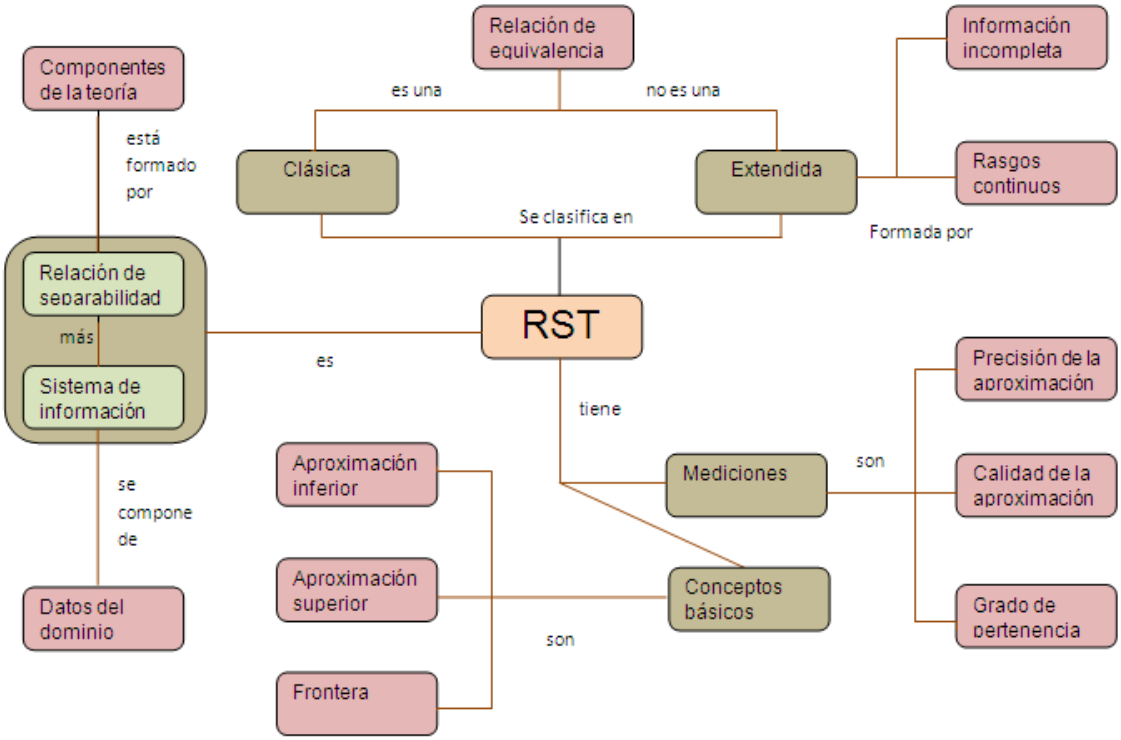


Figura 6: Mapa conceptual sobre RST.
Fuente: Elaboración propia.

1.7 Descripción de las herramientas y técnicas utilizadas

1.7.1 Método Delphi de la consulta de expertos

Ruiz e Ispizua describen la técnica Delphi como un método de investigación sociológica, que independientemente de que pertenece al tipo de entrevista de profundidad en grupo, se aparta de ellas agregando características particulares. Es una técnica grupal de análisis de opinión, parte de un supuesto fundamental y de que el criterio de un individuo particular es menos fiable que el de un grupo de personas en igualdad de condiciones, en general utiliza e investiga la opinión de expertos (21).

Varios son los autores que han aportado una definición de este método, aunque para Konow y Pérez el intento de definirlo es como limitar el alcance y contenido, por lo que es más aconsejable dar una descripción general de sus características, limitaciones, usos y aplicaciones (22).

Parisca (23) considera que el Método Delphi se basa en el principio de la inteligencia colectiva y que trata de lograr un consenso de opiniones expresadas individualmente por un grupo de personas seleccionadas cuidadosamente como expertos calificados en torno al tema, por medio de la iteración sucesiva de un cuestionario retroalimentado de los resultados promedio de la ronda anterior, aplicando cálculos estadísticos.

Las principales características del método están dadas por el anonimato de los participantes (excepto el investigador), iteración (manejar tantas rondas como sean necesarias), retroalimentación (del Inglés *feedback*) controlada, sin presiones para la conformidad, respuesta de grupo en forma estadística (el grado de consenso se procesa por medio de técnicas estadísticas) y justificación de respuestas (discrepancias/consenso).

Suelen distinguirse tres etapas o fases fundamentales en la aplicación del método (24):

1. *Fase preliminar*. Se delimita el contexto, los objetivos, el diseño, los elementos básicos del trabajo y la selección de los expertos.

2. *Fase exploratoria*. Elaboración y aplicación de los cuestionarios según sucesivas vueltas, de tal forma que con las respuestas más comunes de la primera se confecciona la siguiente.

3. *Fase final*. Análisis estadísticos y presentación de la información.

Para la aplicación del método es necesario considerar metodológicamente dos aspectos básicos de su caracterización sobre los cuales se sustenta, que son:

La selección del grupo de expertos a encuestar: personas conocedoras, con reconocida competencia y con experiencia en el tema que garantice la confiabilidad de los resultados, creativos e interesados en participar.

Por la limitación de tiempo y recursos se recomienda que el número de expertos participantes en la investigación que se desarrolla no sea muy numeroso.

Elaboración de los cuestionarios: tener en cuenta la teoría de la comunicación, con mecanismos que reduzcan los sesgos en las respuestas, preguntas claras, precisas e independientes. Suelen ser preguntas cuantitativas para calcular medias y rangos, y cualitativas para la justificación de sus opiniones.

1.7.2 STATGRAPHICS

STATGRAPHICS es una herramienta diseñada para el estudio y análisis estadístico, es sencillo de aprender y utilizar gracias a su diseño intuitivo que facilita la realización de los diversos análisis. Se aplica en los siguientes campos: estadística descriptiva, calidad.

Está compuesto por las siguientes partes:

- Editor de datos**.**

El editor de datos consiste en una hoja de cálculo con la que se introducen y gestionan los datos.

- StatFolio.

Es una herramienta que permite almacenar trabajos completos realizados que pueden ser recuperados posteriormente con todos sus procedimientos y análisis sin tener que crear macros. Permite salvar y recuperar ficheros de datos, así como gráficos y las opciones del análisis. Así, cuando se deseen realizar análisis anteriores en otros conjuntos de datos diferentes, es suficiente con abrir el StatFolio correspondiente y elegir las nuevas variables, para obtener rápidamente actualizados los nuevos resultados.

- StatAdvisor.

Es un intérprete estadístico que analiza las salidas obtenidas, determinando si los resultados son estadísticamente significativos o no y subrayando cualquier tipo de anomalías en el análisis. La explicación depende de las variables utilizadas en el análisis. Por lo tanto, es una herramienta que sirve de ayuda para interpretar las salidas de los análisis y procedimientos realizados.

- StatGallery.

Esta herramienta permite combinar e imprimir en sucesivas pantallas textos y gráficos combinados facilitando la realización de informes. Su finalidad es la de archivar los resultados obtenidos con un análisis.

- StatReporter.

Es una herramienta intermedia entre un block de notas y un procesador de textos completo, que permite combinar informes a los que añadir tus propias notas o incluir un StatGallery.

El informe puede ser personalizado usando opciones de edición (cortar, copiar), reordenando los textos y los gráficos, cambiando el estilo de las fuentes de texto así como su tamaño y color. Además, cuando se hacen cambios en los gráficos en un análisis, los cambios se reflejan automáticamente en el StatReporter (25).

1.7.3 WEKA

Es una extensa colección de algoritmos de máquinas de conocimiento desarrollados por la universidad de Waikato (Nueva Zelanda) implementada en Java; útil para ser

aplicados sobre datos mediante las interfaces que ofrece o para insertarlos dentro de cualquier aplicación. Además WEKA contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización. WEKA está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla.

Este programa es de libre distribución y difusión. Además, ya que WEKA está programado en Java, es independiente de la arquitectura, funciona en cualquier plataforma sobre la que haya una máquina virtual Java disponible (26).

El paquete WEKA contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades.

Los puntos fuertes de WEKA son:

- Está disponible libremente bajo la licencia pública general de GNU.
- Es portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado.
- Es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario (27).

1.7.3.1 La interfaz de usuario

Al ejecutar la aplicación nos aparece el **selector de interfaz de WEKA** (*WEKA GUI Chooser*) que da la opción de seleccionar entre cuatro posibles interfaces de usuario para acceder a las funcionalidades del programa, éstas son "*Simple CLI*", "*Explorer*", "*Experimenter*" y "*Knowledge Flow*".

1.7.3.2 Simple CLI

Simple CLI es la abreviatura de *Simple Command-Line Interface* (Interfaz Simple de Línea de Comandos); se trata de una consola que permite acceder a todas las opciones de WEKA desde línea de comandos.

1.7.3.3 Explorer

La interfaz **Explorer** (Explorador) dispone de varios paneles que dan acceso a los componentes principales del banco de trabajo:

- El panel "*Preprocess*" dispone de opciones para importar datos de una base de datos, de un fichero CSV, y para preprocesar estos datos utilizando los denominados algoritmos de *filtrado*. Estos filtros se pueden utilizar para transformar los datos (por ejemplo convirtiendo datos numéricos en valores discretos) y para eliminar registros o atributos según ciertos criterios previamente especificados.

- El panel "*Classify*" permite al usuario aplicar algoritmos de clasificación estadística y análisis de regresión (denominados todos *clasificadores* en WEKA) a los conjuntos de datos resultantes, para estimar la exactitud del modelo predictivo resultante, y para visualizar predicciones erróneas, curvas ROC, o el propio modelo (si este es susceptible de ser visualizado, como por ejemplo un árbol de decisión).

- El panel "*Associate*" proporciona acceso a las reglas de asociación aprendidas que intentan identificar todas las interrelaciones importantes entre los atributos de los datos.

1.7.3.4 Experimenter

La interfaz **Experimenter** (Experimentador) permite la comparación sistemática de una ejecución de los algoritmos predictivos de WEKA sobre una colección de conjuntos de datos.

1.7.3.5 Knowledge Flow

Knowledge Flow (Flujo de Conocimiento) es una interfaz que soporta esencialmente las mismas funciones que el *Explorer* pero con una interfaz que permite "*arrastrar y soltar*". Una ventaja es que ofrece soporte para el aprendizaje incremental.

1.7.4 UML

El Lenguaje Unificado de Modelado (UML) es ante todo un lenguaje. Un lenguaje proporciona un vocabulario y unas reglas para permitir una comunicación. En este caso, este lenguaje se centra en la representación gráfica de un sistema.

Este lenguaje indica cómo leer los modelos, pero no dice cómo crearlos. Esto último es el objetivo de las metodologías de desarrollo.

Los objetivos de UML son muchos, pero se pueden sintetizar sus funciones:

- Visualizar: UML permite expresar de una forma gráfica un sistema de forma que otro lo puede entender.
- Especificar: UML permite especificar cuáles son las características de un sistema antes de su construcción.
- Construir: A partir de los modelos especificados se pueden construir los sistemas diseñados.
- Documentar: Los propios elementos gráficos sirven como documentación del sistema desarrollado que pueden servir para su futura revisión.

