UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 3

Extracción de correlaciones entre el test vocacional CHASIDE y la carrera de Ingeniería en Ciencias Informáticas.

Tesis presentada para optar por el título Ingeniero en Ciencias Informáticas

Autor: Samuel Ojeda Pereira

Tutores: Ing. Guillermo Manuel Negrín Ortiz

MSc. Julio C. Diaz Vera

La Habana, 24 de junio de 2017

DATOS DE CONTACTO

MSc. Julio Cesar Diaz Vera

Universidad Martha Abreu, Villa Clara, Cuba.

Correo electrónico: jcdiaz@uci.cu

Ing. Guillermo Manuel Negrín Ortiz

Universidad de las Ciencias Informáticas, La Habana, Cuba.

Correo electrónico: gmnegrin@uci.cu

"*...Al mundo no lo va a podrirlo la gente mala, sino la que se dice BUENA!!!! pero no hace nada....nadita nada!!!...* "

Buena Fe

DECLARACIÓN DE AUTORÍA

Declaro ser autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste se firma la presente a los ___ días del mes de _____ del año 20___ .

Samuel Ojeda Pereira

MSc. Julio Cesar Diaz Vera

Ing. Guillermo Manuel Negrín Ortiz

AGRADECIMIENTOS

Agradecer por los logros de estos 5 años:

En primer lugar a mis tutores Julio Cesár y el Guille los cuáles son como un padre y un hermano para mí.

A mi tío José el cuál, ha sido durante estos 5 años, guía para todos mis pasos.

A mi compañera de vida Danet la que ha estado apoyándome en estos años y en toda mi labor en eventos.

A alguien que es como una hermanita de la universidad y que quiero mucho: Leidy Rosa.

A mi segundo papá Jorge quien, con su resabio ha sabido educarme ante la vida.

A mi prima Maidy y el Cabe que siempre confiaron en mí.

Mi tía Maira que siempre me ha dado la fuerza espiritual para salir adelante.

A mi hermano Fabian el cuál, a pesar de la distancia, me ayuda siempre.

A Selita la negrita más bullera de todo el 57, que es una amiga muy especial.

A alguien que extraño mucho porque fue una buena amiga y compañera de cuarto: ARY.

A Patry (la lagartijita) y su flaco Esteban quienes han sido amigos incondicionales.

A mis compañeros de aula que me han ayudado a ser mejor, a lo largo de los años

A la brigada FI21 por su ayuda y creer en mí como alumno ayudante.

A los profes departamento de ISW por su ayuda incondicional y tenerme todos los días a su lado (no menciono nombres para que no hayan celos).

A la profe Maybel la cuál me ha apoyado en mi investigación.

A todos los profes que han contribuido en mi carrera (Eliober, Hugo, Michel Álvarez Cancio, Michel Sariol, Dina, Pedro Arango,Yorgüy, Yadira Beatriz...) en especial a Mónica y Dariela que siempre han estado pendiente de mí.

A los profesores del tribunal los cuales estuvieron a su cargo la revisión de este documento. En especial a las profesoras Ailía y Olga.

A mi suegra Yamilet y Ana María que han sido como una madre y una tía para mí.

A Humbelina y al profesor Fabra que con mucha disposición me han ayudado cada una de las veces que fui a pedir su socorro.

Samuel Ojeda Pereira

DEDICATORIA

A la memoria de dos grandes educadoras que forman parte de mis logros:

Celia García Bellos(Mi ABUELITA)

y

Ángela (tía Laly)

Esta investigación también está dedicada a mis padres:

Zunilda Ibis Pereira (Mi gordita)

Samuel Ojeda García(Mi viejo)

Samuel Ojeda Pereira

RESUMEN

La sociedad cubana está comprometida con el éxito personal de cada uno de sus ciudadanos y el sistema de educación superior del país es un reflejo de ese compromiso. A pesar de ello los niveles de fracaso escolar no se corresponden con los deseados ya que sobrepasan los esperados. Una causa para esta problemática puede estar asociada a que los estudiantes no matriculan carreras correlacionadas favorablemente con sus habilidades, competencias y preferencias. En este trabajo se propone relacionar estas habilidades, competencias y preferencias presentes en el test de orientación vocacional CHASIDE con la carrera de Ingeniería en Ciencias Informáticas. El objetivo de este trabajo es obtener un modelo que permita hacer un análisis de la correlación que existe entre los elementos antes mencionados. El proceso de desarrollo de software orientado al análisis de datos es el de software de predominio de cómputo que usa como parte de su metodología la extracción de conocimiento de una base de datos, representado en forma de reglas de asociación que permiten obtener un modelo de análisis en forma de correlaciones. Se usó un dataset con 300 tuplas para obtener el modelo que permite establecer las correlaciones entre las habilidades, competencias y preferencias y la carrera de Ingeniería en Ciencias Informáticas.

Palabras claves: Competencias, correlación, habilidades, nivel de fracaso .

Índice General

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La orientación vocacional comienza a considerarse un elemento importante en el desarrollo de capacidades creadas en los trabajadores a partir del año 1908 con la creación, en Boston, Estados Unidos, del Primer Buró de Orientación Vocacional a cargo de Frank Parsons. En aquel momento se introduce el término "Vocational Guidance" para agrupar un conjunto de elementos que permitían escoger qué profesión podría resultar más adecuada para cada persona en particular[1].

A medida que la orientación vocacional ha evolucionado se han utilizado técnicas que permiten establecer determinadas correspondencias entre las aptitudes naturales del ser humano y competencias necesarias para el correcto desempeño de la profesión. Las técnicas que han acaparado la mayor atención en este tipo de tareas están asociadas a la aplicación de test que siguen las teorías factorialistas [2].

Cuba ha dedicado esfuerzos en el área de la orientación vocacional. Este es un proceso que se lleva a cabo en el sector de la educación, y que sirve de ayuda para que los estudiantes de enseñanza media y media superior puedan seleccionar una profesión. Sin embargo, a pesar de ello, los niveles de fracaso escolar en las universidades cubanas todavía alcanzan cuotas superiores a las esperadas[3]. La insuficiente orientación vocacional es uno de los factores que provoca que los estudiantes seleccionen carreras para las que no tienen las aptitudes necesarias o que coinciden con sus intereses básicos. Esto provoca limitar la capacidad de los estudiantes para enfrentar las tareas, ya que no están motivados y les dificulta que alcancen resultados positivos.

Se han desarrollado varios instrumentos para determinar las habilidades, competencias y preferencias de los estudiantes y establecer la relación entre estas y las áreas del conocimiento. Especialmente interesante en este sentido es el test CHASIDE [3]. Este test sirve de apoyo para identificar los intereses y aptitudes de los estudiantes enfocado a una orientación vocacional y profesional, el test consta de 98 preguntas con actividades que pueden ser del interés del estudiante, y así poder tener una guía de los intereses que el estudiante podría tener (ver anexo A). Sin embargo, estos resultados no han sido correlacionados, de manera experimental, con el éxito en el estudio de la carrera de Ingeniería

en Ciencias Informáticas. Lo que constituye una limitación importante si se pretenden desarrollar sistemas inteligentes y/o de recomendación que contribuyan a la orientación de los estudiantes con vistas a decidir la carrera más favorable para ellos.

Lo antes expresado conduce a plantear:

-Problema a resolver: ¿Cómo correlacionar las habilidades, competencias y preferencias determinadas mediante el test de orientación vocacional CHASIDE con la carrera de Ingeniería en Ciencias Informáticas?

-Objeto de estudio: Proceso de desarrollo de software

-Objetivo general: Establecer un grupo de correlaciones entre las habilidades, competencias y preferencias detectadas en el test vocacional CHASIDE con la carrera de Ingeniería en Ciencias Informáticas.

Teniendo como campo de acción: Proceso de desarrollo de software de predominio de cómputo.

Objetivos específicos:

1. Establecer el marco conceptual para el desarrollo de la investigación.
2. Recopilar los resultados de la aplicación del test CHASIDE.
3. Determinar las técnicas de exploración de datos a utilizar mediante la metodología Catalyst.
4. Implementar los algoritmos asociados a las técnicas seleccionadas.
5. Validar los resultados obtenidos.

Posibles resultados: Modelo de correlación entre las habilidades, competencias y preferencias detectadas en el test vocacional CHASIDE y la carrera de Ingeniería en Ciencias Informáticas.

Capítulo 1

Fundamentación teórica

1. Fundamentación Teórica

En este capítulo se describen los elementos fundamentales asociados a la investigación, se detallan los diferentes acercamientos para determinar correlaciones y se exponen las dificultades desprendidas de cada una. Se presenta un acercamiento a la minería de datos como una alternativa viable para el cómputo de las correlaciones, al mismo tiempo se ubica a este tipo de proyecto dentro de la taxonomía de software propuesta en [4].

Por último, se presentan los elementos metodológicos fundamentales asociados al desarrollo de proyectos de minería de datos. Describiéndose las características fundamentales de la metodología escogida.

1.1. Alternativas para el cómputo de correlaciones

Las competencias, habilidades y preferencias de un ser humano están estrechamente correlacionadas, este es un factor determinante a la hora de definir las correlaciones existentes de estos elementos y el éxito en la carrera de Ingeniería en Ciencias Informáticas (desempeño académico).

La primera aproximación que debe ser mencionada está asociada al uso de modelos de regresión, ya sea de regresión gradual [5] o de regresión hacia adelante [6], que permiten obtener un modelo de correlación reducido. Sin embargo, estos métodos estadísticos estándares, consideran el impacto de cada factor de manera independiente y tienden a introducir errores de estimación o interpretación. Por ello, han sido evitados en este tipo de análisis, favoreciendo el uso de modelos más complejos[7].

En [8] se presenta una revisión exhaustiva de los métodos utilizados para operar sobre el efecto de la alta correlación de los factores a la hora de estimar la correlación con el efecto. El estudio se realiza en particular para el campo del análisis del efecto de la solución sobre la salud respiratoria de los pacientes. Las correlaciones reportadas entre los contaminantes PM_10, NO_2, SO_2 y CO medidas entre el 70 y el 85 $\%$ [9] son equivalentes a las encontradas en las preguntas del test CHASIDE.

Los Modelos Bayesianos, una de las principales alternativas revisadas por [8], superan significativamente a las aproximaciones más tradicionales. Los mismos, no son capaces de solucionar

el problema de la varianza entre los elementos del modelo debido a la necesidad de que el investigador aporte la información asociada a la probabilidad condicional de los factores que puede no estar disponible o ser desconocida. En segundo, lugar la gran cantidad de información necesaria para definir a priori la distribución exacta de probabilidades la convierte en un enfoque poco realista.

El particionamiento recursivo, es otra de las alternativas que ha sido utilizada para relacionar elementos. En este caso se construye un modelo de regresión basado en árboles que permite detectar interacciones complejas y múltiples al mismo tiempo, que se beneficia de una interfaz visualmente intuitiva. La desventaja fundamental está asociada a su imposibilidad de gestionar variables, que tienen efectos muy similares y en estos casos es común que una de la dos no sea considerada dentro del árbol[10].

Otra de las variantes importantes es el análisis con agrupamientos (cluster analysis) que ha obtenido resultados favorables [11] pero presenta una gran dificultad a la hora de seleccionar los grupos, tanto en la cantidad de grupos a crear, como en los elementos a tener en cuenta para su construcción. Existe mucha subjetividad en el momento de etiquetar los grupos y al mismo tiempo, es extremadamente difícil, resumir los grupos a partir de una única etiqueta.

Una variante que ha tenido resultados muy favorables, y que constituye la motivación técnica principal de este trabajo, fue presentada en [12] donde se listan un grupo de ventajas del uso de las reglas de asociación para el análisis de correlaciones.

- Las reglas de asociación tienen mayor interpretación en comparación con el resto de las opciones.
- Las reglas de asociación son capaces de resumir el impacto de la combinación de múltiples factores sin recurrir a jerarquías.
- Se limitan las creencias no verificadas por parte del investigador.

Con vistas a desarrollar una herramienta informática que facilite el cómputo de correlaciones haciendo uso de técnicas propias del minado de reglas de asociación es necesario caracterizar la misma dentro del ámbito del desarrollo de software.

1.2. Desarrollo de software para el minado de reglas de asociación

El problema de la construcción de software es tratado a nivel académico en la disciplina Ingeniería de Software [13]. Sin embargo, los principios, prácticas y postulados que conforman el cuerpo de conocimiento de esta disciplina no pueden ser adoptados simplemente porque alguien los ha enunciado antes. La adopción de los postulados de la Ingeniería de Software requiere de contexto y evidencias.

En los últimos años, ha ganado protagonismo, una línea de trabajo enfocada a la realización de investigaciones empíricas en el campo de la Ingeniería de Software [14]. Las mismas están orientadas a proporcionar el contexto necesario para la adopción de sus postulados. Este elemento ha impulsado el desarrollo de trabajos que pretenden establecer categorías para los tipos de software. La aspiración fundamental es lograr enmarcar el contexto en el que son aplicadas las evidencias empíricas.
Uno de los resultados indispensables en esta área es el desarrollo de taxonomías de tipos de software que permitan establecer un vocabulario estructurado al mismo tiempo que define una teoría de acuerdo a la que se organizan las etiquetas del vocabulario. A juicio del investigador es especialmente importante, en este sentido, la taxonomía de tipos de software presentada en [15].

1.2.1. Taxonomía de tipos de software

La taxonomía propuesta presenta cuatro niveles. En el primer nivel se agrupan cuatro categorías diferentes que se etiquetan de la A a la D y que permiten agrupar a los software teniendo en cuenta el predominio de la gestión de datos o del cómputo[16]. Las categorías definidas son las siguientes:

- A: Sistemas con predominio de los datos: en esta categoría se agrupan los sistemas de propósito general desde los orientados a consumo hasta algunos con perfiles más específicos dentro de un negocio.
- B: Software de Sistemas: el centro de esta categoría se centra en el nivel de abstracción desde el bajo nivel asociado a los sistemas operativos hasta el más alto asociado a ejemplo entornos integrados de desarrollo.
- C: Sistemas con predominio del control: estos se centran en cuánto es necesario controlar el sistema desde pequeños programas embebidos hasta grandes sistemas que gestionan y controlan operaciones complejas como los SCADA.
- D: Sistemas con predominio del cómputo: el eje central de esta línea está en cuán conceptual es el problema desde aspectos prácticos hasta áreas totalmente teóricas como la inteligencia artificial.

El problema planteado en este trabajo clasifica dentro del item D, un sistema donde predomina el cómputo. Por ello esta rama de la taxonomía es la que se describe a continuación.

1.2.1.1. Sistemas con predominio del cómputo

Según la taxonomía planteada, en el segundo nivel se utilizan dos o tres letras minúsculas que representan un código nemotécnico para la categoría. El tercer nivel se etiqueta con un número

mientras que, el cuarto, es etiquetado con una letra minúscula[16]. Es necesario puntualizar que para el software de predominio de cómputo la taxonomía tiene tres niveles como puede apreciarse en la figura 1.1.

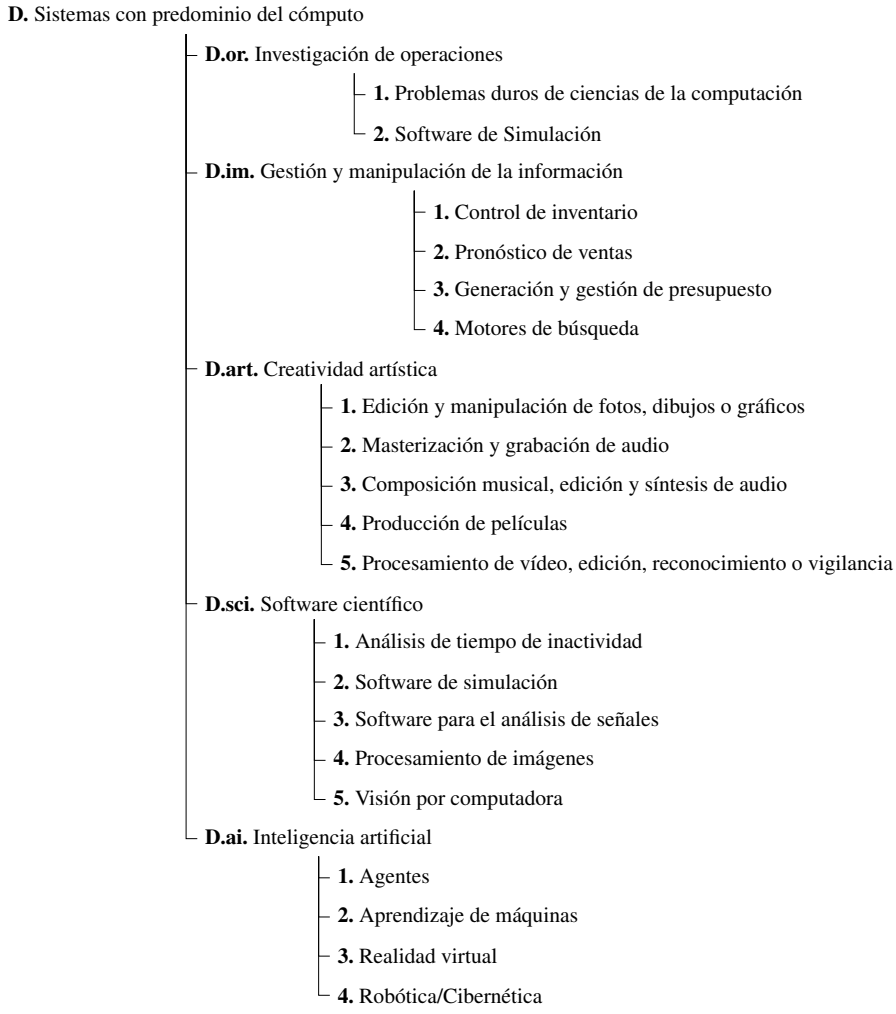


Figura 1.1: *Sistemas con predominio del cómputo*

Elaboración propia

El cómputo de reglas de asociación es una de las áreas más estudiadas y aplicadas de la minería de datos, una rama multidisciplinaria de las Ciencias de la Computación que a juicio de varios investigadores tiene una estrecha relación con el aprendizaje de máquinas [17]. Para la inteligencia artificial, entre el aprendizaje de máquinas y la minería de datos, es posible establecer una relación jerárquica de uso en la que:

- La minería de datos se encarga del descubrimiento de patrones ocultos en los datos, lo que permite explicar algún fenómeno. Permite intuir lo que realmente está pasando en un conjunto

de datos.

- El aprendizaje de máquinas utiliza las técnicas de minería de datos y otros algoritmos de aprendizaje para construir modelos acerca de lo que está pasando en un conjunto de datos de forma que se pueda utilizar el modelo para predecir el futuro.
- La inteligencia artificial utiliza los modelos creados con aprendizaje de máquinas y otras formas de razonamiento para permitir un comportamiento inteligente en las computadoras.

A partir de lo antes planteado es posible establecer la clasificación del software para el cómputo de reglas de asociación como **D.ai.1** Software con predominio del cómputo - inteligencia artificial - aprendizaje de máquinas.

De manera general el uso de la taxonomía permitirá comparar las soluciones que han funcionado y las que no lo han hecho para un tipo particular de desarrollo de software.

La clasificación de la aplicación a desarrollar, facilita la tarea de relacionar varias herramientas, técnicas y métodos, con lo que se facilitará que, los desarrolladores, puedan tomar mejores decisiones acerca de:

- Metodologías de desarrollo
- Patrones de diseño y componentes de software
- Algoritmos
- Entornos integrados de desarrollo
- Formatos de intercambio de datos
- Arquitecturas y estilos arquitectónicos
- Técnicas de prueba

1.3. Metodologías para el desarrollo de proyectos de minería de datos

Las metodologías para el desarrollo de proyectos de minería de datos alcanzaron su punto cumbre en la década del 2000. Donde apareció CRISP-DM [18] la metodología más utilizada para el desarrollo de proyectos de minería de datos y que se considera un estándar de facto para la industria. Sin embargo, CRISP-DM define qué hacer, pero no cómo hacerlo y no incluye actividades asociadas a la gestión de los proyectos lo que provocó que aparecieran varias metodologías alternativas dentro de las que se pueden destacar SEMMA, propuesta por el instituto SAS, Catalyst (también conocida como P3TQ) [19] y la propia evolución de CRISP-DM denominada CRISP-DM 2.0.

En [20] se presenta una revisión exhaustiva de las metodologías de minería de datos más comúnmente utilizadas hasta el año 2010. Escoger entre una de ellas implica la existencia de un marco comparativo que permita seleccionar la que mejor se adapte al proyecto que se desea realizar.

Para ello es necesario definir los elementos fundamentales asociados al proceso de descubrimiento de conocimiento en bases de datos (KDD por sus siglas en inglés) propuesto por primera vez en [21].

1.3.0.2. Proceso de descubrimiento de conocimiento en bases de datos

KDD es definido como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, entendibles en los datos. Es un proceso iterativo, ya que algunas de las salidas de las actividades en las fases puede retroceder a actividades anteriores, e interactivo ya que el usuario puede participar en la preparación de los datos y en la validación de los modelos obtenidos[22].

El proceso originalmente presenta nueve fases, pero puede resumirse en cuatro fases para una mejor comprensión sin pérdida de generalidad, como se muestra en la figura 1.2.

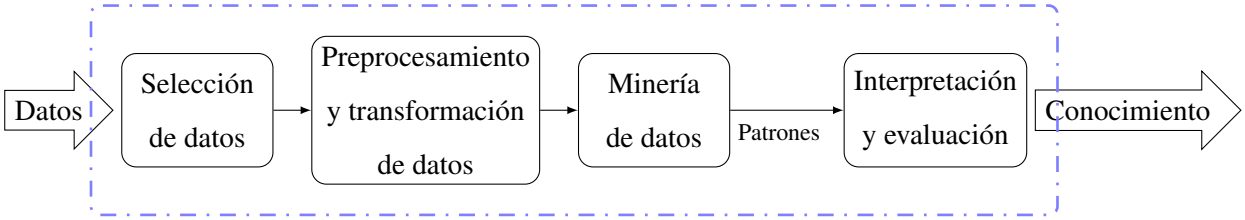


Figura 1.2: *Resumen del proceso KDD*

Elaboración propia

Cada una de las fases presentes en la figura 1.2 es responsable de las siguientes actividades[23]:

- Selección de datos: se eligen las fuentes de datos que serán utilizadas, se integran las fuentes y se determinan los atributos que formarán parte del conjunto de datos al que se aplicarán las técnicas de minería de datos. En muchos casos es recomendable la construcción de un almacén de datos.
- Preprocesamiento y transformación de los datos: se realizan tareas asociadas a la limpieza de los datos, eliminación de ruido o datos anómalos y el tratamiento a los datos faltantes. Al mismo tiempo se ejecutan tareas para la reducción de la dimensionalidad es posible transformar los datos calculando nuevos atributos o bien redefiniendo los existentes con otro formato.
- Minería de datos: se determina la tarea de minería con la que abordará el estudio, se selecciona el algoritmo de minado y se aplica el mismo a los datos seleccionados.

- Interpretación y evaluación: se aplican técnicas de evaluación para la validar lo patrones encontrados y finalmente son interpretados y traducidos a un formato amigable para el usuario.

Si bien el modelado en KDD puede resumirse en cinco fases principales mencionadas anteriormente, en [23] y [22] se definen nueve etapas para llevar a cabo todo el proceso, definidas a continuación:

1. Comprensión del dominio de aplicación: en esta primera etapa, se debería recolectar todo el conocimiento disponible y relevante sobre el dominio de aplicación e identificar los objetivos del proceso KDD desde el punto de vista del usuario.
2. Creación del conjunto de datos: esta etapa consiste en la elección de las fuentes de datos que se utilizarán, la integración de las mismas y la selección de las observaciones/atributos que conformarán la vista minable (datos que se van a procesar). Aunque no es estrictamente necesario, en este paso podría requerirse la construcción de un almacén de datos.
3. Limpieza y pre-procesamiento de los datos: en esta fase se deberían llevar a cabo tareas como limpieza de ruido o datos anómalos (outliers) y tratamiento de datos faltantes (missing values).
4. Reducción y proyección de los datos: en este paso se detectan características útiles de representación de los datos dependiendo del objetivo de la tarea de minería (descripción o predicción). Se incluye la utilización de técnicas de reducción de la dimensionalidad y métodos de transformación de los datos para reducir la cantidad de variables en discusión o para encontrar representaciones invariantes de los datos. En esta etapa es frecuente la transformación de los datos, calculando nuevos atributos o bien re-definiendo los existentes con otro formato.
5. Determinar la tarea de minería de datos: en esta fase, se deberá determinar la tarea de minería con la que se abordará el estudio (como agrupamiento, regresión, clasificación, o asociación) teniendo en cuenta los objetivos definidos en la etapa 1.
6. Determinar el algoritmo de minería. De acuerdo a la tarea de minería establecida en el punto anterior, en esta etapa se define el algoritmo (o algoritmos) que se aplicarán para la búsqueda de patrones sobre los datos. Incluye la determinación de qué modelos y parámetros son los más adecuados según la naturaleza del problema y de los datos disponibles.
7. Minería de datos (MD): etapa en la que se aplican los algoritmos y técnicas seleccionadas al conjunto de datos en búsqueda de los patrones de interés.
8. Interpretación. Comprende la interpretación de los patrones encontrados, visualizando y traduciendo los mismos en términos comprensibles por el usuario.
9. Utilización del nuevo conocimiento: en esta fase se implementa el conocimiento descubierto, apoyando con el mismo la toma de decisiones o bien reportándolo a las partes interesadas.

Incluye la verificación y resolución de potenciales conflictos con conocimiento descubierto previamente.

Es posible apreciar que KDD define las etapas de desarrollo de un proyecto de minería de datos, pero no aporta nada en cuestión de las actividades concretas que deben realizarse. Esta responsabilidad queda en manos del equipo de desarrollo.

1.3.0.3. Marco comparativo

Previamente se comentaba la necesidad de contar con un marco comparativo apropiado para la selección de la metodología de desarrollo de proyectos de minería de datos. En este apartado se presentan los elementos fundamentales a tener en cuenta para la selección, partiendo del proyecto concreto que se pretende desarrollar. Existen cuatro elementos fundamentales para comparar las metodologías de desarrollo[24].