Aunque UML está pensado para modelar sistemas complejos con gran cantidad de software, el lenguaje es los suficientemente expresivo como para modelar sistemas que no son informáticos, como flujos de trabajo (del Inglés *workflow*) en una empresa, diseño de la estructura de una organización y por supuesto, en el diseño de hardware.

Un modelo UML está compuesto por tres clases de bloques de construcción:

- *Elementos:* son abstracciones de cosas reales o ficticias (objetos, acciones).
- *Relaciones:* relacionan los elementos entre sí.
- *Diagramas:* Son colecciones de elementos con sus relaciones.

Los diagramas más utilizados son los de casos de uso, clases y secuencia (28).

1.7.5 Lenguaje de programación Python

Python es un lenguaje de scripting independiente de plataforma y orientado a objetos, preparado para realizar cualquier tipo de programa, desde aplicaciones Windows a servidores de red o incluso, páginas web. Es un lenguaje interpretado, lo que significa que no se necesita compilar el código fuente para poder ejecutarlo, lo que ofrece ventajas como el ahorro del tiempo de desarrollo e inconvenientes como una menor velocidad en tiempo de ejecución del programa.

En los últimos años el lenguaje se ha hecho muy popular, gracias a varias razones como:

- La cantidad de librerías que contiene, tipos de datos y funciones incorporadas en el propio lenguaje, que ayudan a realizar muchas tareas habituales sin necesidad de tener que programarlas desde cero.
- La sencillez y velocidad con la que se crean los programas. Un programa en Python suele tener menos líneas de código que su equivalente en *Java o C*.
- La cantidad de plataformas en las que se puede desarrollar, como Unix, Windows, OS/2, Mac, Amiga y otros.
- Además, Python es gratuito, incluso para propósitos empresariales.

Se pueden crear todo tipo de programas. No es un lenguaje creado específicamente para la web, aunque entre sus posibilidades sí se encuentra el desarrollo de páginas web (29).

Hay versiones disponibles de Python en muchos sistemas informáticos distintos. Originalmente se desarrolló para Unix, aunque cualquier sistema es compatible con el lenguaje siempre y cuando exista un intérprete programado para él.
Python dispone de un intérprete por línea de comandos en el que se pueden introducir sentencias. Cada sentencia se ejecuta y produce un resultado visible, que puede ayudar a entender mejor el lenguaje y probar los resultados de la ejecución de porciones de código rápidamente.La programación orientada a objetos está soportada en Python y ofrece en muchos casos una manera sencilla de crear programas con componentes reutilizables. Dispone de muchas funciones incorporadas en el propio

lenguaje, para el tratamiento de strings, números, archivos, entre otros. Además, existen librerías que se pueden importar en los programas para tratar temas específicos como la programación de ventanas o sistemas en red o cosas tan interesantes como crear archivos comprimidos en .zip.

Python tiene una sintaxis visual, gracias a una notación identada (con márgenes) de obligado cumplimiento. En muchos lenguajes, para separar porciones de código, se utilizan elementos como las llaves o las palabras clave *begin* y *end*. Para separar las porciones de código en Python se debe tabular hacia dentro, colocando un margen al código que iría dentro de una función o un ciclo. Esto ayuda a que todos los programadores adopten unas mismas notaciones y que los programas de cualquier persona tengan un aspecto muy similar.
Python está en movimiento y en pleno desarrollo, pero ya es una realidad y una interesante opción para realizar todo tipo de programas que se ejecuten en cualquier máquina. El equipo de desarrollo está trabajando de manera cada vez más organizada y cuentan con el apoyo de una comunidad que está creciendo rápidamente (30).

2 Materiales y métodos

Introducción

En este capítulo se muestra la encuesta dirigida a los expertos y el proceso seguido hasta llegar a un consenso final guiado por el método Delphi, que incluye la matriz booleana, el grafo de relación y la tabla de poderes. Se hace un análisis estadístico de la información en el que se definen cuáles pares de variables siguen una distribución normal, como condición final del método. Además se exponen los datos pertenecientes a la base de hechos y su origen. Se desarrolla la aplicación del método RSReduct a la base de hechos existente para obtener finalmente un reducto de la misma y así disminuir costo en tiempo de ejecución y espacio de memoria de los programas.

2.1 Método Delphi de la consulta de expertos

La Consulta de Expertos consiste en buscar los criterios de varios especialistas de disímiles disciplinas, con el propósito de validar y de tener una base de apoyo que respalde el diseño a utilizar. Para darle así un carácter científico al análisis experto, ya que los niveles de conocimientos que se involucran en este método son muy particulares y específicos. La correcta aplicación de elementos del método Delphi permite el alcance de otras vías de análisis en la investigación, además de esclarecer la necesidad de incorporar a este sistema conceptos de la lógica difusa que permiten la mayor solidez al cuestionario de preguntas a responder una vez hecho el análisis estadístico correspondiente (21).

Esta técnica tiene la ventaja de eliminar el efecto líder de otros métodos de expertos, pues los encuestados son anónimos entre sí, pero es muy importante para un correcto resultado seleccionar bien a los consultados y definir bien el campo de investigación, con preguntas precisas, cuantificables e independientes.

Objetivo: Obtener un criterio de especialistas para guiar la investigación.
Elementos básicos: Encuesta.
Selección de los expertos: Ingenieros informáticos, profesores universitarios y dirigentes.

2.1.1 Encuesta aplicada a los Expertos

2.1.1.1 Cuestionario de preguntas

¿Cuál es su ocupación?

¿Años de experiencia?

¿Cree usted que se pueda pronosticar el éxito de un estudiante mediante resultados docentes en tres años de estudios?

¿Qué aspectos de los que se muestran a continuación tendría en cuenta para dicha medida?

Lugar de procedencia

Tipo del centro de procedencia

Sexo

Vía de ingreso

Otro

Marque con una *X* la respuesta que usted considere correcta a la siguiente interrogante ¿Las variables a continuación guardan relación?

	Ninguna relación	Poca relación	Alguna relación	Relación	Mucha relación
Tipo del centro de procedencia – Promedio.					
Sexo – Promedio.					
Sexo – Programación.					
Programación – Promedio.					
Asignaturas básicas – Promedio.					
Vía de ingreso – Promedio.					
IPI – Programación.					
ESPA- Programación.					
ESPA- Asignaturas básicas.					
ESPA- Promedio.					

IPI- Promedio.					
IPI – Asignaturas básicas.					
Procedencia familiar- Promedio.					
Lugar de procedencia - Promedio.					

En la tabla 3 se muestran los pares de variables presentes en la encuesta aplicada a los expertos y el valor de mayor coincidencia en la opinión de los expertos.

Tabla 3: Resultados obtenidos en la encuesta aplicada a los expertos

	Ninguna relación	Poca relación	Alguna relación	Relación	Mucha relación
Tipo del centro de procedencia – Promedio.	X				
Sexo – Promedio.				X	
Sexo – Programación.	X				
Programación – Promedio.					X
Asignaturas básicas – Promedio.				X	
Vía de ingreso – Promedio.				X	
IPI – Programación.				X	
ESPA- Programación.				X	
ESPA- Asignaturas básicas.				X	
ESPA- Promedio.				X	
IPI- Promedio.				X	
IPI – Asignaturas básicas.				X	
Procedencia familiar- Promedio.			X		
Lugar de procedencia - Promedio.	X				

Una vez obtenidos los resultados de los expertos se le asigna un identificador a cada variable.

A1: Lugar de procedencia

A2: Promedio

A3: Tipo del centro de procedencia

A4: Sexo

A5: Programación

A6: Asignaturas básicas

A7: Vía de ingreso

A8: IPI

A9: ESPA

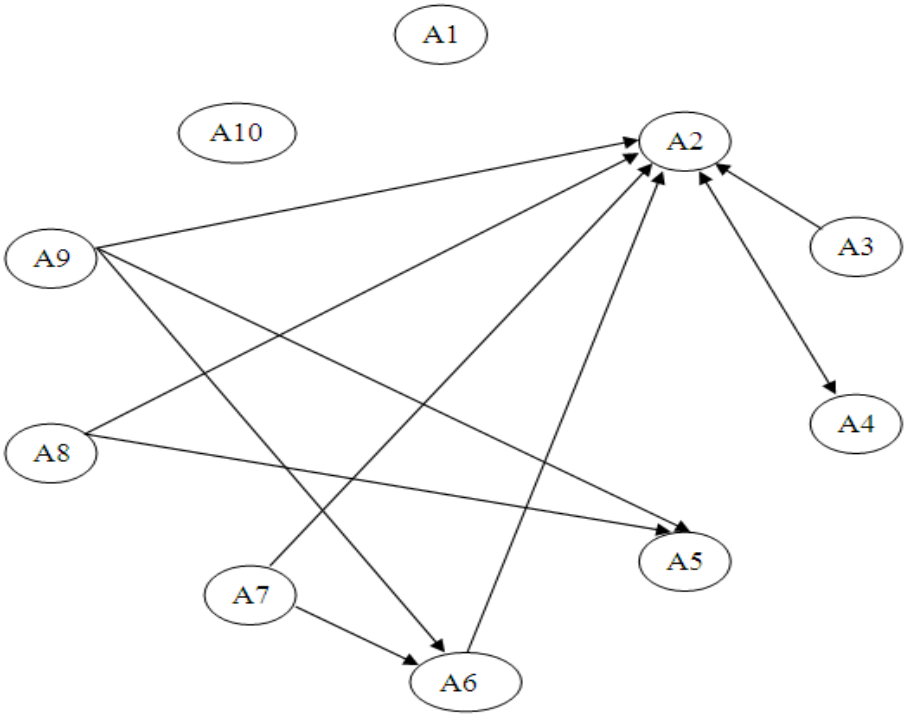
A10: Procedencia familiar

En la tabla 4, que se muestra a continuación se exponen, según la relación entre las variables, los resultados obtenidos en la encuesta aplicada a los expertos, dónde se le asigna a: ninguna relación: 0, poca relación: 0, alguna relación: 1, relación: 1, mucha relación: 1 (en ambas direcciones por ser una relación fuerte).

Tabla 4: Matriz Booleana

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1	0	0	0	0	0	0	0	0	0	0
A2	0	0	0	1	0	0	0	0	0	0
A3	0	1	0	0	0	0	0	0	0	0
A4	0	1	0	0	0	0	0	0	0	0
A5	0	0	0	0	0	0	0	0	0	0
A6	0	1	0	0	0	0	0	0	0	0
A7	0	1	0	0	0	1	0	0	0	0
A8	0	1	0	0	1	0	0	0	0	0
A9	0	1	0	0	1	1	0	0	0	0
A10	0	1	0	0	0	0	0	0	0	0

A raíz de la matriz booleana anterior se crea el grafo de dependencia siguiente, en el que se representan las relaciones entre los nodos, que este caso son las variables en cuestión. Los puntos A1, A2, A3, A4, A5, A6, A7, A8, A9 y A10 reciben el nombre de vértices y las líneas en un sentido (o dos), arcos.



La tabla 5, a continuación, muestra los poderes de cada vértice, así el vértice con mayor poder es la variable con más dependencia y de la que más dependen las otras.

Tabla 5: Tabla de poderes.

	PA	*PB*	*PA+PB*
A1	0	0	0
A2	6	0	6
A3	0	1	1
A4	1	1	2
A5	2	0	2
A6	2	1	3
A7	0	2	2
A8	0	2	2
A9	0	3	3
A10	0	0	0

Las variables con más peso, según el criterio de los expertos, son A2, A6 y A9 correspondientes a: Promedio, Asignaturas básicas y ESPA

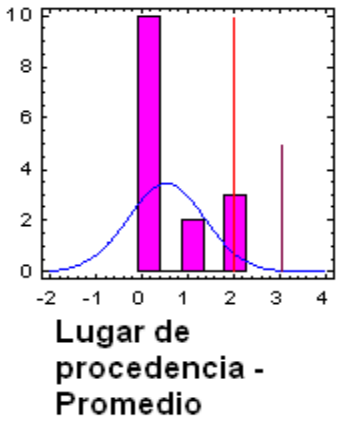
2.1.2 Análisis estadístico de la información

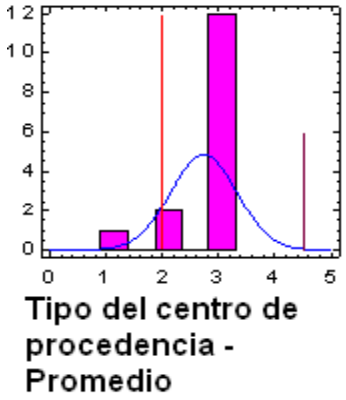
Para realizar el análisis estadístico de la información obtenida de la encuesta aplicada a los expertos se utilizó el asistente estadístico STATGRAPHICS, mediante el cual se obtuvieron los siguientes resultados a partir de las respuestas de los expertos en cada una de las variables que pertenecían al cuestionario.

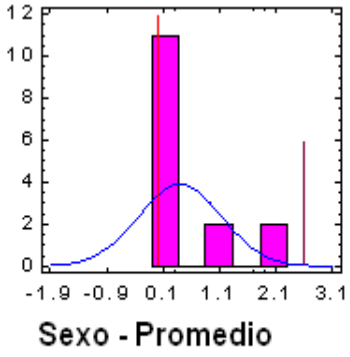
Las siguientes gráficas muestran cómo los pares de variables asociadas a la encuesta siguen una distribución normal.

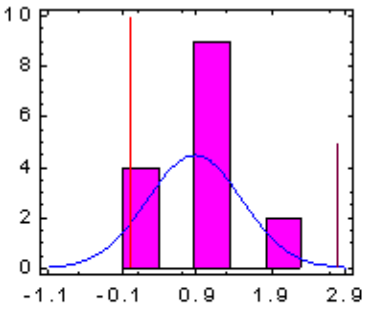
En cada gráfica se relacionan los pares de variabres en cuestión con los valores de las respuestas de los expertos según la siguiente relación:

Ninguna relación: 0.
Poca relación: 1
Alguna relación: 2
Relación: 3
Mucha relación: 4

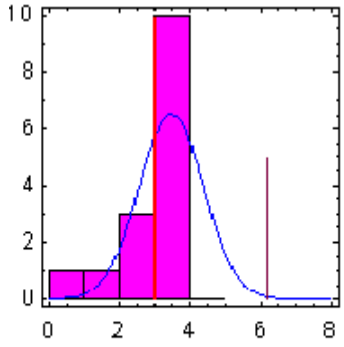




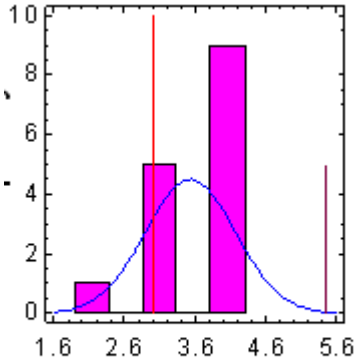




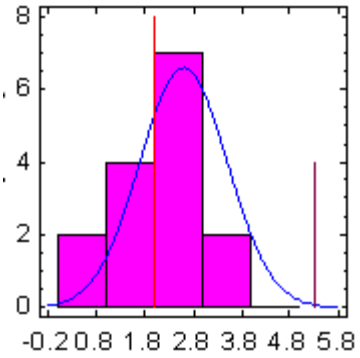
Sexo - Programación



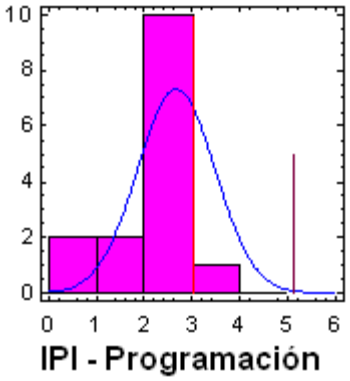
Programación - Promedio



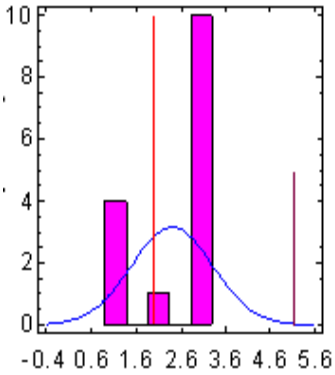
Asignaturas básicas - Promedio



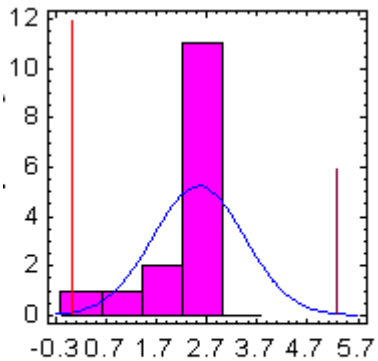
Vía de ingreso - Promedio



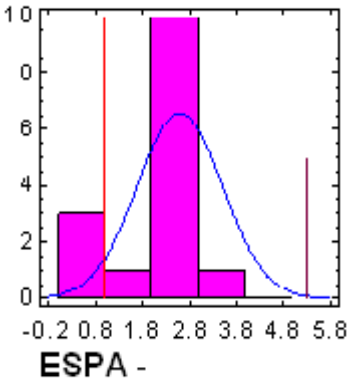
IPI - Programación



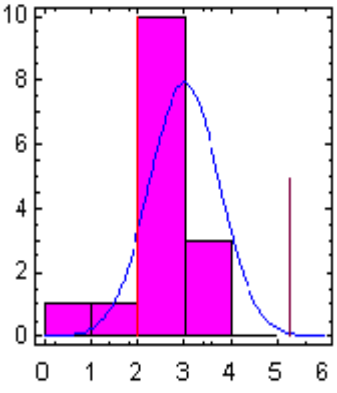
ESPA - Programación



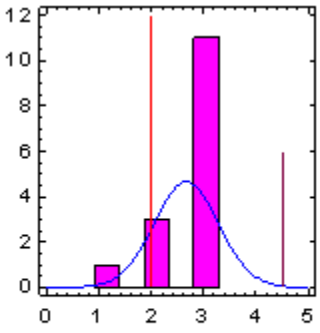
ESPA - Asignaturas básicas

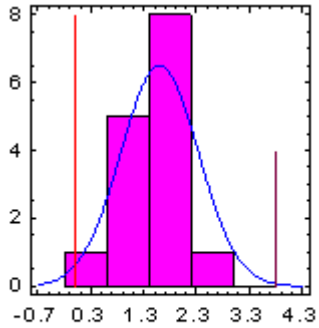


ESPA - Promedio



IPI - Promedio

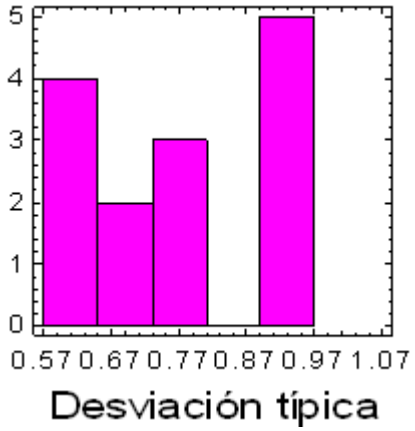




En el siguiente histograma se muestra el rango en el que se encuentran cada una de las desviaciones típicas *S* de los pares de variables graficadas anteriormente, lo que demuestra que a cada par de variables le corresponde una desviación típica con valor mayor que 0.57 y menor que 0.97, por lo que se puede afirmar que todos los pares de variables siguen una distribución normal.

Como S<1, se acepta el criterio de los expertos.

Se aceptó el criterio de los expertos de acuerdo al valor obtenido en la desviación típica.



2.2 Método RSReduct

2.2.1 Aplicación del método RSReduct para obtener un reducto

Para la aplicación de la RST en la predicción académica de estudiantes se tuvieron en cuenta una muestra de 621 estudiantes del primer semestre de primer año de la

facultad 5 de la Universidad de las Ciencias Informáticas correspondiente a los cursos 2006-2007, 2007-2008 y 2008-2009. De estos estudiantes se conoce:
Sexo (Masculino, Femenino); Provincia (Pinar del Río, Ciudad Habana, La Habana, Granma, Matanzas, Sancti Spíritus, Villa Clara, Las Tunas, Santiago de Cuba, Camagüey, Guantánamo, Holguín, Cienfuegos, Ciego de Ávila, Isla de la Juventud); Vía de ingreso (Preuniversitario, Instituto Politécnico, Politécnico, MININT, MINFAR, Orden 18), Tipo del Centro de procedencia (IPVCE, IPUEC, Técnico Medio Informática, EMCC, DEPORTE, IPU, Técnico Medio); Actividad laboral de la madre y del padre (Obrero, Técnico, Ama de casa, Profesional, Otra actividad, Jubilado, Campesino); Nivel escolar de la madre y del padre (Primaria, Secundaria, Técnico Medio, Preuniversitario, Universitario, Ninguno terminado, Obrero calificado); Salario de la madre y del padre (0_100, 101_200, 201_300, 301_400, 401_500, 501_600, 601_700, 701_800, 801_900, 901_1000); Notas de las asignaturas Matemática1, Idioma Extranjero 1, Educación Física1, Introducción a la Programación, Matemática Discreta, las que toman los valores (2, 3, 4 ó 5); Además cuenta con un atributo de decisión que puede tomar los valores (si, no), estos valores están a asociados a los resultados finales del semestre (aprobado, desaprobado). Los datos obtenidos están expuestos en el anexo 1 (ver anexo 1).

2.2.2 RSReduct como método de selección de rasgos

Inicialmente los datos obtenidos, como toda la información que se obtiene de una base de datos tan compleja como la base se datos de la UCI Akademos, en algunos casos son redundantes y contienen muchos datos que en realidad su única función es ocupar espacio en memoria; sin embargo se puede obtener el mismo conocimiento de una base de datos sin información que no es necesaria que de una que contiene toda la información, el problema está en qué datos seleccionar. Para esto en este caso se utilizan algoritmos de selección de rasgos basado en la teoría de los conjuntos aproximados.

Los conjuntos aproximados se aplican eficientemente en la reducción de atributos o selección de rasgos sobre la base del concepto de reducto.