- Nivel de detalle de las actividades donde es necesario, evaluar el grado de profundidad con el que se describe cada actividad. En una metodología es necesario que por fase se propongan todas las actividades específicas que la componen y una guía de cómo ejecutar cada actividad.
- Escenarios de aplicación. La metodología debe detallar cómo se adapta a cada uno de los diferentes escenarios que pueden constituir punto de partida del proceso de desarrollo.
- Actividades específicas que componen las fases. En este punto se evalúan las actividades principales que deberían estar presentes en cada fase.
- Actividades de gestión del proyecto. En esta sección se evalúan la incorporación de actividades que tienen por objetivo aumentar las probabilidades de que el proyecto finalice con éxito en el tiempo estimado y con el presupuesto aprobado.

El proyecto que se presenta en este trabajo se desarrolla por una persona, en un marco mayoritariamente académico y sin especificaciones de presupuesto. Estos elementos permiten descartar las actividades de gestión del marco comparativo ya que el impacto de las mismas sobre el resultado final es prácticamente nulo. En la figura 1.3 se presentan los elementos que conforman el marco comparativo y en las siguientes secciones se desarrolla cada uno de manera independiente.

1.3.0.4. Nivel de detalle de las actividades de cada fase

Las metodologías están compuestas por actividades agrupadas en paquetes de mayor abstracción denominados fases. Las actividades específicas de cada fase establecen la secuencia de trabajo a

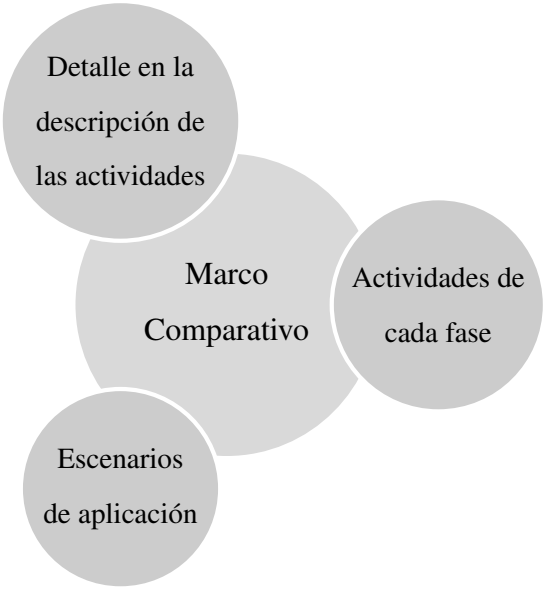


Figura 1.3: *Elementos que conforman el marco comparativo*

realizar. Una metodología completa no solo describe las actividades, sino que específica la forma en que estas deben ejecutarse. El resultado concreto de la ejecución de una actividad es un entregable. En el marco comparativo se evalúa el nivel de detalle con que la metodología define las actividades de cada fase a partir de las siguientes características[22].

- Característica 1.1: ¿Se definen actividades específicas para cada fase del proceso?

  Una metodología de minería de datos debería proponer un conjunto de actividades específicas que detallan el trabajo a realizar en cada fase.

  Resultado esperado: En cada fase del proceso se definen actividades específicas de menor nivel.
- Característica 1.2: ¿Se presentan de manera explícita los pasos a seguir para desarrollar cada actividad?

  Una metodología completa debería describir los pasos necesarios para ejecutar cada una de las actividades específicas. Esto disminuye la subjetividad y da mayor apoyo al usuario.

  Resultado esperado: Se indica paso a paso cómo ejecutar cada una de las actividades que se propone.
- Característica 1.3: ¿Se definen las entradas de cada actividad?

  La definición de las entradas de una actividad le permitirá al equipo de trabajo saber cuándo está en condiciones de comenzar la misma y qué elementos requiere para ello (requisitos). Una

entrada puede ser el resultado de una actividad anterior.

Resultado esperado: Se indican las entradas necesarias para cada actividad.

- Característica 1.4: ¿Se definen las salidas de cada actividad?

  La salida de una actividad generalmente se materializa mediante uno o más entregables, los cuales representan los resultados obtenidos luego de su ejecución.

  Resultado esperado: Se especifican los entregables de cada actividad.

- Característica 1.5: ¿Se provee una guía de buenas prácticas para cada una de las actividades específicas?

  Para el usuario es recomendable contar no sólo con la definición de las actividades específicas sino también con una serie de consejos desde el punto de vista práctico que reduzcan los inconvenientes que le pudieran surgir durante su ejecución.

  Resultado esperado: Se proponen consejos prácticos a tener en cuenta en cada actividad.

1.3.0.5. Escenarios de aplicación

Los proyectos de minería de datos pueden ser desarrollados en diferentes escenarios. En algunos casos se desea obtener conocimiento para abordar algún problema y en otros se desea explorar los datos transaccionales en busca de patrones útiles[22].

- Característica 2.1: ¿Se especifican actividades para la definición y el análisis del problema con el que colaborará la minería de datos?

  El punto de partida del proyecto debe ser el problema para el que se desea obtener nuevo conocimiento. Esta actividad permite definir las fuentes de datos necesarias y las técnicas de minería más apropiadas. Requiere interiorizar la situación del usuario y definir claramente los objetivos del proyecto.

  Resultado esperado: En las primeras fases del proceso se consideran actividades para definir y analizar el problema en el que se enmarcará el proyecto.

- Característica 2.2: ¿Se consideran puntos de partida alternativos donde el usuario no refiere un problema, sino que solo desea explorar sus datos?

  En ocasiones el usuario manifiesta que desea implementar un proyecto de minería de datos para encontrar patrones ocultos en la información de su organización. Bajo esta situación, el analista estaría navegando por la información, buscando cualquier tipo de conocimiento

novedoso. En estos casos siempre existe una problemática o interés de trasfondo que motiva a dicha exploración.

Resultado esperado: En aquellos casos donde aparentemente se desea "explorar" la información organizacional, la metodología provee una guía de actividades para identificar problemas latentes que el usuario desconoce y que permitirán definir objetivos claros para el proyecto.

- Característica 2.3: ¿La metodología es independiente del dominio de aplicación?

  Los proyectos de minería de datos pueden llevarse a cabo en diversas áreas, tales como salud, industria, comercio, deporte o educación. Si bien cada dominio requiere de conocimientos específicos, las actividades principales del proceso de explotación de información deberían ser independientes del ámbito de aplicación.

  Resultado esperado: La metodología es general y no está condicionada a un dominio de aplicación en particular.

- Característica 2.4: ¿La metodología es aplicable a proyectos de diferente tamaño?

  Al igual que en otras disciplinas, los proyectos de minería de datos pueden ser de baja, mediana o gran envergadura. Una metodología debería poder gestionar proyectos de cualquier tamaño. El equipo de trabajo podrá seleccionar aquellas actividades que crea conveniente según el tamaño de su proyecto.

  Resultado esperado: La metodología puede utilizarse para proyectos de cualquier tamaño.

1.3.0.6. Actividades específicas que componen cada fase

En este punto, es necesario evaluar la incorporación de ciertas actividades relevantes que deberían estar presentes en el proceso de minería de datos. Se deben tener en cuenta las fases generales que componen el proceso y los detalles particulares del proyecto[22].
Fase de análisis del problema.

- Característica 3.1: ¿Se especifican los objetivos del proyecto y define un criterio de éxito para el proyecto?

  Una vez definido el problema u oportunidad de negocio, se debe especificar qué objetivos tendrá el proyecto desde el punto de vista organizacional y técnico. Consiste en explicitar cuándo se considerará que el proyecto de minería de datos ha logrado resultados aceptables para satisfacer los objetivos planteados.

Resultado esperado: Se especifican los objetivos del negocio y los objetivos técnicos del proyecto. Además, se definen los criterios de éxito para el proyecto en el plano organizacional y en el plano técnico.

Fase de selección y preparación de los datos.

- Característica 3.2: ¿Se propone un análisis exploratorio inicial de los datos, contemplan actividades para la transformación de variables y la creación de atributos derivados? Una vez seleccionados y recolectados los datos, es conveniente realizar un análisis exploratorio para familiarizarse con los mismos. Un estudio de la distribución y comportamiento de las variables, identificando además las tareas de limpieza que se deberán llevar a cabo. Generalmente los datos originales no tienen el formato que se necesita para la vista minable, razón por la cual resulta de gran importancia la transformación de las variables existentes y el cálculo de atributos derivados. Los atributos derivados son aquellos cuyo valor surge a partir de otros campos del mismo registro.

  Resultado esperado: Se efectúa un análisis exploratorio/descriptiva inicial de los datos recolectados y se especifican actividades para la transformación de los datos.

Fase de modelado.

- Característica 3.3: ¿Se efectúa una selección de las técnicas que se aplicarán y se planifica de qué forma se evaluarán los resultados?

  En función de la tarea de minería que se haya planteado (como clasificación o agrupamiento) existen diversas técnicas que el modelador puede utilizar. Sin embargo, no todas las técnicas son aplicables en todos los casos. Algunos factores pueden influir en la selección de las mismas, como el tamaño del conjunto de datos o bien la naturaleza de las variables en estudio.

  Luego de analizar detalladamente la factibilidad de cada técnica, se debería efectuar la selección final de aquellas que se utilizarán en esta fase de modelado. Es importante establecer cómo se evaluarán los resultados obtenidos por las mismas. En este punto debería definirse cómo se dividirá el conjunto de datos para el entrenamiento y prueba de los modelos. Por otro lado, debería establecerse cuál será el criterio para la ponderación de los resultados.

  Resultado esperado: Se efectúa un análisis y selección final de las técnicas de minería que se implementarán para la creación de los modelos. Una vez definida la técnica a aplicar se propone, la planificación de cómo se evaluarán los resultados obtenidos.

Fase de evaluación.

- Característica 3.4: ¿Se comparan y ponderan los modelos obtenidos?

  Luego de analizar cada modelo individualmente, es necesario establecer una ponderación para determinar cuáles son los más robustos y los que mejor se adecuan a los objetivos planteados.

  Resultado esperado: Se sugiere una comparación entre los modelos obtenidos, para posteriormente establecer una ponderación de los mismos en función de los objetivos técnicos y organizacionales.

1.3.0.7. Comparación entre las metodologías

Cada una de las características presentes en el marco comparativo puede evaluarse como ($\surd$) positiva si está presente en la metodología o de negativa en el caso de no estar marcada. No se definen niveles de cumplimiento debido al elevado grado de subjetividad. La evaluación final de la metodología estará asociada al por ciento de características que cumplimente, dentro del marco comparativo. En la tabla 1.1 se presenta la comparación realizada para las cinco metodologías más referenciadas en la bibliografía. Los datos necesarios para realizar la evaluación se obtuvieron de [20].

Características	CRISP-DM[18]	Catalyst[19]	SEMMA[25]	Cios et al[26]	Five A[27]
característica 1.1	√	√	√	√	√
característica 1.2		√			
característica 1.3	√	√	√	√	
característica 1.4	√	√		√	
característica 1.5	√				
característica 2.1	√	√	√	√	
característica 2.2		√		√	
característica 2.3	√	√	√	√	√
característica 2.4	√	√	√	√	√
característica 3.1	√	√	√	√	√
característica 3.2	√	√	√	√	√
característica 3.3	√	√		√	
característica 3.4	√	√	√		√

Tabla 1.1: *Comparación entre las metodologías*

Elaboración propia

Entre todas las metodologías comparadas la de mejor comportamiento, de acuerdo con el marco de comparación establecido para este trabajo es la metodología Catalyst con un 93.7 % de las características cumplidas. Esta es la principal razón que sustenta la selección de Catalyst como metodología para desarrollar el proyecto de correlación entre el test CHASIDE y el éxito en la carrera de Ingeniería en Ciencias Informáticas.

1.3.1. Metodología Catalyst

En el año 2003, Dorian Pyle propone en su libro "Business modelling and data mining" una metodología para el proceso de extracción de conocimiento en bases de datos llamada "Catalyst". En cuanto a su estructura, la metodología Catalyst está formada por dos partes (o sub-metodologías): Metodología para el Modelado del Negocio y Metodología para la Minería de Datos[19].

La sub-metodología de Modelado del Negocio cuenta con una guía de pasos para modelar el problema u oportunidad de negocio que abordará el proyecto. Contempla diferentes ámbitos en el proyecto de

minería de datos, recomendando una guía de pasos para llevar a cabo en cada escenario específico. Pyle propone cinco situaciones o puntos de partida diferentes para el proyecto:

- Escenario 1: Datos

  Explorar los datos en búsqueda de relaciones útiles e interesantes.

  1. Determinar las fuentes de donde se recolectarán los datos.
  2. Identificar al personal interesado (stakeholders) en el proyecto.
  3. Discutir el proyecto original con el personal interesado.
  4. Caracterizar el conjunto de datos en función de las relaciones P3TQ por las cuales fueron recolectados.
  5. Caracterizar la motivación del negocio para recolectar y almacenar los datos.
  6. Descubrir quién o qué departamento originó el proyecto y qué expectativas tienen sobre el mismo.
  7. Descubrimiento del problema:

     a) Identificar las principales relaciones P3TQ que dan origen a los datos.
     b) Identificar y caracterizar al personal interesado.
     c) Identificar los objetos organizacionales que los datos representan.
     d) Enmarcar el problema u oportunidad.
     e) Preparar un esbozo del caso de negocio.
     f) Presentar el nuevo proyecto al personal interesado.
     g) Armar el caso de negocio completo, si es necesario.
     h) Enmarcar y describir la situación del negocio.
     i) Definir los requerimientos de la implementación.

- Escenario 2: PROBLEMA/OPORTUNIDAD

  Dado un problema u oportunidad de negocio, ver cómo la minería de datos puede colaborar con la misma.

  1. Identificar y caracterizar al personal interesado relevante.
  2. Explorar la situación de negocio con el personal interesado.
  3. Enmarcar y describir la situación del negocio.
  4. Identificar los objetivos de negocio relevantes para el proyecto.
  5. Buscar los datos que se explotarán.
  6. Armar el caso de negocio.
  7. Presentar el caso de negocio al personal interesado.
  8. Describir la situación del negocio para el proceso de minería.
  9. Definir los requerimientos de implementación.

- Escenario 3: PROSPECCIÓN

  Proyecto diseñado para descubrir dónde la minería de datos puede aportar valor en la organización.

  1. Caracterizar las relaciones P3TQ claves de la organización.
  2. Identificar el flujo de los principales procesos de la organización.
  3. Identificar el personal interesado.
  4. Entrevistar al personal interesado.
  5. Descubrir qué "cambios estratégicos" pueden resultar de mayor interés para cada usuario.
  6. Caracterizar los modelos de minería que pueden dar soporte a los cambios estratégicos.
  7. Explorar las fuentes de datos.
  8. Preparar un borrador del caso de negocio para cada oportunidad significativa.
  9. Presentar los casos de negocio al personal interesado.
  10. Enmarcar la situación de negocio que se abordará.
  11. Definir los requerimientos de implementación.

- Escenario 4: MODELO DEFINIDO

  Utilizar la minería de datos para construir un modelo específico para una situación determinada.

  1. Identificar al personal interesado.
  2. Discutir los requerimientos con el personal interesado.
  3. Enmarcar la situación del negocio.
  4. Buscar los datos a minar.
  5. Definir los requerimientos de la implementación.

- Escenario 5: ESTRATEGIA

  Dada una situación estratégica, analizar si la minería de datos puede ser útil para explicar la situación actual y descubrir cuáles son las opciones para resolverla. Es decir, el proyecto se inicia requiriendo un análisis estratégico para apoyar la planificación de escenarios corporativos.

  1. Identificar al personal interesado.
  2. Entrevistar al personal interesado.
  3. Enmarcar la situación de negocio.
  4. Si es necesario, trabajar iterativamente con el personal interesado para crear un mapa de los escenarios estratégicos.
  5. A partir del mapa crear un modelo de la situación estratégica, mediante un enfoque de sistema.
  6. Caracterizar las relaciones P3TQ más importantes de la organización.

7. Relacionar el mapa con las relaciones identificadas.
8. Si es necesario, simular la situación estratégica para identificar ambigüedades, incertidumbres, errores y descubrir relaciones.
9. Caracterizar las relaciones P3TQ en términos de los "cambios estratégicos".
10. Descubrir qué "cambios estratégicos" pueden resultar de mayor interés para cada usuario.
11. Caracterizar los "cambios estratégicos" más viables.
12. Explorar las fuentes de datos.
13. Enmarcar cada oportunidad/problema de negocio en el modelo estratégico con particular atención a las estrategias y sus interacciones, incluyendo los riesgos que implican cada una.

Tomando algunos de estos cinco puntos o escenarios de partida, el autor de la metodología Pyle propone una serie de pasos y herramientas para llegar a descubrir el problema y los requerimientos organizacionales que abordará el proyecto, así como los datos necesarios para efectuar el análisis. Cualquiera sea el escenario de partida, siempre se modela el problema que el proyecto de minería abordará, obteniendo:

1. Los datos necesarios para minar.
2. Los requerimientos reales, considerando las expectativas y necesidades del usuario.

Al aplicar la metodología, el segundo paso o parte de esta es la Metodología para la Minería de Datos en la cual se proporciona una guía de pasos para el descubrimiento de patrones/relaciones de acuerdo al problema de negocio identificado. Estos pasos se hacen llamar muy parecidos a los descrito por el modelado de KDD y siguen la siguiente estructura:

1. Preparación de los datos:

    - Evaluar las variables en estudio (medidas de posición y dispersión, outliers, datos faltantes, etc).
    - Chequear problemas básicos en las variables (variables con muchas categorías por ejemplo).
    - Chequear problemas en la base de datos completa, análisis multivariado (CHAID analysis).
    - Chequear variables anacrónicas (que no aportan información).
    - Chequear que haya suficientes datos.
    - Chequear que se cubran todos los valores posibles de las variables, aun los que no son de interés.

- Chequear la necesidad de recodificar variables.

2. Selección de herramientas y modelado inicial

   - Estructurar los datos para el proceso de minería (dividir los datos de entrenamiento, prueba y evaluación).
   - Caracterizar las variables de entrada y salida.
   - Seleccionar el algoritmo de minería.
   - Evaluar el impacto de los datos faltantes mediante un MVCM (Missing Value Check Model).
   - Crear un modelo inicial:

     a) Exploratorio / Descriptivo: Si se realiza minería para entender la situación del negocio.
     b) De Clasificación: si se realiza minería para clasificar.
     c) De Predicción: si se realiza minería para predecir.

3. Refinar el modelo

   - Si el método es exploratorio, describir los resultados encontrados sobre la situación actual.
   - Si el modelo es predictivo o de clasificación, verificar la capacidad predictiva del modelo (por ejemplo, con matrices de confusión o gráficos de "valores predichos vs observados", según el caso).
   - Verificar el modelo con el personal interesado.

4. Implementar el modelo

   - Si el modelo es exploratorio: se deben revisar los requerimientos del problema, elaborar un informe con los resultados del descubrimiento, contabilizar los valores extremos, incorporar evidencia negativa, incorporar evidencia empírica/experimental y obtener realimentación de los usuarios.
   - Si el modelo es de clasificación/predicción: se deben revisar los requerimientos de la implementación planteados antes de la minería, revisar los requerimientos del problema, preparar una explicación del modelo y revisar los requerimientos finales de implementación.
   - Comunicación y difusión de resultados.

1.4. Conclusiones del capítulo

Luego de un estudio detallado de los conceptos básicos de los tipos de software, específicamente los de predominio de cómputo, se decide implementar una herramienta que permita realizar el proceso de extracción de conocimiento en bases de datos. Esta se realiza mediante el uso de los modelos de reglas de asociación, para obtener como resultado un modelo que describa las aptitudes e intereses presentes en el ingeniero en Ciencias Informáticas. El proceso de extracción de conocimiento se realizará mediante la metodología Catalyst, la cual describe detalladamente los pasos a realizar y se ajustan las tareas a las características del problema.

Capítulo 2

Extracción de las correlaciones

2. OBTENCIÓN DEL MODELO DE REGLAS DE ASOCIACIÓN

EN este capítulo se aplica la metodología Catalyst para construir el modelo de reglas de asociación que facilita la obtención de correlaciones entre el test de orientación vocacional CHASIDE y los resultados alcanzados por estudiantes de la carrera de Ingeniería en Ciencias Informáticas. El modelo responde a la necesidad de la universidad de mejorar los resultados docentes alcanzados, especialmente en la disminución del fracaso escolar. La figura 2.1 presenta la relación del modelo de reglas con los objetivos de la institución.

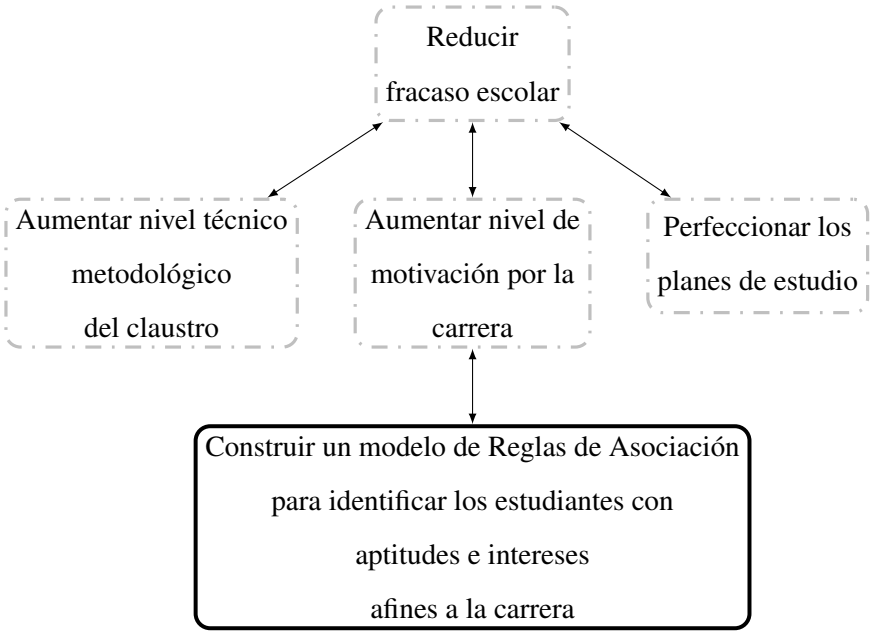


Figura 2.1: *Objetivos del negocio relevantes para el proyecto*

Elaboración propia

La figura pretende explicar cómo darle solución a unos de los problemas fundamentales por lo cual ocurre el fracaso escolar. El modelo de Reglas de Asociación que se pretende construir para identificar los estudiantes con aptitudes e intereses a fines a la carrera, permitirá combatir la mala orientación vocacional que pueda tener un estudiante hacia la carrera.

2.1. Obtención de los datos a explorar

Los datos necesarios para realizar la exploración se obtienen de la aplicación del test de orientación vocacional CHASIDE y la clasificación de los estudiantes se realiza de acuerdo a sus resultados en la carrera como sigue a continuación:

- Rendimiento normal índices entre 3.16 y 3.99.
- Buen rendimiento para valores entre 4 y 4.75.
- Alto rendimiento promedio por encima de 4.75.

Estas clasificaciones responden a la intención de un primer análisis exploratorio y no pretenden ser un criterio único para determinar el éxito de un estudiante en la carrera, esta fue definida a partir de el criterio de un expertos.

Se implementó una aplicación Web para recopilar los datos, la aplicación cuenta con tres vistas, la principal se muestra en la figura A.1, donde se responde las preguntas referentes al test CHASIDE.

Figura 2.2: *Vista principal del test CHASIDE*

Elaboración propia

Desde el punto de vista del análisis de las correlaciones el elemento fundamental a tener en cuenta de esta aplicación Web es el modelo de datos. El mismo es presentado en la figura 2.3. La base de datos contiene los resultados de cada una de las encuestas. El desarrollo de la base de datos se realizó utilizando PostgreSQL como motor de persistencia.

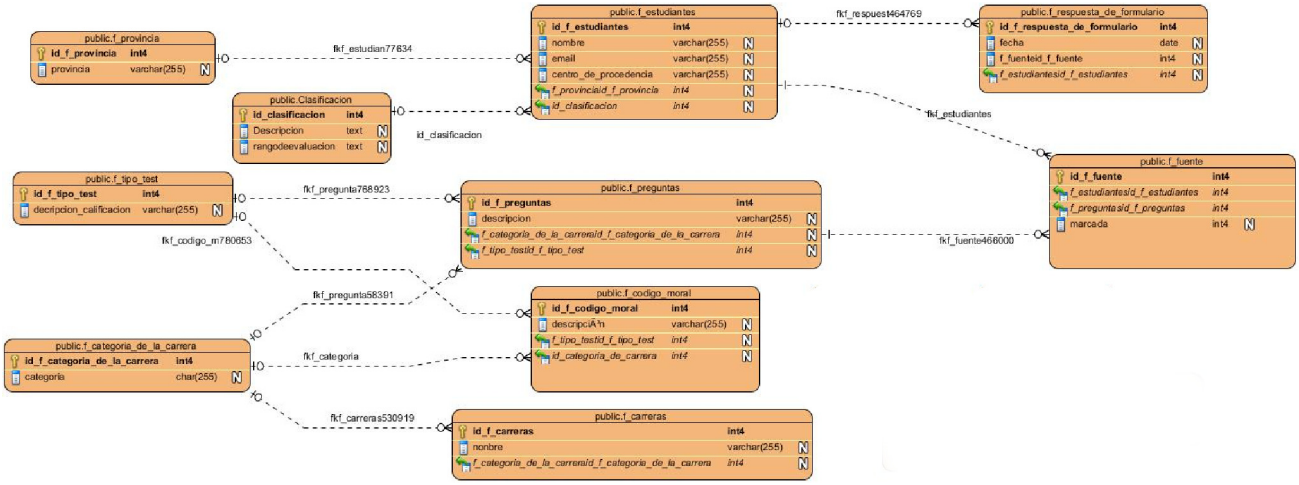


Figura 2.3: *Modelo físico de la base de datos del sistema CHASIDE*

Elaboración propia

La información relevante a la investigación se almacena en las siguientes tablas:

- Tabla f_estudiantes figura 2.4, cuenta con los siguientes campos:

  -id_f_estudiantes: este campo identifica a cada encuestado.