Sin embargo, esta beneficiosa alternativa se encuentra limitada por el hecho de que encontrar esos conjuntos mínimos de atributos constituye un problema de la teoría de

conjuntos aproximados, dado que computar todos los reductos es una tarea con alto costo computacional. Diversos autores han propuesto métodos para el cálculo de reductos a través de los conjuntos aproximados.

El método aplicado se llama RSReduct y lo desarrollaron los doctores en ciencias Yailé Caballero, Delia Álvarez, Analay Baltá, Rafael Bello y María García. RSReduct es un método que trata de encontrar un reducto de manera que éste sea lo suficientemente bueno para el análisis de datos en tiempos aceptables. Para ello se utiliza la búsqueda heurística como estrategia de búsqueda, debido a que si se tomara en cuenta la variante exhaustiva o completa, el consumo de tiempo y recursos de cómputo sería bastante grande y la no determinística dificulta saber cuándo aparece un subconjunto mínimo.

El método consiste en un algoritmo glotón que comienza por un conjunto vacío de atributos, y a través de heurísticas va formando un reducto mediante la selección de los atributos uno a uno de una lista, hasta que se cumple la condición de parada; en la lista de atributos; estos se encuentran ordenados según el valor arrojado por la función de evaluación heurística para cada uno de estos. Para la construcción de las funciones de evaluación heurística se siguen criterios del método ID3 con respecto a la entropía y la ganancia de los atributos, dependencia entre atributos mediante los conjuntos aproximados, así como la opción de otorgar costos a los atributos, es decir, manipulación de atributos con costos diferentes (31).

En este algoritmo se utilizan las medidas R(A) y H(A).

$$R(A) = \sum_{i=1}^{k} \frac{S_i}{S} e^{(1-c)}$$

En la expresión *k* es el número de valores diferentes del rasgo *A*. C_i es el número de clases diferentes presentes en los objetos que tienen el valor *i* para el rasgo *A* y $\frac{S_i}{S}$ es la frecuencia relativa del valor *i* en *S* (cantidad de objetos con el valor *i* en el rasgo *A* sobre la cantidad de objetos de toda la muestra). La principal idea de esta medida es maximizar la heterogeneidad entre objetos que pertenecen a clases diferentes y

minimizar la homogeneidad entre aquellos que son de la misma clase, además $0 \leq R(A) \leq 1$

H(A) se obtiene a través del siguiente algoritmo:

I. Se calcula el vector *R (T)= (R (A1) R(A2),...).* Para todos los atributos del problema se calcula su R(A) y así con todos los valores se forma el vector R (T).

II. Se determinan los *n* mejores atributos por los cálculos del paso anterior. El valor de *n* se puede seleccionar por el usuario. Como resultado de este paso se obtiene el vector RM = (R (Ai), R (Aj),...) con n = $|RM|$

III. Se determinan las combinaciones de *n* en *p, siendo p* la cantidad de elementos que contendrá cada combinación. Ambos valores(*n* y *p*) son seleccionados por el usuario. $Comb = (\{A_i, A_j, A_k\},...,\{A_l, A_t, A_p\})$ desde los atributos seleccionados en el paso II se obtiene el vector de combinaciones $C_p^n = \frac{n!}{p!(n-p)!}$.

IV. Se calcula el grado de dependencia de las clases con respecto a cada una de las combinaciones obtenidas en el paso anterior. Como resultado de este paso se obtiene el vector de dependencias: $DEP = (k(comb1, d),...,k(combn, d))$ donde k representa la medida para el grado de dependencia entre atributos de los conjuntos aproximados con respecto a los valores de decisión *d.*

V. Para cada atributo A se calcula H(A) según la siguiente ecuación:

$H(A) = \sum_{\forall i | A \in combi} k(combi, d)$

Una vez obtenida la heurística *RG(A) = H(A) + G(A)* se sigue con el RSReduct.

P1. Formar la tabla de distinción.

Sea B matriz binaria ($M^2 - M$)/ *2x (N+1).* Cada fila corresponde a un par de objetos diferentes. Cada columna de esta matriz corresponde a un atributo, la última columna corresponde a la decisión (tratada como un atributo).

Sea *b ((k, n), i)* un elemento de *B* correspondiente al par *Ok, On*, y al atributo *i*, para $i \in \{1,... N\}$:

$$b((k,n),i)=\begin{Bmatrix}1, sia_i(O_k)-\Re a_i(O_n)\\ 0, sia_i(O_k)\Re a_i(O_n)\end{Bmatrix} para i\in 1,...,N$$

$$b((k,n),N+1)=\begin{Bmatrix}0, sid_i(O_k)\neq d_i(O_n)\\ 0, sid_i(O_k)=d_i(O_n)\end{Bmatrix}$$

Donde $\Re$ es una relación de similaridad en dependencia del tipo del atributo a_i.

P2. Para cada atributo *A* se calcula el valor de RG(A) utilizando la heurística. Se forma una lista ordenada de atributos comenzando por el atributo más relevante (el que maximice RG(A)).

Heurística. RG(A) = R(A) + H(A)

P3. Se tiene *i = 1*, *R* = conjunto vacío y se tiene $A_1, A_2, ..., A_n$ lista ordenada de atributos según el paso 2, si *i≤n* entonces $R = R \cup A_i$, *i=i+1.*

P4. Si *R* satisface la Condición I parar (32).

(Condición I) $\forall k,n \ \forall a_i \in R \quad a_i(o_k)\Re a_i(o_n) \Rightarrow d(o_k)=d(o_n)$

2.2.3 Implementación del método RSReduct

Tabla 6. Descripción de la base de casos Estudiantes

Nombre del atributo	*Valores de su dominio*
Sexo	Masculino, Femenino
Provincia	Pinar del Rio, Ciudad Habana, La Habana, Granma, Matanzas, Sancti Espíritus, Villa Clara, Las_Tunas, Santiago de Cuba, Camagüey, Guantánamo, Holguín, Cienfuegos, Ciego de Ávila, Isla de la Juventud
Vía de ingreso	Preuniversitario, Instituto Politécnico, Politécnico, MININT, MINFAR, Orden 18
Tipo del centro	IPVCE, IPUEC, Técnico Medio Informática, EMCC,

de procedencia	DEPORTE, IPU, Técnico Medio
Actividad laboral de la madre	Obrero, Tecnico, Ama de Casa, Profesional, Otra Actividad, Jubilado, Campesino
Actividad laboral del padre	Obrero, Tecnico, Ama de Casa, Profesional, Otra Actividad, Jubilado, Campesino, Secundaria, Tecnico Medio, Preuniversitario, Universitario
Nivel escolar de la madre.	Primaria, Secundaria, Tecnico Medio, Preuniversitario, Universitario, Ninguno Terminado, Obrero Calificado
Nivel escolar del padre	Primaria, Secundaria, Tecnico Medio, Preuniversitario, Universitario, Ninguno Terminado, Obrero Calificado
Salario de la madre	0_100, 101_200, 201_300, 301_400, 401_500, 501_600, 601_700, 701_800, 801_900, 901_1000
Salario del padre	0_100, 101_200, 201_300, 301_400, 401_500, 501_600, 601_700, 701_800, 801_900, 901_1000
Matemática I	2, 3, 4, 5
Idioma Extranjero I	2, 3, 4, 5
Educación Física I	2, 3, 4, 5
Introducción a la Programación	2, 3, 4, 5
Matemática Discreta	2, 3, 4, 5
Atributo decisión	si, no

Para la implementación del método se crearon cuatro clases en Python cuyo diagrama UML se muestra en la Figura 7: Diagrama UML de las clases del método RSReduct.

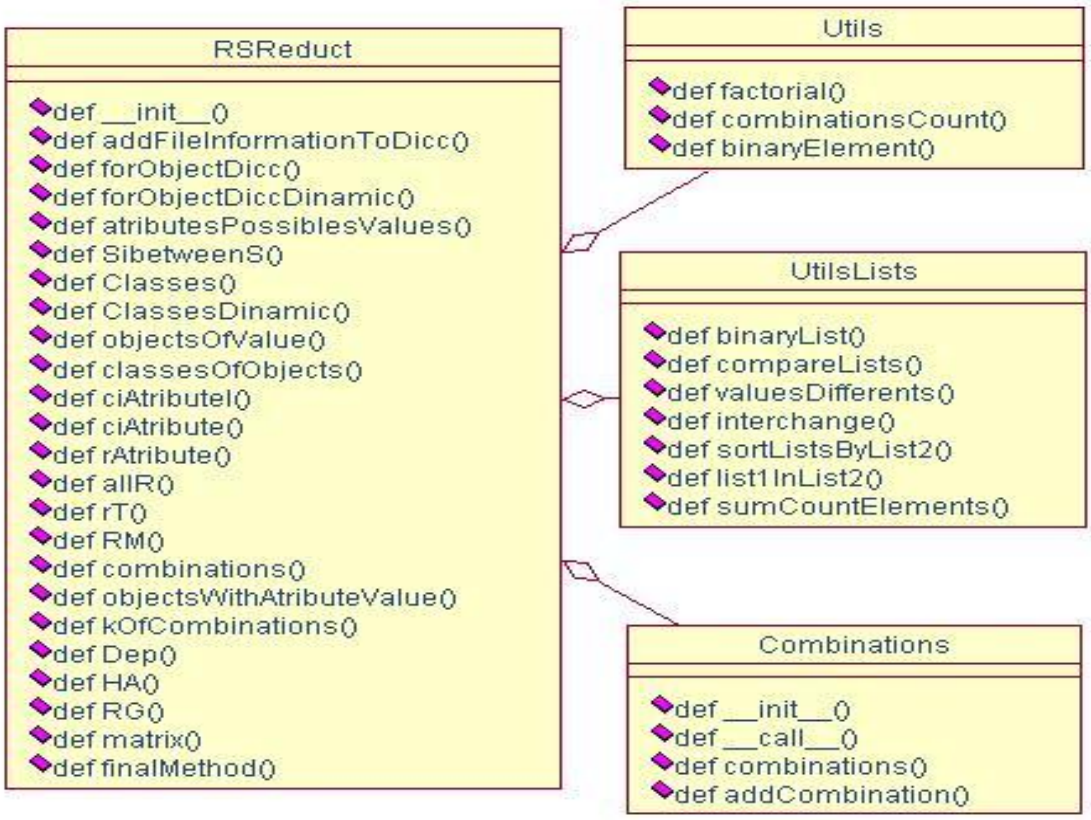


Figura 7: Diagrama UML de las clases del método RSReduct.

Fuente: elaboración propia.

Una vez ejecutado el código para la base de casos se obtuvo el siguiente subconjunto de rasgos: {Matemática I, Nivel escolar de la madre, Nivel escolar del padre Tipo del Centro de procedencia, sexo}, para un valor de n = 5 y p = 3.

3 Resultados

Introducción

En este capítulo se establecen comparaciones entre los métodos existentes, utilizando como software de apoyo el WEKA, y el método RSReduct. Se obtienen resultados al calcular la calidad de la clasificación del sistema de información con todos los datos y con los datos obtenidos después de aplicar el algoritmo de selección de rasgos. Además se obtienen reglas que clasifican los objetos de la base de hechos.

3.1 Análisis de los resultados

Para analizar los resultados se utilizaron varios clasificadores de tipo árbol de decisión. En la tabla a continuación se muestran los resultados obtenidos.

Tabla 7: Comparación entre agunos clasificadores de tipo árbol

Tipo del clasificador	Calidad de la clasificación	Tiempo de ejecución (s)
ADTree	0.99	0.10
BFTree	0.98	1.50
DecisionStump	0.98	0.08
FT	0.99	1.00
Id3	0.97	0.50
J48	0.98	0.90
J48fraft	0.98	0.20
RandomTree	0.94	0.10

Aplicando un clasificador de tipo árbol de decisión Adtree, por ser el de mayor calidad de la clasificación se obtuvo el árbol representado en la Figura 8: ADtree para la base de hechos completa.

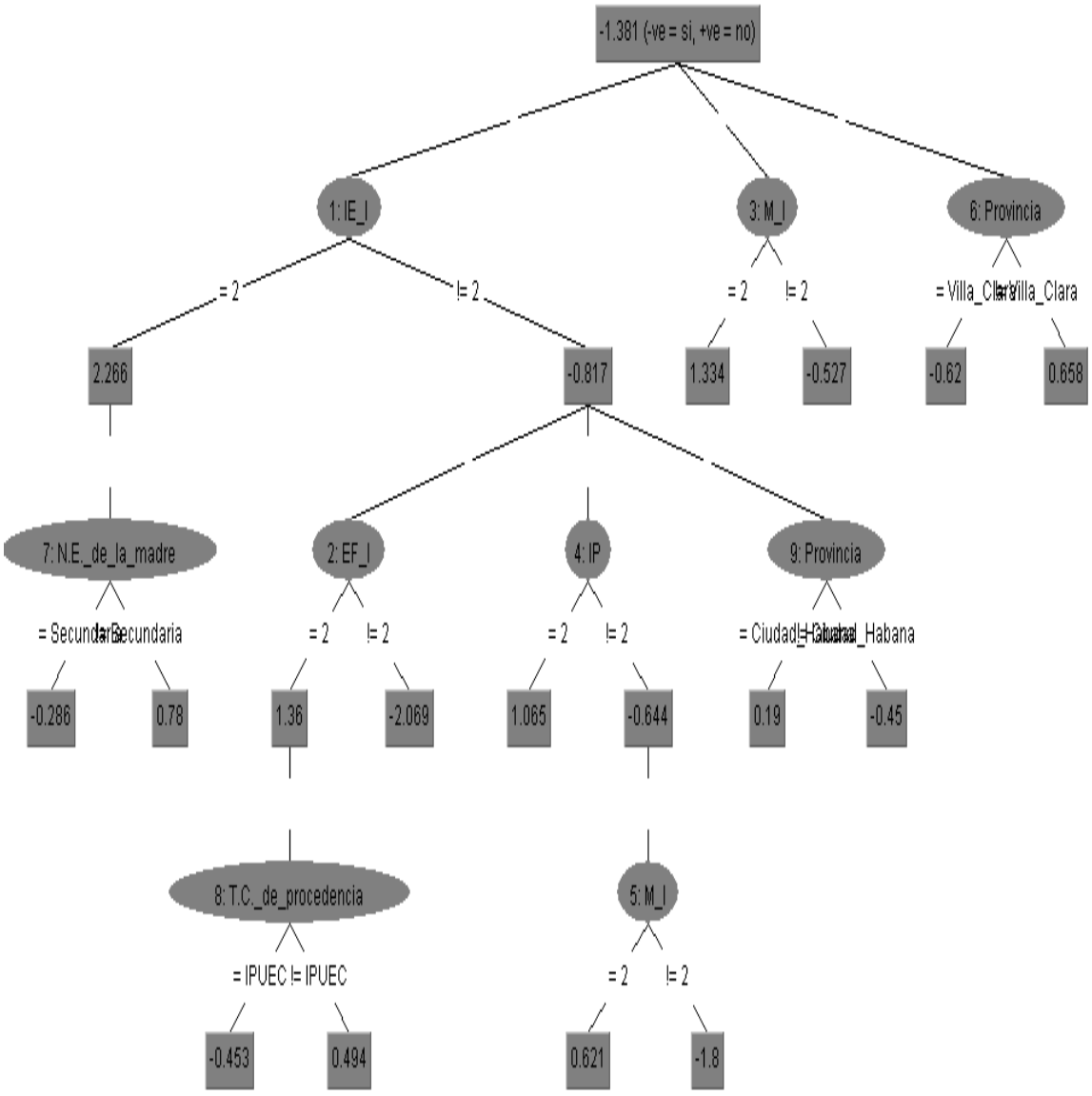


Figura 8: ADtree para la base de hechos completa

Fuente: resultado de aplicar Adtree en el software WEKA.

La calidad de la clasificación del ADtree para la base de hechos completa es de 0.99. Se aplicó el mismo procedimiento con la base de hechos resultante después de aplicar el método RSReduct. El resultado se muestra en la Figura 9: ADtree para la base de hechos con los atributos seleccionados

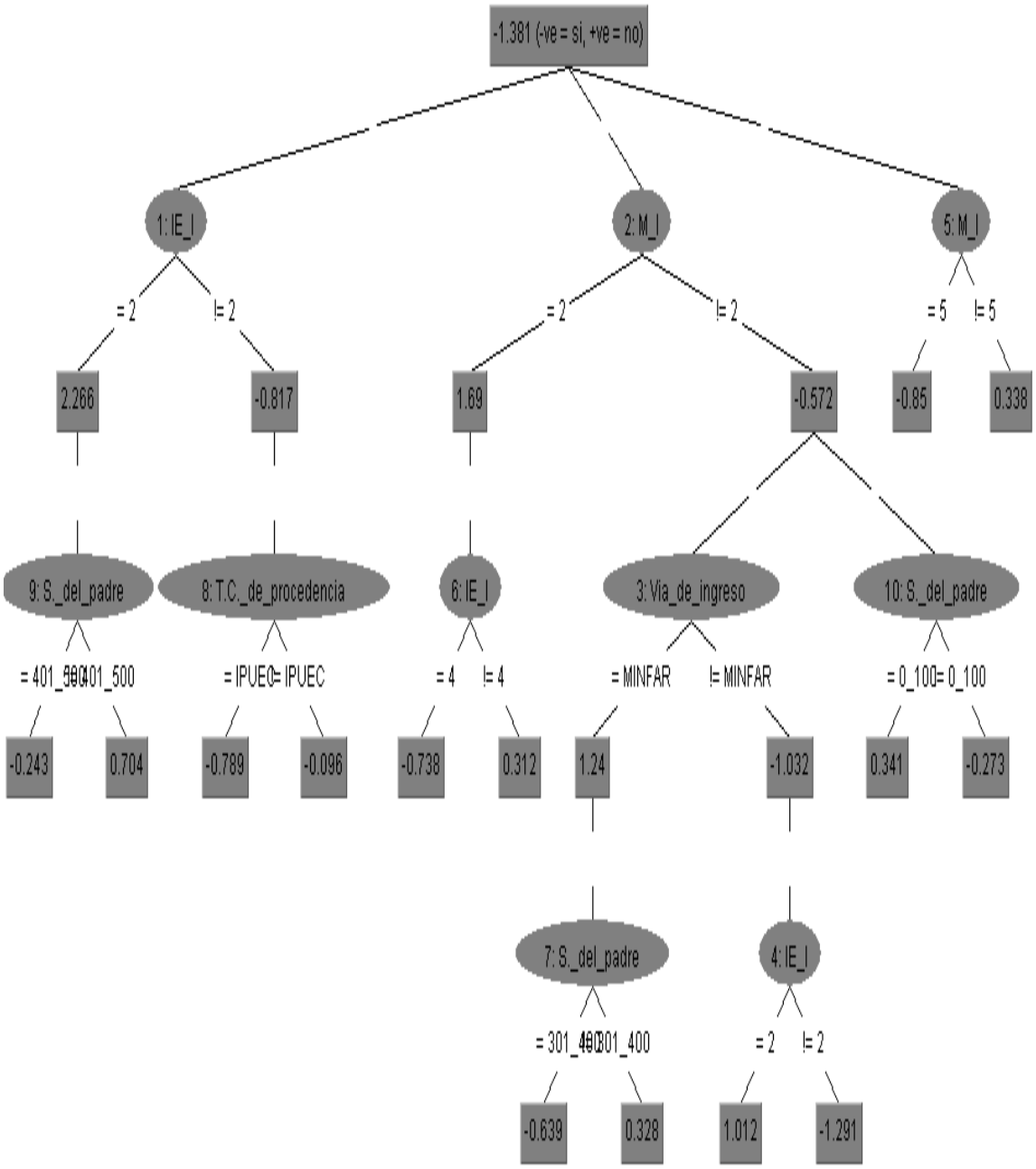


Figura 9: ADtree para la base de hechos con los atributos seleccionados

Fuente: resultado de aplicar Adtree en el software WEKA.

La calidad de la clasificación del ADtree para la base de hechos con los atributos seleccionados es de 0.987. Véase que la diferencia es menor a 0.01, esto indica que el reducto es efectivo.

Además se generaron un conjunto de reglas aplicando varios clasificadores de tipo reglas. En la siguiente tabla se muestran los resultados obtenidos.

Tabla 8: Comparación entre algunos clasificadores de tipo reglas.

Tipo del clasificador	Calidad de la clasificación	Tiempo de ejecución (s)
NNge	0.99	0.10
Conjuntive hule	0.98	46.70
Decision tablee	0.98	15.08
One R	0.98	1.00
Oidor	0.97	0.30
Error	0.94	10.00

Por haberse obtenido la mayor calidad de la clasificación para el clasificador de tipo regla NNge se generaron con el software WEKA las siguientes reglas.