  -nombre: este campo almacena el Nombre del encuestado

  -email:este campo almacena el correo del encuestado.

  -centro_de_procedencia: este campo almacena la procedencia de cada encuestado.

  -f_provinciaid_f_provincia: este campo almacena el identificador de la provincia.

  -id_clasificacion: este campo almacena el identificador de la clasificación dada al encuestado y solo va a estar lleno para los encuestados de interés a ser analizados.

Figura 2.4: *Diagrama físico de la tabla: f_estudiantes*

- Tabla Clasificacion, figura 2.5, cuenta con los siguientes campos:

  -id_clasificacion: este campo identifica las cuatro clasificaciones posibles.

-Descripcion: este campo describe la clasificación dada.

-rangodeevaluacion: este campo almacena los principales rasgos por lo cual se dio la clasificación.

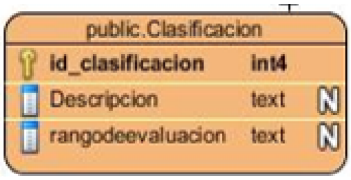


Figura 2.5: *Diagrama físico de la tabla: Clasificacion*

- Tabla f_fuente figura 2.6, cuenta con los siguientes campos:

  -id_f_fuente: este campo identifica de cada pregunta por cada encuestado la respuesta dada.

  -f_estudiantesid_f_estudiantes: este campo hace referencia al encuestado.

  -f_preguntasid_f_preguntas: este campo hace referencia a la pregunta que marcó o no el encuestado.

  -marcada: este campo almacena 1 si el encuestado marcó la pregunta o 0 si no.

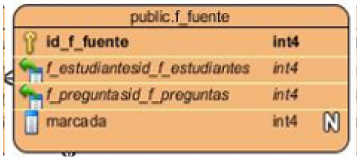


Figura 2.6: *Diagrama físico de la tabla: f_fuente*

- Tabla f_pregunta figura 2.7, cuenta con los siguientes campos:

  -id_f_preguntas: este campo identifica a la pregunta del test CHASIDE.

  -descripcion: este campo describe la pregunta.

  -f_categoria_de_la_carreraid_f_categoria_de_la_carrera: este campo hace referencia a la categoría de la pregunta dentro del test CHASIDE.

  -f_tipo_testid_f_tipo_test: este campo hace referencia al tipo de la pregunta dentro del test CHASIDE la cual puede ser 1 para aptitud o 2 para interés.

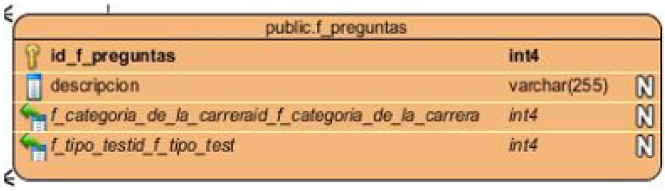


Figura 2.7: *Diagrama físico de la tabla: f_fuente*

2.2. Caracterización de las variables

La caracterización de las variables, para este caso, consiste en hacer un análisis exploratorio de los resultados obtenidos del test CHASIDE y verificar por cada encuestado el estado en que está cada pregunta, que constituyen las variables a tener en cuenta a la hora de construir el modelo. Lo primero que se hace para realizar esta tarea, es extraer de los datos que se van a analizar para luego, realizar un reporte del estado de los datos.

2.2.1. Extracción de los datos desde las fuentes

Los datos a explorar, que constituyen los valores de las variables del problema, se obtiene mediante la función SQL presentada en la figura 2.8. En la misma se genera un registro para cada estudiante que realizó el test y se le asocia el valor(true o false) a cada pregunta del test. La representación final contendrá en cada fila los números de las preguntas marcadas en true. El total de preguntas es de 98 más un atributo en el que se refleja una categoría asociada al aprendizaje.

```
CREATE OR REPLACE FUNCTION pregunta_por_estudiante()
  RETURNS SETOF text AS
$BODY$
declare
f record;
est integer;
res text;
z text;
clas integer;
begin
FOR est IN (SELECT f_estudiantes.id_f_estudiantes FROM  public.f_estudiantes) loop
res:=est||'->';
FOR f IN (SELECT f_preguntasid_f_preguntas as id_es FROM public.f_fuente
where f_estudiantesid_f_estudiantes=est)loop
res=res || f.id_es|| ',';
end loop;
SELECT f_estudiantes.id_clasificacion into clas
FROM public.f_estudiantes WHERE id_f_estudiantes=est;
case clas
when 1 then res:=res||'99';
when 2 then res:=res||'100';
when 3 then res:=res||'101';
when 4 then res:=res||'102';
else
res:='borrar';
END CASE;
z:=res;
return next z;
end loop;
end
$BODY$
  LANGUAGE plpgsql;
```

Figura 2.8: *Función SQL para explorar los datos.*

Elaboración propia

2.2.2. Reporte de descripción de los datos:

Después del análisis exploratorio de los datos se puede afirmar que no existen datos ausentes debido a que todas las preguntas fueron contestadas por cada encuestado. El 100 % de los encuestados están clasificados de acuerdo a su promedio de notas.
La diversidad de la distribución de las clases es homogénea, hay un 35 % clasificados de "Avanzado", un 33 % clasificados de "Satisfactorio" y un 32 % evaluado de "Bueno". Las principales variables que son las clasificaciones de los encuestado y las preguntas están lo suficientemente completas para poder seleccionar la información deseada. Lo próximo que se debe hacer es verificar en qué estado está la información que se va a extraer.
También tras el análisis exploratorio de los datos se puede afirmar que:

- variables con un único valor en el 95 % de las instancias son: (50,69,8,67,74,21,29,7,63)
- no aparecen variables con datos ausentes en las instancias.
- no aparecen variables con muchas categorías únicas.

El análisis exploratorio de las variables arroja que para la construcción del modelo no se van a tener en cuenta las variables con valor único ya que siempre van a estar incluidas en todas las reglas.

2.3. Chequear problemas en el conjunto de datos.

Al analizar los datos no se encontraron patrones inesperados durante la fase exploratoria, por lo que no es necesario una limpieza de los datos. Las preguntas 50, 69, 8, 67, 74, 21, 29, 7 y 63 siempre tienen el mismo valor, por ello se consideran variables irrelevantes para el modelo ya que van a ser fijas en las reglas, no tienen peso para la construcción del modelo.
Tras dividir el conjunto de datos en diferentes partes para observar el comportamiento de las variables, se ha evidenciado que este se mantiene con respecto a la estructura, por lo cual se concluye que los datos son suficientes para la clasificación y están correctamente estructurados. Afirmándose que se puede proceder a realizar la preparación y modelado de los datos.
Tras el análisis de las variables y sus características, se analizan los software existentes actualmente en el mercado para el minado de reglas de asociación. Atendiendo a algunas características como son: Licencia, Acceso a SQL, Modelos de clasificación, entre otros, los cuales tienen un gran número de herramientas para la explotación de la minería de datos, algunas con mejores características en algunos

aspectos, pero casi todas son diseñadas con la misma finalidad, se muestra un cuadro comparativo a continuación:

Características	Clementine[28]	SAS Enterprise Miner[29]	Tariykdd[30]	Weka[31]
Licencia libre			√	√
Requiere conocimientos avanzados de Minería de datos			√	
Acceso a SQL	√		√	√
Multiplataforma		√	√	√
Requiere bases de datos especializadas		–		
Modelos de clasificación	√	√	√	√
Puede combinar modelos	√	√	√	√

Tabla 2.1: *Comparación entre los softwares para procesar datos.*

Elaboración propia

La comparación de los software que procesan datos para realizar la extracción del conocimiento, arroja que usando estos software no se solucionara el problema de correlacionar habilidades y competencias que representan las preguntas del test CHASIDE con el rendimiento académico ya que ninguno procesa datos de bases de datos centralizadas. Además, es engorroso procesar esta cantidad de datos con los software convencionales porque son poco adecuados para analizar gran cantidad de columnas. También, aunque el Weka es uno de los software más utilizado para la extracción de conocimientos por sus características, este no se actualiza desde el año 2011. Otras de sus desventajas, es el algoritmo que tiene implementado para el minado de reglas de asociación el cuál, solo permite como entrada datos binarios desfavoreciendo la estructura que se desea usar para correlacionar los datos. [31].

Lo planteado anteriormente hace que se decida implementar un algoritmo que permita controlar los parámetros de uso de memoria y almacenamiento, que son problemas comunes que surgen en el

proceso de datos con gran cantidad de columnas. Además, que permita procesar datos que no sean binarios para lo cual los datos se estructuran de la manera siguiente, para que puedan ser procesados por el algoritmo que se implementará.
Estructura de los datos:

- √ Preguntas de tipo aptitud(identificador).
- √ Preguntas de tipo interés(identificador).
- √ Clasificador.

La estructura planteada para el procesamiento de los datos cuenta con 98 preguntas que se representan de manera numérica por el identificador de la pregunta, además se le agrega una columna para poder clasificar cada encuestado. Esto trae como inconveniente una estructura engorrosa de procesar por el número de columnas que tiene, haciendo necesario buscar un mecanismo capaz de disminuir la dimensionalidad de las columnas. Después de analizados los datos del problema se caracterizan las variables de entrada y salida.

2.3.1. Caracterización de las variables de entrada y salida.

Teniendo como entrada la estructura anteriormente definida, se caracterizan las variables que la conforman. Al ser las preguntas parte de estas variables, se decide agruparlas por perfiles, teniendo en cuenta la clasificación que plantea el test CHASIDE (tomando como rango para el perfilado las columnas de cada categoría). Los perfiles se hallaron al recorrer las preguntas de cada encuestado, tomando la lista de preguntas por cada rango definido y comparándolas con la lista de perfiles creados, en caso de no ser iguales, se crea un perfil con dicha lista. Al perfilar los datos, se reduce el número de columnas a 15, 7 perfiles de tipo aptitud, 7 perfiles de tipo interés y la clasificación del encuestado, así se garantiza que hallan menos filas para procesar.
En la solución se van a omitir todas aquellas preguntas o variables que tengan un valor único o esté su soporte por encima del $95\,\%$. La variable de salida es un conjunto de reglas de asociación, el cual va estar conformado por la combinación de los perfiles que tienen en común cada clase o clasificación que va a implicar a dicha clasificación. La variable de salida se va a visualizar en forma de reglas de asociación, la cual debe tener un soporte y una confianza(Ejemplo: $\{Lista\_de\_perfiles(combinacion\_de\_perfiles)\} \rightarrow clasificador, Sop : X, Conf : Y$, osea 69,214,215,14,48 $\rightarrow$ 3,Confianza:0.571414,Soporte:1.0).

El resultado esperado es un modelo de reglas de asociación que permite correlacionar el comportamiento académico de los estudiantes en la carrera de Ingeniería en Ciencias Informáticas. El modelo será implementado como parte del sistema donde están almacenados los datos del test CHASIDE, para que sea posible hallar la correlación entre el test con otras carreras, convirtiéndose en una herramienta adaptable para el estudio de cualquier carrera. Con esta información se podrá predecir el comportamiento de un egresado del nivel medio superior en una carrera donde se halla hecho el estudio, este caso en particular, para la carrera de Ingenierías en Ciencias Informáticas. El resultado también, servirá para evaluar las competencias e intereses de los profesionales en una carrera.

2.3.2. Seleccionar un algoritmo de extracción de reglas

En las técnicas de minería de datos existentes para encontrar correlaciones es esencial la selección del algoritmo, partiendo del enfoque conceptual para extraer información de los datos y llegando al resultado esperado. Cada algoritmo representa la manera de desarrollar paso a paso una determinada técnica o tarea, por tanto, es preciso entender los parámetros y características del algoritmo para preparar los datos a analizar. A continuación, se describe el funcionamiento de algunos algoritmos.

2.3.3. Algoritmo Eclat

El algoritmo Eclat aprovecha los item directamente para la computación de soporte. El paso de conteo de soporte se puede mejorar significativamente indexado la base de datos de tal manera que permita el cálculo de frecuencia rápida. Observe que en el enfoque de nivel-sabio, para contar el soporte, tenemos que generar subconjuntos de cada transacción y comprobar si existen en el árbol de prefijos. Esto puede ser costoso porque puede generar muchos subconjuntos que no existen en el árbol de prefijos[32].

La idea básica es que el apoyo de un candidato itemset se puede calcular mediante la intersección de los item de subconjuntos elegidos adecuadamente. En general, dado t (X) y t (Y) para dos conjuntos de elementos frecuentes X e Y, teniendo:

$T(XY) = t(X) \cap t(Y)$

El apoyo del candidato XY es simplemente la cardinalidad de t (XY), es decir, $sup(XY) = |t(XY)|$. Eclat intersecta los item sólo si los conjuntos de elementos frecuentes comparten un prefijo común y atraviesa el árbol de búsqueda de prefijo, procesando un grupo de conjuntos de elementos que tienen el mismo prefijo, también llamado una clase de equivalencia de prefijo[33].