=== Classifier model (full training set) ===

NNGE classifier

Rules generated:

1. class si IF : Sexo in {Masculino,Femenino} ^ Provincia in {Pinar_del_Rio,Ciudad_Habana,La_Habana,Granma,Matanzas,Sancti_Spiritus,Villa_Clara,Santiago_de_Cuba,Camaguey,Guantanamo,Holguin,Cienfuegos} ^ Via_de_ingreso in {MINFAR} ^ T.C._de_procedencia in {EMCC} ^ A.L._de_la_madre in {Obrero,Tecnico,Ama_de_Casa,Profesional,Otra_Actividad} ^ A.L._del_padre in {Obrero,Tecnico,Profesional,Otra_Actividad,Jubilado,Campesino,Universitario} ^ N.E._de_la_madre in {Primaria,Secundaria,Tecnico_Medio,Preuniversitario,Universitario} ^ N.E._del_padre in {Secundaria,Tecnico_Medio,Preuniversitario,Universitario,Ninguno_Terminado} ^ S._de_la_madre in {0_100,201_300,301_400,401_500,501_600,701_800} ^ S._del_padre in {0_100,101_200,201_300,301_400,401_500,501_600,601_700,701_800,901_1000} ^ M_I in {3,4,5} ^ IE_I in {3,4,5} ^ EF_I in {2,3,4,5} ^ IP in {3,4,5} ^ MD in {3,4,5} (29)

2. class si IF : Sexo in {Masculino,Femenino} ^ Provincia in {Pinar_del_Rio,Ciudad_Habana,La_Habana,Granma,Matanzas,Sancti_Spiritus,Villa_Clara,Las_Tunas,Santiago_de_Cuba,Camaguey,Guantanamo,Holguin,Cienfuegos,Ciego_de_Avila,Isla_de_la_Juventud} ^ Via_de_ingreso in {Preuniversitario,Instituto_PoliTecnico,PoliTecnico,MININT,Orden_18} ^ T.C._de_procedencia in {IPVCE,IPUEC,Tecnico_Medio_Informatica,EMCC,DEPORTE,IPU,Tecnico_Medio} ^ A.L._de_la_madre in {Obrero,Tecnico,Ama_de_Casa,Profesional,Otra_Actividad,Jubilado,Campesino} ^ A.L._del_padre in {Obrero,Tecnico,Ama_de_Casa,Profesional,Otra_Actividad,Jubilado,Campesino,Secundaria,Preuniversitario,Universitario} ^ N.E._de_la_madre in {Primaria,Secundaria,Tecnico_Medio,Preuniversitario,Universitario,Ninguno_Terminado,Obrero_Calificado} ^ N.E._del_padre in {Primaria,Secundaria,Tecnico_Medio,Preuniversitario,Universitario,Obrero_Calificado} ^ S._de_la_madre in {0_100,101_200,201_300,301_400,401_500,501_600,601_700,701_800,801_900,901_1000} ^ S._del_padre in {0_100,101_200,201_300,301_400,401_500,501_600,601_700,701_800,801_900,901_1000} ^ M_I in {3,4,5} ^ IE_I in {3,4,5} ^ EF_I in {2,3,4,5} ^ IP in {2,3,4,5} ^ MD in {2,3,4,5} (544)

3. class si IF : Sexo in {Femenino} ^ Provincia in {Ciudad_Habana,Santiago_de_Cuba,Camaguey} ^ Via_de_ingreso in {Preuniversitario,Instituto_PoliTecnico} ^ T.C._de_procedencia in {IPUEC,Tecnico_Medio_Informatica} ^ A.L._de_la_madre in {Profesional} ^ A.L._del_padre in {Tecnico_Medio} ^ N.E._de_la_madre in {Tecnico_Medio,Preuniversitario} ^ N.E._del_padre in {Universitario} ^ S._de_la_madre in {201_300,301_400} ^ S._del_padre in {301_400,401_500} ^ M_I in {3,5} ^ IE_I in {3,4,5} ^ EF_I in {4,5} ^ IP in {3,5} ^ MD in {4} (3)

4. class si IF : Sexo in {Masculino,Femenino} ^ Provincia in {La_Habana,Camaguey,Guantanamo} ^ Via_de_ingreso in {Preuniversitario,Instituto_PoliTecnico} ^ T.C._de_procedencia in {IPVCE,Tecnico_Medio_Informatica,DEPORTE} ^ A.L._de_la_madre in {Obrero,Tecnico,Profesional} ^ A.L._del_padre in {Obrero,Tecnico,Profesional} ^ N.E._de_la_madre in {Tecnico_Medio,Universitario,Obrero_Calificado} ^ N.E._del_padre in {Universitario,Obrero_Calificado} ^ S._de_la_madre in {101_200,201_300,901_1000} ^ S._del_padre in {201_300,501_600,601_700} ^ M_I in {2} ^ IE_I in {4} ^ EF_I in {3,4,5} ^ IP in {4} ^ MD in {3,5} (3)

5. class si IF : Sexo in {Masculino} ^ Provincia in {Ciudad_Habana} ^ Via_de_ingreso in {Preuniversitario} ^ T.C._de_procedencia in {IPUEC} ^ A.L._de_la_madre in {Tecnico} ^ A.L._del_padre in {Profesional} ^ N.E._de_la_madre in {Preuniversitario} ^ N.E._del_padre in {Preuniversitario} ^ S._de_la_madre in {301_400} ^ S._del_padre in {301_400} ^ M_I in {2} ^ IE_I in {3} ^ EF_I in {3} ^ IP in {4} ^ MD in {5} (1)

6. class si IF : Sexo in {Femenino} ^ Provincia in {Pinar_del_Rio,Las_Tunas} ^ Via_de_ingreso in {Preuniversitario,Instituto_PoliTecnico} ^ T.C._de_procedencia in {IPUEC,Tecnico_Medio_Informatica} ^ A.L._de_la_madre in {Obrero,Profesional} ^ A.L._del_padre in {Obrero,Profesional,Otra_Actividad} ^ N.E._de_la_madre in {Secundaria,Universitario} ^ N.E._del_padre in {Secundaria,Universitario} ^ S._de_la_madre in {101_200,201_300,401_500} ^ S._del_padre in {101_200,401_500,701_800} ^ M_I in {2} ^ IE_I in {3,4} ^ EF_I in {3} ^ IP in {3,5} ^ MD in {5} (3)

7. class si IF : Sexo in {Femenino} ^ Provincia in {Villa_Clara} ^ Via_de_ingreso in {Preuniversitario} ^ T.C._de_procedencia in {IPUEC} ^ A.L._de_la_madre in {Profesional} ^ A.L._del_padre in {Profesional} ^ N.E._de_la_madre in {Universitario} ^ N.E._del_padre in {Universitario} ^ S._de_la_madre in {401_500} ^ S._del_padre in {901_1000} ^ M_I in {2} ^ IE_I in {3} ^ EF_I in {2} ^ IP in {3} ^ MD in {5} (1)

8. class no IF : Sexo in {Masculino,Femenino} ^ Provincia in {Ciudad_Habana,Granma,Matanzas,Sancti_Spiritus,Villa_Clara,Santiago_de_Cuba,Camaguey,Guantanamo,Holguin,Cienfuegos} ^ Via_de_ingreso in {Preuniversitario,Instituto_PoliTecnico} ^ T.C._de_procedencia in {IPVCE,IPUEC,Tecnico_Medio_Informatica,DEPORTE,IPU} ^ A.L._de_la_madre in {Obrero,Ama_de_Casa,Profesional,Otra_Actividad} ^ A.L._del_padre in {Obrero,Tecnico,Profesional,Otra_Actividad,Jubilado,Preuniversitario,Universitario} ^ N.E._de_la_madre in {Secundaria,Tecnico_Medio,Preuniversitario,Universitario} ^ N.E._del_padre in {Secundaria,Tecnico_Medio,Preuniversitario,Universitario,Obrero_Calificado} ^ S._de_la_madre in {0_100,201_300,301_400,401_500,501_600} ^ S._del_padre in {0_100,301_400,401_500,501_600,701_800} ^ M_I in {2} ^ IE_I in {2,3} ^ EF_I in {2,3} ^ IP in {2,3,4,5} ^ MD in {2,3,5} (26)

9. class no IF : Sexo in {Femenino} ^ Provincia in {Pinar_del_Rio} ^ Via_de_ingreso in {Instituto_PoliTecnico} ^ T.C._de_procedencia in {Tecnico_Medio_Informatica} ^ A.L._de_la_madre in {Otra_Actividad} ^ A.L._del_padre in {Obrero} ^ N.E._de_la_madre in {Tecnico_Medio} ^ N.E._del_padre in {Tecnico_Medio} ^ S._de_la_madre in {701_800} ^ S._del_padre in {301_400} ^ M_I in {2} ^ IE_I in {5} ^ EF_I in {2} ^ IP in {3} ^ MD in {5} (1)

10. class no IF : Sexo in {Masculino,Femenino} ^ Provincia in {Ciudad_Habana} ^ Via_de_ingreso in {MINFAR} ^ T.C._de_procedencia in {EMCC} ^ A.L._de_la_madre in {Obrero,Profesional} ^ A.L._del_padre in {Profesional,Otra_Actividad} ^ N.E._de_la_madre in {Secundaria,Universitario} ^ N.E._del_padre in {Preuniversitario,Universitario} ^ S._de_la_madre in {201_300,401_500} ^ S._del_padre in {0_100,401_500} ^ M_I in {3,4} ^ IE_I in {3,4} ^ EF_I in {2} ^ IP in {2} ^ MD in {5} (2)

11. class no IF : Sexo in {Masculino,Femenino} ^ Provincia in {Ciudad_Habana,Matanzas,Santiago_de_Cuba,Camaguey,Holguin} ^ Via_de_ingreso in {Preuniversitario,MINFAR} ^ T.C._de_procedencia in {IPVCE,IPUEC,EMCC,DEPORTE,IPU} ^ A.L._de_la_madre in {Obrero,Profesional,Otra_Actividad} ^ A.L._del_padre in

{Obrero,Profesional,Preuniversitario} ^ N.E._de_la_madre in {Tecnico_Medio,Preuniversitario,Universitario} ^ N.E._del_padre in {Tecnico_Medio,Universitario} ^ S._de_la_madre in {301_400,401_500} ^ S._del_padre in {301_400,401_500} ^ M_I in {3} ^ IE_I in {2} ^ EF_I in {2,4} ^ IP in {2,3,4,5} ^ MD in {3,5} (5)

12. class si IF : Sexo in {Masculino} ^ Provincia in {Villa_Clara} ^ Via_de_ingreso in {Preuniversitario} ^ T.C._de_procedencia in {IPUEC} ^ A.L._de_la_madre in {Ama_de_Casa} ^ A.L._del_padre in {Otra_Actividad} ^ N.E._de_la_madre in {Secundaria} ^ N.E._del_padre in {Tecnico_Medio} ^ S._de_la_madre in {0_100} ^ S._del_padre in {401_500} ^ M_I in {3} ^ IE_I in {2} ^ EF_I in {3} ^ IP in {4} ^ MD in {5} (1)

13. class no IF : Sexo in {Masculino} ^ Provincia in {Las_Tunas} ^ Via_de_ingreso in {Preuniversitario} ^ T.C._de_procedencia in {IPVCE} ^ A.L._de_la_madre in {Tecnico} ^ A.L._del_padre in {Profesional} ^ N.E._de_la_madre in {Tecnico_Medio} ^ N.E._del_padre in {Universitario} ^ S._de_la_madre in {301_400} ^ S._del_padre in {301_400} ^ M_I in {2} ^ IE_I in {2} ^ EF_I in {2} ^ IP in {2} ^ MD in {5} (1)

14. class no IF : Sexo in {Masculino} ^ Provincia in {Holguin} ^ Via_de_ingreso in {Preuniversitario} ^ T.C._de_procedencia in {IPUEC} ^ A.L._de_la_madre in {Profesional} ^ A.L._del_padre in {Tecnico_Medio} ^ N.E._de_la_madre in {Preuniversitario} ^ N.E._del_padre in {Universitario} ^ S._de_la_madre in {401_500} ^ S._del_padre in {0_100} ^ M_I in {3} ^ IE_I in {2} ^ EF_I in {3} ^ IP in {5} ^ MD in {4} (1)

Stat :

class si : 8 exemplar(s) including 5 Hyperrectangle(s) and 3 Single(s).
class no : 6 exemplar(s) including 3 Hyperrectangle(s) and 3 Single(s).

Total : 14 exemplars(s) including 8 Hyperrectangle(s) and 6 Single(s).

Time taken to build model: 0.2 seconds

=== Summary ===

Correctly Classified Instances	615	99.0338 %
Incorrectly Classified Instances	6	0.9662 %
Kappa statistic	0.9092	
Mean absolute error	0.0097	
Root mean squared error	0.0983	
Relative absolute error	8.7337 %	
Root relative squared error	42.0558 %	
Total Number of Instances	621	

=== Confusion Matrix ===

```
  a   b   <-- classified as
583   2 |   a = si
  4  32 |   b = no
```

Como se muestra en la matriz de confusión sólo están mal clasificados 6 elementos de los 621 del total; dos que pertenecen a la clase si y cuatro que pertenecen a la clase no, lo que brinda un 0.99% de la información bien clasificada.

A partir de los datos obtenidos en las reglas anteriormente expuestas se llegó a la siguiente relación: En la tabla 9 se muestran los resultados.

Tabla 9: Relación entre las reglas, sus clases y los elementos que clasifica.

Numero de la regla	**Clase a la que pertenece**		**Cantidad de objetos (estudiantes)que clasifica**
	si	no	
1	X		29

2	X		544
3	X		3
4	X		3
5	X		1
6	X		3
7	X		1
8		X	26
9		X	1
10		X	2
11		X	5
12	X		1
13		X	1
14		X	1
Total	**8**	**6**	**621**

A continuación se hace una descripción de las reglas que clasifican la mayor cantidad de objetos. Estas reglas son: la 2, la 1 y la 8, según la cantidad de elementos que clasifiquen.

Descripción de la regla 2: Se puede pronosticar que un estudiante es académicamente exitoso si no tiene vía de ingreso del Ministerio de las Fuerzas Armadas Revolucionarias (MINFAR) y aprobó todas las asignaturas.

Descripción de la regla 1: Se puede pronosticar que un estudiante es académicamente exitoso si no pertenece a Isla de la juventud, Cienfuegos o Ciego de Ávila; tiene vía de ingreso del MINFAR; proviene de una EMCC (Escuela Militar Camilo Cienfuegos); sus padres no son jubilados o campesinos; los padres tienen algún nivel escolar; el salario de la madre es mayor a $200 y aprobó todas las asignaturas.

Descripción de la regla 8: Se puede pronosticar que un estudiante no es académicamente exitoso si: no pertenece a Pinar del Río, La Habana, Las Tunas, Ciego de Ávila o Isla de la juventud; tiene vía de ingreso preuniversitario o un instituto politécnico; no proviene de una EMCC; sus padres son obreros, profesionales o

realizan otra actividad; los padres tienen nivel escolar primaria; el salario de los padres es menor a $500 y suspendió alguna de las asignaturas.

En la tabla 6 se muestra un estudio cuantitativo de los reductos obtenidos utilizando RSReduct para la base de casos en cuestión y utilizando el software WEKA; se calcularon la longitud del reducto y el tiempo de ejecución, en segundos, para cada caso.

Tabla 10: Resultados de calcular un reducto con RSReduct y con WEKA

	Reducto	Longitud del reducto	Tiempo de ejecución
RSReduct	Matemática I, Nivel escolar de la madre, Nivel escolar del padre, Tipo del Centro de procedencia, sexo.	5	2.15 min
Software WEKA	Matemática I, Nivel escolar del padre, Tipo del Centro de procedencia, Introducción a la programació, Actividad Laboral de la madre, Matemática Discreta.	6	2 min

Para ilustrar un poco mejor cuán buena fue la reducción del método utilizado, la Figura 10: Representación gráfica de los resultados experimentales de RSReduct para la base de casos muestra el tamaño original de la base de casos y el tamaño al que fue reducida al usar la heurística de RSReduct.

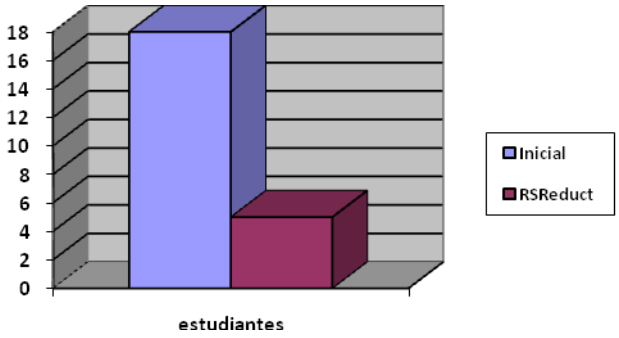


Figura 10: Representación gráfica de los resultados experimentales de RSReduct para la base de casos.
Fuente: elaboración propia.

Estos resultados experimentales se compararon estadísticamente en aras de buscar diferencias con otros métodos de selección de rasgos implementados con técnicas de reconocimiento de patrones. Inicialmente se probaron cada uno de los algoritmos mencionados anteriormente en el mismo procesador y para la misma base de casos que RSReduct. En la siguiente tabla se muestran los resultados obtenidos.

Tabla 11: Selección de rasgos con diferentes técnicas de reconocimiento de patrones

Método	**Longitud del reducto**
Primero el Mejor	3
Búsqueda exhaustiva	3
Búsqueda FCBF	16
Búsqueda genética	5
Búsqueda aleatoria	5

Conclusiones

La aplicación de métodos de reducción de atributos para problemas de Minería de datos es un paso de avance en el procesamiento de datos. Específicamente en el campo de la predición académica resulta de gran utilidad debido a que se eliminan los datos que menos valor tienen para este proceso.

Se aplicó la teoría de conjuntos aproximados en la predicción académica de estudiantes teniendo como algoritmo de selección de rasgos el método RSReduct, el cual fue implementado. El criterio de los expertos sirvió para verificar que los atributos seleccionados por el método RSReduct fueron los adecuados (en su mayoría). Se establecieron comparaciones entre la calidad de la clasificación del sistema completo y sobre el reducto para mostrar que no existen diferencias considerables en dicha medida. Se aplicó el clasificador de tipo árbol ADTree al caso de estudio por ser este el mejor clasificador de tipo árbol en este caso. Se obtuvo el subconjunto de atributos más importantes para la predición académica, los que indican que no sólo las notas académicas son importantes en los resultados finales.

Recomendaciones

- Continuar la implementación de otros campos de la teoría de conjuntos aproximados.
- Vincular los resultados de esta investigación en otros campos como el deporte y la salud.
- Implementar un sistema capaz de diagnosticar el éxito académico de estudiantes teniendo como base esta investigación.
- Incorporar los resultados obtenidos en esta investigación en proyectos investigativos y/o productivos en la UCI.

Citas bibliográficas

1. **Blanco, Visitación García Jiménez Alvarado Izquierdo y Amelia Jiménez.** *La predicción del rendimiento académico: regresión lineal versus regresión logística .* [Disponible] [Citado: febrero 16, 2009.] http://www.psicothema.com/pdf/558.pdf.

2. **Ciria, Miguel Ángel Broc Cavero y Carmen Gil.** Predicción del rendimiento académico en alumnos de ESO y Bachillerato mediante el Inventario Clínico para Adolescentes de Millon (escala MACI). [Disponible] [Citado: marzo 23, 2009.] http://www.doredin..

3. **Samper, Julián De Zubiría.** *Cinco mitos sobre la inteligencia y el talento.* [Disponible] 2006. [Citado: febrero 13, 2009.] http://www.institutomerani.edu.co/publicaciones/articulos/2009/Mitos%20sobre%20la%20Inteligenica%20y%20el%20talento%20De%20Zubir%C3%ADa.pdf.

4. **Rodríguez, Carlos Alberto Román y Yenima Hernández.** *Variables psicosociales y su relación con el desempeño académico de estudiantes de primer año de la Escuela Latinoamericana de Medicina.* [Disponible] [Citado: marzo 25, 2009.] http://www.rieoei.org/1085.htm .

5. **Cavero, Sonia Damiani.** *"Resultados diferenciales de la prueba diagnóstica sobre gráficos según procedencia de educación media superior.* [Disponible] [Citado: marzo 10, 2009.] http://bvs.sld.cu/revistas/ems/vol16_3_02/ems05302.htm.

6. **Pawlak, Z.** *Rough Sets. International journal of Computer and Information Sciences.* 1982.

7. **Zadeh, Lotfi A.** *Texto del discurso pronunciado para la recepción del Doctorado Honaris Causa por la Facultad de ciencias de la Universidad de Ovideo.* España : s.n., 1995.

8. **Parsons, S.** *"Current approaches to handling imperfect information indata and knowledge bases", en IEEE Trans. On knowledge and data enginnering.* 2006.

.

9. **Carrillo, Fernando Virseda Benito y Javier Román.** *Minería de datos y aplicaciones.* [Disponible] Universidad Carlos III. [Citado: marzo 13, 2009.] http://www.it.uc3m.es/jvillena/irc/practicas/06-07/22.pdf .

10. **Zhang S., Zhang C. , Yang Q.** *Data preparation for data mining. Applied Artificial Intelligence.* 2005.

11. **Langley.** *Selection of Relevant Features in Machine Learning. Procs. Of the AAAI Fal Symposiumon Relevance.* New Orleans, LA : s.n., 2004.

12. **Bahamonde, Antonio.** [Disponible] [Citado: marzo 26, 2009.] www.aic.uniovi.es.

13. **Sopena.** Enciclopedia. [book auth.] Provenza 95. Barcelona : Editorial Ramon Sopena, 2006.

14. **Rosas, Fransisco días.** *la predicción del rendimiento académico en la universidad: un ejemplo de aplicaciónde la regresión múltip.* [Disponible] [Citado: abril 10, 2009.] http://e-spacio.uned.es/fez/eserv.php?pid=bibliuned:20476&dsID=prediccion_rendimiento.pdf.

15. **Vidal, Tomás Arredondo.** *Árboles de Decisión* [Disponible] marzo 26, 2008 [Citado: marzo 16, 2009.] http://profesores.elo.utfsm.cl/~tarredondo/info/soft-omp/Arboles%20de%20Decision.pdf

16. **Amézquita, Marcela Quintero Edgar.** *Herramienta "Arboles de Decisión": Alternativas de Uso de la Tierra para los Llanos Orientales de Colombia.* [Disponible] 2000. [Citado: marzo 16, 2009.] http://ciat-library.ciat.cgiar.org/documentos_electronicos_ciat/articulos_ciat/Manual_Arboles.pdf.

17. **Pawlak, Z.** *Vagueness and uncertainty: a rough set perspective en Computational Intelligence.* 1995.

18. **Pawlak, Z.** *Rough sets.* 1995.

19. **D.A SAVIC, J.W DAVIDSON, DAVIS.** *Data Mining and Knowledge discovery for the water industry", Water Industry Systems, modelling and optimisation applications.* s.l. : Dragan A. Savic and Godfrey A. Walter, 2001.

20. **Pérez, Rafael Bello.** *La teoría de conjuntos aproximados en el contexto de la inteligancia artificial.* Universidad de las Ciencias Informáticas. Cuba. 2008.