2.3.4. Algoritmo Frequent Pattern Tree Approach: FPGrowth

El método FPGrowth indexa la base de datos o vistas minables para el cómputo de soporte rápido mediante el uso de un árbol de prefijo aumentado llamado árbol de patrón frecuente (árbol FP). Cada nodo en el árbol está etiquetado con un único elemento y cada nodo secundario representa un elemento diferente[34].

Cada nodo almacena también la información de soporte para el conjunto de items que comprende los elementos la ruta de la raíz a ese nodo. El árbol FP se construye de la siguiente manera. Inicialmente el árbol contiene como elemento inicial el elemento nulo $\emptyset$. A continuación, para cada tupla $t, X \in D, dondeX = i(t)$, se inserta el conjunto de elementos X en el árbol FP, incrementando el recuento de todos los nodos a lo largo de la ruta que representa X. Si X comparte un prefijo con alguna transacción insertada anteriormente, entonces X seguirá el mismo camino hasta el prefijo común. Para los artículos restantes en X, los nuevos nodos se crean bajo el prefijo común, con los recuentos inicializados a 1. El árbol FP está completo cuando se han insertado todas las transacciones[33].

El FP-árbol puede ser considerado como un prefijo comprimido representación de D. Se quiere un árbol que sea lo más compacto posible y tener los elementos más frecuentes en la parte superior del árbol. Para lograr esto, FPGrowth reordena los items en orden de soporte, es decir, a partir de la base de datos inicial, primero calcula el apoyo de elementos individuales $i \in I$. A continuación, descarta los elementos poco frecuentes y ordena los elementos frecuentes disminuyendo el apoyo. Finalmente, cada tupla $t, X \in D$ se inserta en el árbol FP después de reordenando de X disminuyendo el soporte del item.

2.3.5. Apriori

Este algoritmo está basado en la reducción de conjuntos, centrándose en un soporte mínimo(sop_min) introducido al igual que el Algoritmo BUTE FORCE [35]. En [36] se define, el sop(X) como la proporción de transacciones que contienen el conjunto X, donde I es un conjunto de elementos, y se utilizará |A| para denotar la cardinalidad del conjunto se utiliza, A(sop(X)=($|\{I|I \in D \wedge I \supseteq X\}|$)/$N$). De la definición de soporte se tiene que, si sop(A∪C)$\leqslant$ sop_min entonces sop(A $\rightarrow$ C) $\leqslant$ sop_min. Apriori genera todos los conjuntos que cumplen con la condición de tener un soporte menor o igual a sop_min. Para cada conjunto frecuente X se generan todas las reglas de asociación A $\rightarrow$ C tal que A∪C= A y A∩ C = 0. Cualquier regla que no satisfaga las restricciones impuestas por el usuario, como por ejemplo la confianza mínima, se desecha, y las reglas que sí las cumplen se conservan.

Como el sop(A) $\geq$ sop(A$\rightarrow$C) y sop(C)$\geq$ sop(A$\rightarrow$C), si A $\bigcup$ C es un conjunto frecuente entonces tanto A como C son conjuntos frecuentes. El soporte, la confianza y otras métricas por las cuales las reglas de asociación A $\rightarrow$ C son evaluadas y se puede usar el sop(A), sop(C) y sop(A $\cup$ C) como referencias. El algoritmo Apriori se utiliza para la generación de items-set frecuentes. Este algoritmo se resume en dos pasos fundamentales para obtener los resultados[33]:

Paso1:

Generación de todos los item-sets que contienen un solo elemento, utilización de estos para generar item-sets que contengan dos elementos, y así sucesivamente. Se toman todos los posibles pares de items que cumplen con las medidas mínimas de soporte inicialmente preestablecidas; esto permite ir eliminando posibles combinaciones: aquellas que no cumplan con los requerimientos de soporte no entrarán en el análisis.

Paso2:

Generación de las reglas revisando que cumplan con el criterio mínimo de confianza. Es interesante observar que, si una conjunción de consecuentes de una regla cumple con los niveles mínimos de soporte y confianza, sus subconjuntos (consecuentes) también los cumplen; en el caso contrario, si algún ítem no los cumple no tiene caso considerar sus súper conjuntos.

Después del análisis de los principales aspectos técnicos de los algoritmos más utilizados en tareas descriptivas y las reglas de asociación, se puede concluir que de los algoritmos analizados el más sencillo de implementar es Apriori, por lo que se escoge para ser utilizado en la investigación. Aprovechando las características de este algoritmo después de realizar el primer paso se pudiera restringir en el segundo paso a que solo genere las reglas donde el consecuente sea uno de los clasificadores. Generando las reglas donde aparezca un solo elemento como consecuente implicado a un conjunto, si restringimos al algoritmo a la primera iteración podríamos obtener reglas de clasificación. Estas reglas son muy parecidas a las de asociación con la diferencia de que un solo consecuente implica a un conjunto de item-set. Ejemplo:(x$\rightarrow$ y,z,v,n).

Una vez definido el modelado del problema se procede a la implementación, para ello se analiza en que lenguaje se debe implementar el algoritmo qué se va a utilizar. Entre los lenguajes de programación más populares para la solución de problemas de minerías de datos se encuentra Java, C++, R y Python. Se escoge Python como lenguaje para la implantación del algoritmo ya que es muy fácil integrarlo a la solución del problema por su compatibilidad con las bases de datos en especial con PostgreSQL.

2.4. Construcción del modelo

Como se ha establecido en las secciones anteriores lo primero que se necesita es seleccionar los datos con que se va a trabajar para luego crear la vista minable. Teniendo en cuenta que se necesitan todos los perfiles por cada tipo de test dígase interés y aptitud con la clasificación dada a cada encuestado, se construyen 2 funciones para poder extraer los datos y transformarlos en la vista minable que se necesita. Luego de estructurar los datos se implementa el algoritmo Apriori en Python el cual tiene como entrada la vista minable, el soporte mínimo, la confianza mínima y los items, devolviendo como salida una lista de reglas de asociación la cual se guardan en la base de datos.

2.4.1. Estructurar los datos para el proceso

La primera función (figura 2.9), se crean y devuelven todos los perfiles asociados a un estudiante por un tipo de test específico, obviando las preguntas que tienen como soporte entre 0.95 a 1

```
CREATE OR REPLACE FUNCTION perfil(est integer, tipo integer)
RETURNS integer[] AS
$BODY$
declare
f record;
res integer[];
tip_p integer;
resul integer[];
begin
for tip_p IN(SELECT f_categoria_de_la_carrera.id_f_categoria_de_la_carrera as p
FROM   public.f_categoria_de_la_carrera order by f_categoria_de_la_carrera.id_f_categoria_de_la_carrera)loop
res:=ARRAY []::integer[];
FOR f IN (SELECT
f_preguntas.id_f_preguntas as preg,f_fuente.marcada  FROM public.f_fuente, public.f_preguntas
WHERE
f_preguntas.id_f_preguntas = f_fuente.f_preguntasid_f_preguntas and f_estudiantesid_f_estudiantes= $1 and
f_preguntas.id_f_preguntas not in(SELECT f_fuente.f_preguntasid_f_preguntas as borrar
 FROM public.f_fuente Group by f_fuente.f_preguntasid_f_preguntas, marcada Having count(*)>30)
and f_preguntas.f_categoria_de_la_carreraid_f_categoria_de_la_carrera=tip_p and
f_preguntas.f_tipo_testid_f_tipo_test=$2 and marcada=1 order by f_preguntas.id_f_preguntas) loop
res:=res||f.preg;
end loop;
if(exists(SELECT prefil FROM perfil where prefil=res))then
resul:=resul||(SELECT id_perfil FROM perfil where prefil=res);
else
INSERT INTO perfil(prefil) VALUES (res);
resul:=resul||(SELECT id_perfil FROM perfil where prefil=res);
end if;
end loop;
return resul;
end
$BODY$
LANGUAGE plpgsql;
```

Figura 2.9: *Función SQL para perfilar las preguntas y devolver los perfiles dado un estudiante y un tipo de test. Elaboración propia*

Una vez que se perfilan los datos se crea otra función (figura 2.10) para automatizar el proceso con el objetivo que recorra todos los estudiantes. La salida de la función se exporta a un archivo en el formato csv que permite ser utilizado por el algoritmo implementado.

```
CREATE OR REPLACE FUNCTION perfilest()
RETURNS SETOF text AS
$BODY$
declare
est record;
resul integer[];

begin
FOR est IN SELECT f_estudiantes.id_f_estudiantes as e FROM  public.f_estudiantes loop
resul:=(select perfil(est.e, 1));
resul:=resul||(select perfil(est.e,2));
resul:=resul||(SELECT id_clasificacion FROM  public.f_estudiantes
 where f_estudiantes.id_f_estudiantes = est.e);
return next array_to_string(resul, ',', '*');
end loop;
end;
$BODY$
LANGUAGE plpgsql;
```

Figura 2.10: *Función SQL para crear la vista minable.*

Elaboración propia

La estructura de la vista minable creada está compuesta por una fila que constituye un encuestado con los perfiles 7 para aptitud, 7 para interés y la última columna hace referencia a la clasificación. A continuación, se describe la estructura:

C H A S I D E							C H A S I D E							Clase
Perfil 1	Perfil 2	Perfil 3	Perfil 4	Perfil 5	Perfil 6	Perfil 7	Perfil 8	Perfil 9	Perfil 10	Perfil 11	Perfil 12	Perfil 13	Perfil 14	Clasificación
Intereses							Aptitudes							

Tabla 2.2: *Estructura de la vista minable.*

Elaboración propia

2.4.2. Proceso de minería de datos a través del algoritmo Apriori.

Una vez definida la estructura ya están listos los datos para ser procesados por el algoritmo seleccionado(Algoritmo 1). El funcionamiento del algoritmo, se implementó en una clase en Python y se adaptó para obtener reglas de clasificación. A continuación, se muestra el pseudocódigo del

algoritmo, el cual tiene como propósito generar los item frecuentes y luego genera las reglas dado un conjunto de item frecuentes a partir de las clases clasificadoras.

Algoritmo 1 Apriori método básico

Entrada: D=(conjunto de datos a procesar o vista minable) , I=(perfiles y su conjunto de preguntas) Sop=(soporte mínimo), Conf=(confianza mínima) ,Clases=(clase para clasificar)

Salida: Conjunto de reglas(F←Regla n: premisa *to* conclusiones Sop=x ,Conf=y).

1: $F \leftarrow \phi$
2: $C^{(1)} \leftarrow \{\phi\}$ { Inicialisa la lista de Item frecuentes}
3: **para todo** $i \in IdoAumentaricomohijodeφenC^{(1)}consup(i) \leftarrow 0$ **hacer**
4: $k \leftarrow 1$ {k indica el grado del item}
5: **mientras** $C^{(k)} \neq \phi$ **hacer**
6: CALCULA_SOP($C^{(k)}$,D)
7: **para todo** $premisaX \in C^{(k)}do$ **hacer**
8: **si** sop(X)$\geq$ Sop **entonces**
9: **si** Clases $\in$ X **entonces**
10: $F \leftarrow F \cup \{X, sop(X)\}$
11: **fin si**
12: **si no**
13: eliminar X de $C^{(k)}$
14: **fin si**
15: **fin para**
16: $C^{(k+1)} \leftarrow Cacular\_Item\_n(C^{(k)})$
17: $k \leftarrow k + 1$
18: **fin mientras**
19: **devolver** F
20: **fin para**

A continuación, se describen los métodos auxiliares utilizados para calcular el soporte y los item frecuentes :

Algoritmo 2 Apriori: métodos auxiliares

1: CALCULA_SOP($C^{(k)}$,D)
2: **para todo** $\langle t, i(t)\rangle \in Ddo$ **hacer**
3: **si** $X \in C^{(k)}$ **entonces**
4: $sop(X) \leftarrow sop(X) + 1$
5: **fin si**
6: **fin para**
7: $Cacular\_Item\_n(C^{(k)})$
8: **para todo** $premisaX_a \in C^{(k)}do$ **hacer**
9: **para todo** $premisaX_b \in Proximo\_Item(X_a), talqueb > ado$ **hacer**
10: $X_{ab} \leftarrow X_a \cup X_b$ {Podar el candidato si hay subconjuntos infrecuentes}
11: **si** $X_j \in C^{(k)}, paratodoX_j \in X_{ab}, talque\|X_j\| = \|X_{ab}\| - 1$ **entonces**
12: Adicionar $X_{ab}aloshijosdeX_a, consop(X_{ab} \leftarrow 0)$
13: **fin si**
14: **fin para**
15: **si** no exite para X_a **entonces**
16: elimina $X_aytodoslosantecedentesdeX_a, paraC^{(k)}$
17: **fin si**
18: **fin para**
19: **devolver** $C^{(k)}$
20:

Las reglas encontradas a partir del algoritmo 1 son procesadas para convertir los perfiles en las correspondientes listas de preguntas. Este proceso generó 295 reglas de las cuales se escogieron las de mayor confianza y soporte para plantear el modelo.