21. **KONOW, I. y PÉREZ, G.** *Método Delphi.* [Disponible] 1998. [Citado: marzo 23, 2009.] http://geocities.com/Pentagon/Quarters/7578/pros01-03.html

22. **RUIZ OLABUÉNAGA, J. e ISPIZUA, M. A.** La técnica Delphi. *La descodificación de la vida cotidiana. Métodos de investigación cualitativa* 1989.

23. **PARISCA, S**. *Método Delphi. Gestión tecnológica y competitividad Estrategia y filosofía para alcanzar la calidad total y el éxito en la gestión impresional*. La Habana 1997.

24. **Bravo Estévez, María de Lourdes**; **Arrieta Gallastegui, José Joaquín.** *El método Delphi como estrategia didáctica para la enseñansa universitaria***,** Universidad de Cienfuegos. Cuba. Universidad de Oviedo. España. [Disponible] 2006. [Citado: marzo 23, 2009.] http://www.rieoei.org/deloslectores/804Bravo.PDF

25. **Murillo, Bravo***. Integración del STATGRAPHICS en un programa Seis Sigma*. [Disponible] 2004. [Citado: abril 6, 2009.] http://www.STATGRAPHICS.net/SeisSigma.pdf

26. **Morate, Diego García.** *WEKA.* España : [Disponible] 2005 [Citado: enero 21, 2009.] http://metaemotion.com/diego.garcia.morate/download/WEKA.pdf.

27. **Kirkby, Richard; Frank, Eibe; Reutemann, Peter.** *WEKA Explorer User Guide for Version 3-5-8.* [Disponible] 2008 [Citado: enero 22, 2009.] http://www.cs.waikato.ac.nz/~ml/WEKA/

28. **Hernández, Enrique.** *El Lenguaje Unificado de Modelado (UML).*

29. **Van Rossum, Guido**; Ruiz, José María. *PYTHON 3000*. [Disponible] 2008 [Citado: abril 20, 2009.] http://www.linux-magazine.es/issue/42/055-058PythonLM42.pdf

30. **García Corzo, Pablo M.** *Ciencia Abierta y Software Libre* en el aprendizaje [Disponible] 19 de noviembre de 2008 [Citado: abril 20, 2009.] http://www.ucm.es/info/aulasun/archivos/presenta.pdf

31. **Fernández Caballero, Antonio; Gracía, María; Arjona, Manzano.** *Una Perspectiva de la Inteligencia Artificial en su 50 Aniversario.* [Disponible] Julio 14, 2006 [Citado: febrero 19, 2009.] http://www.info-ab.uclm.es/personal/AntonioFdez/download/papers/conference/cmpi2006-volumeII.pdf#page=219

32. **Caballero, Yailé; Bello, Rafael; Alvarez, Delia; Garcia, Maria M.; Baltá, Analay.** *Un nuevo algoritmo de selección de rasgos basado en la Teoría de los Conjuntos Aproximados.* [Disponible] 2007 [Citado: febrero 17, 2009.] http://redalyc.uaemex.mx/redalyc/pdf/430/43004111.pdf

Bibliografía consultada

Amézquita, Marcela Quintero Edgar. *Herramienta "Arboles de Decisión": Alternativas de Uso de la Tierra para los Llanos Orientales de Colombia.* [Disponible] 2000. [Citado: marzo 16, 2009.] http://ciat-library.ciat.cgiar.org/documentos_electronicos_ciat/articulos_ciat/Manual_Arboles.pdf.

Bahamonde, Antonio. [Disponible] [Citado: marzo 26, 2009.] www.aic.uniovi.es.

Blanco, Visitación García Jiménez Alvarado Izquierdo y Amelia Jiménez. *La predicción del rendimiento académico: regresión lineal versus regresión logística .* [Disponible] [Citado: febrero 16, 2009.] http://www.psicothema.com/pdf/558.pdf.

Booch, G. J, Jcobson. *"El lenjuaje Unificado de modelado".* 1998.

Bravo Estévez, María de Lourdes; **Arrieta Gallastegui, José Joaquín.** *El método Delphi como estrategia didáctica para la enseñansa universitaria,* Universidad de Cienfuegos. Cuba. Universidad de Oviedo. España. [Disponible] 2006. [Citado: marzo 23, 2009.] http://www.rieoei.org/deloslectores/804Bravo.PDF

Caballero, Yailé; Bello, Rafael; Alvarez, Delia; Garcia, Maria M.; Baltá, Analay. *Un nuevo algoritmo de selección de rasgos basado en la Teoría de los Conjuntos Aproximados.* [Disponible] 2007 [Citado: febrero 17, 2009.] http://redalyc.uaemex.mx/redalyc/pdf/430/43004111.pdf

Carrillo, Fernando Virseda Benito y Javier Román. *Minería de datos y aplicaciones.* [Disponible] Universidad Carlos III. [Citado: marzo 13, 2009.] http://www.it.uc3m.es/jvillena/irc/practicas/06-07/22.pdf .

Cavero, Sonia Damiani. *"Resultados diferenciales de la prueba diagnóstica sobre gráficos según procedencia de educación media superior.* [Disponible] [Citado: marzo 10, 2009.] http://bvs.sld.cu/revistas/ems/vol16_3_02/ems05302.htm.

Ciria, Miguel Ángel Broc Cavero y Carmen Gil. Predicción del rendimiento académico en alumnos de ESO y Bachillerato mediante el Inventario Clínico para

Adolescentes de Millon (escala MACI). [Disponible] [Citado: marzo 23, 2009.] http://www.doredin.

D.A SAVIC, J.W DAVIDSON, DAVIS. *Data Mining and Knowledge discovery for the water industry", Water Industry Systems, modelling and optimisation applications.* s.l. : Dragan A. Savic and Godfrey A. Walter, 2001.

Fernández Caballero, Antonio; Gracía, María; Arjona, Manzano. *Una Perspectiva de la Inteligencia Artificial en su 50 Aniversario.* [Disponible] Julio 14, 2006 [Citado: febrero 19, 2009.] http://www.info-ab.uclm.es/personal/AntonioFdez/download/papers/conference/cmpi2006-volumeII.pdf#page=219

García Corzo, Pablo M. *Ciencia Abierta y Software Libre* en el aprendizaje [Disponible] 19 de noviembre de 2008 [Citado: abril 20, 2009.] http://www.ucm.es/info/aulasun/archivos/presenta.pdf

Hernández, Enrique. *El Lenguaje Unificado de Modelado (UML).*

KONOW, I. y PÉREZ, G. *Método Delphi .* [Disponible] 1998. [Citado: marzo 23, 2009.] http://geocities.com/Pentagon/Quarters/7578/pros01-03.html

Kirkby, Richard; Frank, Eibe; Reutemann, Peter. *WEKA Explorer User Guide for Version 3-5-8.* [Disponible] 2008 [Citado: enero 22, 2009.] http://www.cs.waikato.ac.nz/~ml/WEKA/

Langley. *Selection of Relevant Features in Machine Learning. Procs. Of the AAAI Fal Symposiumon Relevance.* New Orleans, LA : s.n., 2004.

MARAKAS, G. *Decision Support Systems in the 21st Century.* New York, E.U.A. : Prentice-Hall, 1998.

Morate, Diego García. *WEKA.* España : [Disponible] 2005 [Citado: enero 21, 2009.] http://metaemotion.com/diego.garcia.morate/download/WEKA.pdf.

Murillo, Bravo. *Integración del STATGRAPHICS en un programa Seis Sigma*. [Disponible] 2004. [Citado: abril 6, 2009.] http://www.STATGRAPHICS.net/SeisSigma.pdf

Orlowska, E. (ed.). *Incomplete Information: Rough set analysis.* Physica- Verlag : s.n., 1998.

Parsons, S. *"Current approaches to handling imperfect information indata and knowledge bases", en IEEE Trans. On knowledge and data enginnering.* 2006.

PARISCA, S. *Método Delphi . Gestión tecnológica y competitividad Estrategia y filosofía para alcanzar la calidad total y el éxito en la gestión impresional*. La Habana 1997.

Pawlak, Z. *Rough Sets. International journal of Computer and Information Sciences.* 1982.

Pawlak, Z. *Vagueness and uncertainty: a rough set perspective en Computational Intelligence.* 1995.

Pawlak, Z. *Rough sets.* 1995.

PAWLAK, Z. *Rough Sets, International Journal of Information & Computer Sciences.* 1982. vol. 11, p. 341-356.

PAWLAK, Z. *Rough Sets – Theoretical Aspects of Reasoning about Data .* Boston, London : Kluwer Academic Publishers, 2001.

PAWLAK, Z. and SLOWINSKI, R. *Decision Analysis using Rough Sets. ICS Research Report no 21.* Warsaw, Poland : Institute of Computer Science, Warsaw University of Technology, 1993.

Pérez, Rafael Bello. *La teoría de conjuntos aproximados en el contexto de la inteligancia artificial.* Universidad de las Ciencias Informáticas. Cuba. 2008.

Rosas, Fransisco días. *la predicción del rendimiento académico en la universidad: un ejemplo de aplicaciónde la regresión múltip.* [Disponible] [Citado: abril 10, 2009.] http://e-spacio.uned.es/fez/eserv.php?pid=bibliuned:20476&dsID=prediccion_rendimiento.pdf.

Rodríguez, Carlos Alberto Román y Yenima Hernández. *Variables psicosociales y su relación con el desempeño académico de estudiantes de primer año de la Escuela Latinoamericana de Medicina.* [Disponible] [Citado: marzo 25, 2009.] http://www.rieoei.org/1085.htm .

RUIZ OLABUÉNAGA, J. e ISPIZUA, M. A. La técnica Delphi . *La descodificación de la vida cotidiana. Métodos de investigación cualitativa* 1989.

Samper, Julián De Zubiría. *Cinco mitos sobre la inteligencia y el talento.* [Disponible] 2006. [Citado: febrero 13, 2009.] http://www.institutomerani.edu.co/publicaciones/articulos/2009/Mitos%20sobre%20la%20Inteligenica%20y%20el%20talento%20De%20Zubir%C3%ADa.pdf.

SLOWINSKI, R. and STEFANOWSKI, J. *Rough Classification in Incomplete Information Systems, Mathematical and Computer Modelling.* 1989. vol. 12, no 10/11, p. 1347-1357.

Sopena. Enciclopedia. [book auth.] Provenza 95. Barcelona : Editorial Ramon Sopena, 2006.

Van Rossum, Guido; Ruiz, José María. *PYTHON 3000*. [Disponible] 2008 [Citado: abril 20, 2009.] http://www.linux-magazine.es/issue/42/055-058PythonLM42.pdf

Vidal, Tomás Arredondo. *Árboles de Decisión* [Disponible] marzo 26, 2008 [Citado: marzo 16, 2009.] http://profesores.elo.utfsm.cl/~tarredondo/info/soft-omp/Arboles%20de%20Decision.pdf

Zadeh, Lotfi A. *Texto del discurso pronunciado para la recepción del Doctorado Honaris Causa por la Facultad de ciencias de la Universidad de Ovideo.* España : s.n., 1995.

Zhang S., Zhang C. , Yang Q. *Data preparation for data mining. Applied Artificial Intelligence.* 2005.