2.4.3. Modelo de reglas

Una vez ejecutado el modelo con un soporte mayor a 0.4 y una confianza mayor de 0.3 se obtienen 295 reglas de las cuales 52 reglas clasifican a la categoría de satisfactorio, 64 reglas clasifican a la categoría de bueno y 179 reglas clasifican a la categoría de avanzado. De las reglas encontradas al ejecutar el proceso se muestra un conjunto con algunas de ellas en la tabla 1.1 :

Lista de preguntas	Clasificador	Soporte	Confianza
9,25,59,90,16,31,42,61,77,88,93,72,14,38,51,17,66,0	Avanzado	0.606061	1
72,86,5,18,26,37,47,54,60,83,97,6,55,79,94	Avanzado	0.606061	1
5,18,26,37,47,54,60,83,97,14,38,46,51,6,55,79,94	Avanzado	0.606061	1
9,25,59,10,20,27,35,45,57,81,96,38,46,51	Bueno	0.606061	1
9,25,59,90,16,31,42,61,77,88,93,14,38,51	Avanzado	0.606061	1
15,22,32,44,52,62,70,87,92,72,86,6,55,79,94	Avanzado	0.606061	1
72,86,5,18,26,37,47,54,60,83,97,14,38,46,51	Avanzado	0.606061	1
9,25,59,90,72,86,5,18,26,37,47,54,60,83,97	Avanzado	0.606061	1
9,25,59,90,5,18,26,37,47,54,60,83,97,6,55,79,94	Bueno	0.606061	1
9,25,59,90,5,18,26,37,47,54,60,83,97,14,38,46,51	Bueno	0.606061	1
72,86,15,22,32,44,52,62,70,87,92,9,25,59	Avanzado	0.606061	1
9,25,59,90,5,18,26,37,47,54,60,83,97,14,38,46,51,6,55,79,94	Avanzado	0.0606061	1
9,25,59,90,72,86,5,18,26,37,47,54,60,83,97,6,55,79,94	Avanzado	0.0606061	1
9,25,59,90,16,31,42,61,77,88,93,0,14,38,51	Avanzado	0.0606061	1
9,25,59,90,17,66,16,31,42,61,77,88,93,14,38,51	Avanzado	0.606061	1
9,25,59,90,17,66,16,31,42,61,77,88,93,72	Avanzado	0.0606061	1
9,25,59,72,86,10,20,27,35,45,57,81,96,38,46,51	Bueno	0.0606061	1
9,25,59,90,72,86,14,38,46,51,6,55,79,94	Avanzado	0.0606061	1
9,25,59,90,16,31,42,61,77,88,93,72,14,38,51	Avanzado	0.0606061	1
9,25,59,90,17,66,16,31,42,61,77,88,93,0	Avanzado	0.606061	1
9,25,59,90,72,86,5,18,26,37,47,54,60,83,97,14,38,46,51	Avanzado	0.0606061	1
15,22,32,44,52,62,70,87,92,72,86,6,55,79,94,9,25,59	Avanzado	0.0606061	1
72,86,5,18,26,37,47,54,60,83,97,14,38,46,51,6,55,79,94	Avanzado	0.0606061	1
9,25,59,90,72,86,5,18,26,37,47,54,60,83,97,14,38,46,51,6,55,79,94	Avanzado	0.0606061	1
14,38,51,9,25,59,90,16,31,42,61,77,88,93,0,72	Avanzado	0.0606061	1
0,17,66,16,31,42,61,77,88,93,72,14,38,51	Avanzado	0.0606061	1
9,25,59,90,17,66,16,31,42,61,77,88,93,0,14,38,51	Avanzado	0.0606061	1
9,25,59,90,17,66,16,31,42,61,77,88,93,72,14,38,51	Avanzado	0.0606061	1
9,25,59,90,17,66,16,31,42,61,77,88,93,72,0	Avanzado	0.0606061	1
15,22,32,44,52,62,70,87,92,6,55,79,94,9,25,59	Avanzado	0.0606061	0.666667

Tabla 2.3: *Modelo de Reglas de Asociación.*

2.5. Conclusiones del capítulo

Con la aplicación de las tareas principales de la metodología Catalyst y la implementación del algoritmo Apriori, se logró establecer un modelo que cuenta con 295 reglas, de las cuales, 158 tienen una confianza igual a 1. Del total de reglas, 52 clasifican a la categoría de satisfactorio, 64 reglas, de bueno y 179, de avanzado. El proceso de extracción del conocimiento se realizó, después del perfilado de los datos, para reducir la dimensionalidad de las columnas.

Capítulo 3

Validación de la propuesta

3. VALIDACIÓN

EN el presente capítulo se valida el modelo de reglas encontrado anteriormente, utilizando la técnica de validaciones cruzadas y la métrica de precisión.

3.1. Introducción

La validación de un estudio es la cualidad que lo hace creíble y da testimonio del rigor con que se realizó. Esta implica la relevancia del estudio con respecto a sus objetivos, así como la coherencia lógica entre sus componentes [37].

Existen diversas técnicas para validar los modelos, como son la comparación de los parámetros con los obtenidos mediante modelos físicos teóricos o con simulaciones, utilizar nuevos conjuntos de datos conocidos para comparar con los obtenidos o el uso de técnicas de validación cruzada [38]. Las técnicas que vamos a analizar en este trabajo pertenecen a este último grupo.

3.2. Métodos de validación cruzada

La validación cruzada (cross-validation) es una técnica utilizada para evaluar los resultados obtenidos de un análisis estadístico. Garantizando que los modelos son independientes de la partición entre datos de entrenamiento y prueba que se utilicen garantizando así independencia estadística. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cuán preciso es un modelo que se llevará a cabo en la práctica. Es muy utilizada en proyectos de inteligencia artificial para validar modelos generados [38]. La validación cruzada tiene las aplicaciones siguientes:

- Validar la solidez de un modelo de minería de datos determinado.
- Evaluar varios modelos de una instrucción única.
- Generar varios modelos e identificar a continuación el mejor modelo basándose en estadísticas.

En particular la técnica de validación cruzada que se aplicará, va a permitir evaluar la solidez del modelo de minería de datos, mediante el uso de la métrica de precisión, que se calcula de la siguiente forma:

precisión= $\#deaciertos/\#objetosClasificados$

3.2.1. Selección de la partición

Un elemento determinante en la aplicación de la técnica es el número de clases. El número de clases determina la cantidad de procedimientos de validación que serán ejecutados. El procedimiento de validación se ejecuta n veces con $n = cantidad\_de\_clases$. En cada ejecución se toma $n - 1$ clases para construir el modelo y una clase para validar el modelo. Las n clases tienen que constituir una partición de equivalencia del total de datos disponibles en el dominio.
En este trabajo se decidió utilizar un valor de $n = 3$. Los datos, en este caso, deben cumplir con las características presentadas en la figura 3.1:

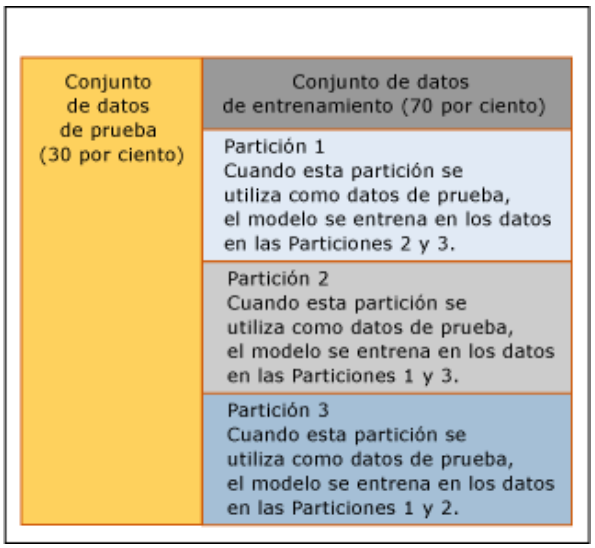
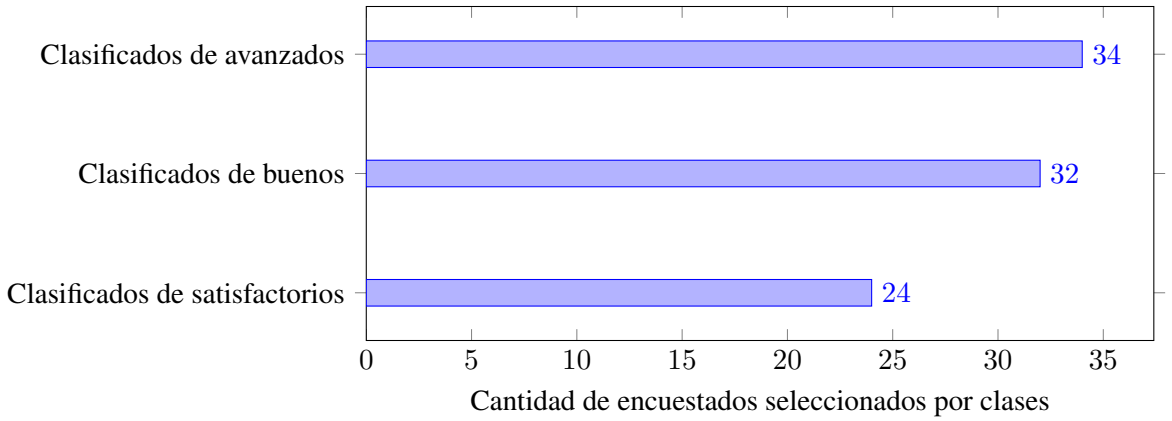


Figura 3.1: *Características de los datos para la validación cruzada con 3 clases.*

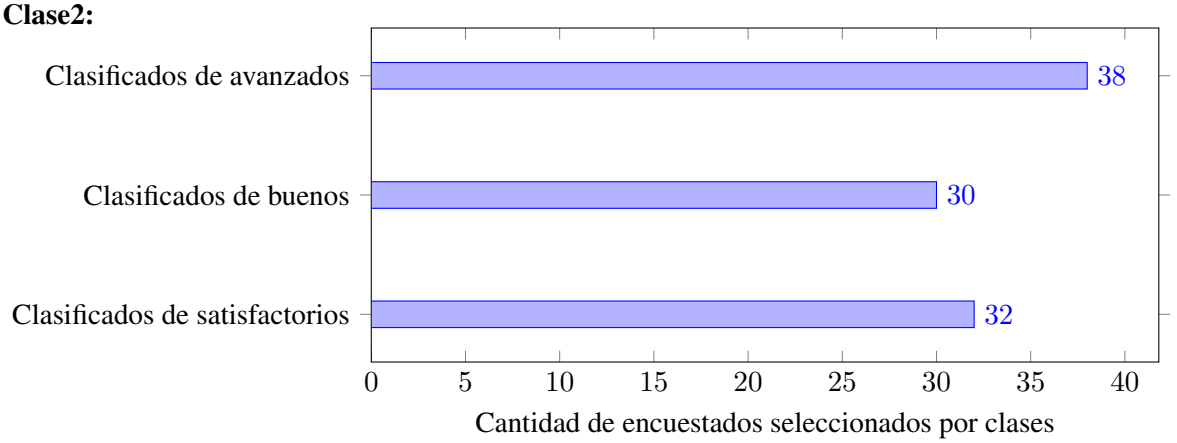
Como se observa en la figura 3.1 el proceso de validación debe de hacerse en tres experimentos, donde se seleccionan dos clases para construir el modelo y se utiliza una tercera clase para clasificar la población que se estudia, o sea la tercera clase se emplea como criterio de clasificación.

3.2.2. Selección los datos

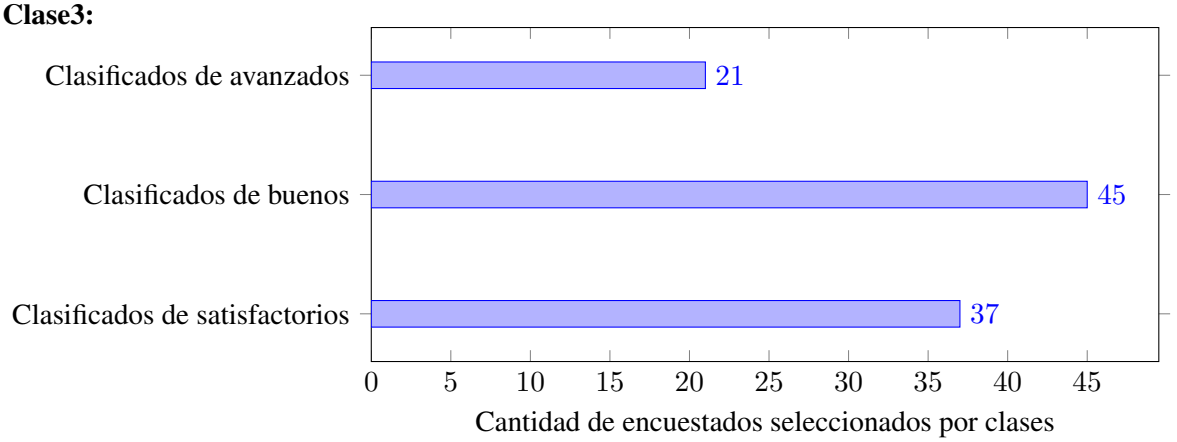
Se realiza una partición con tres clases para ejecutar las del experimento, usando en cada caso 2 clases para construir el modelo y una clase para clasificar, eligiendo de forma aleatoria 100 encuestados por cada clase. La cantidad de etiquetas de clasificación por cada clase quedó de la siguiente manera:
Clase1:



Clase2:



Clase3:



A continuación se describen cada uno de los pasos a seguir para aplicar la técnica de validación, teniendo en cuenta las particiones creadas y los datos seleccionados. Para aplicar dicha técnica se tiene como partida que se va a clasificar cada encuestado teniendo en cuenta el orden siguiente: la mayor confianza para cada regla, en caso que tenga algún empate en la evaluación, se tiene como criterio el soporte y para el caso de igual soporte se tiene como referencias las del conjunto de reglas que más se ajusta.

3.3. Aplicación del método de validación cruzada

Los experimentos se crean con dos pruebas, una construyendo el modelo con las mejores reglas por su soporte y confianza de cada clase y otra incluyendo todas las reglas generadas. Los resultados obtenidos por cada caso fueron:

1. Experimento 1: Se tomaron la clase 1 y la clase 2 para construir el modelo, dejando la clase 3 para clasificar.

   Modelo construido con todas las reglas encontradas.

   -El modelo tiene 192 reglas de asociación

   -De los 100 casos que se usaron para clasificar, se clasificaron correctamente 86 casos

   -Precisión$= 86/100$

   Modelo construido a partir de un conjunto de reglas limitados por cada clase(tabla 3.1).

Lista de preguntas	Clasificador	Soprte	Confianza
0,12,17,43	Satisfactorio	0.0909091	0.666667
9,59,0	Satisfactorio	0.0909091	0.666667
72,14,38,46,51,0	Satisfactorio	0.0909091	0.666667
72,0	Satisfactorio	0.0909091	0.666667
0,17,43	Satisfactorio	0.136364	0.6
3,28,40,0	Satisfactorio	0.136364	0.5
3,0	Satisfactorio	0.0909091	0.5
72,14,38,46,51	Satisfactorio	0.0909091	0.5
17,43,5,18,37,47,54,60,83	Satisfactorio	0.0909091	1
3,9,59	Satisfactorio	0.0909091	1
24,33,41,56,80,89,95,76,82	Bueno	0.0909091	1
24,33,41,56,80,89,95,72,86,76,82	Bueno	0.0909091	1
9,59,39,82	Bueno	0.0909091	1
39,82	Bueno	0.0909091	1
28,40	Bueno	0.0909091	1
16,31,34,42,49,61,68,77,88,93,28,40	Bueno	0.0909091	1
72,86,6,55,79	Bueno	0.0909091	1
6,55,79	Bueno	0.136364	0.75
14,38,46,51,9,25,59	Bueno	0.0909091	0.666667
15,32,44,52,70,87,92	Bueno	0.0909091	0.666667
6,55,79,94	Avanzado	0.136364	1
72,86,9,25,59,90	Avanzado	0.136364	1
6,55,79,94,72,86	Avanzado	0.136364	1
9,25,59,90	Avanzado	0.136364	1
14,38,46	Avanzado	0.0909091	1
14,38,51,9,25,59,90	Avanzado	0.136364	1
6,55,79,94,3,28,40	Avanzado	0.0909091	1
6,55,79,94,15,22,32,44,52,62,70,87,92	Avanzado	0.0909091	1
10,27,35,45,57,81,96	Avanzado	0.0909091	1
9,25,59,90,3,28,40	Avanzado	0.0909091	1

Tabla 3.1: *Modelo construido a partir de un conjunto de reglas limitados por cada clase para el experimento 1*

-El modelo tiene 30 reglas de asociación

-De los 100 casos que se usaron para clasificar, se clasificaron correctamente 84 casos

-Precisión= $84/100$

2. Experimento 2 : Se tomaron la clase 1 y la clase 3 para construir el modelo, dejando la clase 2

para clasificar. El modelo construido arrojó como resultado:

Modelo construido con todas las reglas encontradas.

-El modelo tiene 203 reglas de asociación

-De los 100 casos que se usaron para clasificar, se clasificaron correctamente 90 casos

-Precisión= $90/100$

Modelo construido a partir de un conjunto de reglas limitados por cada clase (tabla 3.2).

Lista de preguntas	Clasificador	Soprte	Confianza
39,82,72,86	Satisfactorio	0.0909091	0.5
39,82	Satisfactorio	0.0909091	0.4
17,43	Satisfactorio	0.0909091	0.285714
72,86	Satisfactorio	0.136364	0.25
3,28,40	Satisfactorio	0.0909091	0.222222
0,12,17,43	Satisfactorio	0.0909091	0.666667
5,18,37,47,54,60,83	Satisfactorio	0.0909091	1
72,14,38,46,51,0	Satisfactorio	0.0909091	0.666667
3,40,17,43	Bueno	0.0909091	1
39,82,9,59	Bueno	0.0909091	0.666667
14,38,46,51,9,25,59	Bueno	0.0909091	0.666667
3,40	Bueno	0.0909091	0.666667
9,59	Bueno	0.0909091	0.666667
39,82	Bueno	0.136364	0.6
24,33,41,56,80,89,95,72,86,76,82	Bueno	0.0909091	1
9,59,39,82	Bueno	0.0909091	1
3,28,40,9,25,59	Bueno	0.0909091	0.5
15,22,32,44,52,62,70,87,92	Bueno	0.0909091	0.5
39,82,72,86	Bueno	0.0909091	0.5
9,25,59,90	Avanzado	0.272727	1
9,25,59,90,72,86	Avanzado	0.181818	1
9,25,59,90,0	Avanzado	0.181818	1
14,38,51	Avanzado	0.181818	1
14,38,46,51,72,86	Avanzado	0.136364	1
14,38,51,72	Avanzado	0.136364	1
14,38,51,9,25,59,90	Avanzado	0.0909091	1
17,66,9,25,59,90	Avanzado	0.0909091	1
3,28,40,9,25,59,90	Avanzado	0.136364	1
14,38,51,9,25,59,90	Avanzado	0.136364	1
14,38,46,51,6,55,79,94,72,86	Avanzado	0.136364	1

Tabla 3.2: *Modelo construido a partir de un conjunto de reglas limitados por cada clase para el experimento 2*

-El modelo tiene 30 reglas de asociación

-De los 100 casos que se usaron para clasificar, se clasificaron correctamente 89 casos

-Precisión= $89/100$

3. Experimento 3 : Se tomaron la clase 3 y la clase 2 para construir el modelo, dejando la clase 1

para clasificar. El modelo construido arrojó como resultado:

Modelo construido con todas las reglas encontradas.

-Un modelo con 228 reglas de asociación

-De los 100 casos que se usaron para clasificar, se clasificaron correctamente 93 casos

-Precisión= $93/100$

Modelo construido a partir de un conjunto de reglas limitados por cada clase(tabla 3.3).

Lista de preguntas	Clasificador	Soprte	Confianza
17,43,3,28,40	Satisfactorio	0.136364	1
17,43,24,33,80,95	Satisfactorio	0.0909091	1
24,33,80,95,3,28,40	Satisfactorio	0.0909091	1
55,5,18,37,47,54,60,83	Satisfactorio	0.0909091	1
17,43,55	Satisfactorio	0.0909091	1
5,18,37,47,54,60,83	Satisfactorio	0.0909091	1
24,33,80,95	Satisfactorio	0.0909091	1
17,43,9,59	Satisfactorio	0.0909091	1
17,43,5,18,37,47,54,60,83	Satisfactorio	0.0909091	1
3,9,59	Satisfactorio	0.0909091	1
24,33,41,56,80,89,95	Bueno	0.136364	1
24,33,41,56,80,89,95,72,86	Bueno	0.0909091	1
24,33,41,56,80,89,95,76,82	Bueno	0.0909091	1
17,43,9,25,59	Bueno	0.0909091	1
10,20,27,35,45,57,81,96,72,86	Bueno	0.0909091	1
17,72	Bueno	0.0909091	1
10,20,27,35,45,57,81,96,9,25,59	Bueno	0.0909091	1
76,82	Bueno	0.0909091	1
10,20,27,35,45,57,81,96,38,46,51	Bueno	0.0909091	1
10,20,27,35,45,57,81,96	Bueno	0.0909091	1
9,25,59,90	Avanzado	0.136364	1
0,9,25,59,90	Avanzado	0.136364	1
72,17,66	Avanzado	0.0909091	1
15,32,44,52,70,87,92,9,25,59,90	Avanzado	0.0909091	1
16,31,42,61,77,88,93,14,38,51	Avanzado	0.0909091	1
16,31,42,61,77,88,93	Avanzado	0.0909091	1
14,38,51,9,25,59,90	Avanzado	0.0909091	1
17,66,9,25,59,90	Avanzado	0.0909091	1
14,38,51,17,66	Avanzado	0.0909091	1
17,66	Avanzado	0.0909091	1

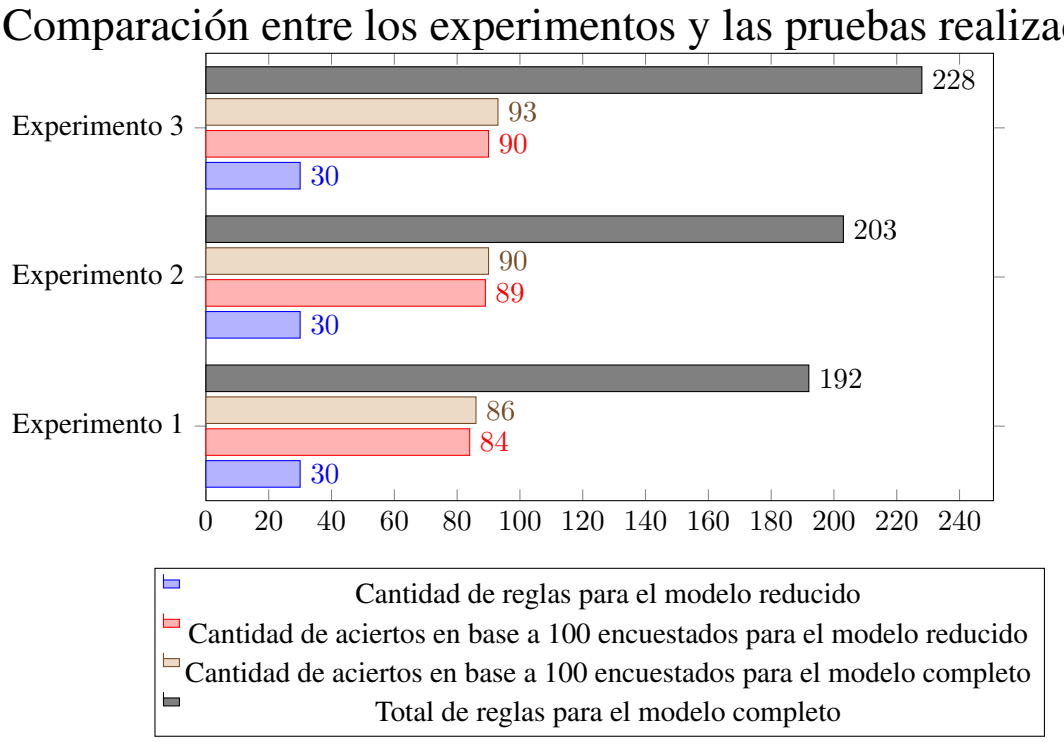
Tabla 3.3: *Modelo construido a partir de un conjunto de reglas limitados por cada clase para el experimento 3*

-El modelo tiene 30 reglas de asociación

-De los 100 casos que se usaron para clasificar, se clasificaron correctamente 90 casos

-Precisión= $90/100$

Las pruebas realizadas arrojaron como resultado final que, el promedio de precisión a partir del proceso de extracción de reglas de asociación, fue de un 89.33 % para el modelo completo y un 88.66 % para el modelo con la selección de reglas. A continuación, se muestra el comportamiento de la variable precisión para cada tipo de modelo, los experimentos realizados para todas las reglas y para un conjunto de reglas:



3.4. Discusión de los resultados

Al aplicar el proceso para extracción reglas de asociación al test de orientación vocacional CHASIDE con relación al desempeño académico, que se le aplicó a profesionales y estudiantes de la carrera se obtiene un modelo de reglas para cada combinación de los umbrales de soporte y confianza. A partir del total de reglas generadas se puede determinar qué combinaciones de intereses y aptitudes debe tener un estudiante con relación a su desempeño. El modelo planteado logra establecer un conjunto de 295 reglas las cuales describen características principales del ingeniero en Ciencias Informáticas. Logrando así correlacionar las aptitudes, habilidades y competencias del estudiante con el desarrollo académico en la carrera. El 56 % de las reglas encontradas tienen una confianza de 1 y el modelo construido a partir de la validación cruzada arrojó que para los 3 experimentos la precisión fue mayor que el 84 %, llegando a alcanzar un promedio de 89,6 %.

Tras la validación, se llega a la conclusión que el modelo se puede reducir a 30 reglas ya que las

pruebas con los modelos reducidos arrojan un porciento de precisión muy similar a los encontrados con el modelo completo. Adoptándose el modelo de 30 reglas como alternativo para definir el conjunto de reglas presentes que describe las aptitudes e intereses presente en el ingeniero en Ciencias Informáticas según el desempeño académico.

Un ejemplo de las reglas de asociación encontradas es el estudiante que tiene las aptitudes e intereses: 15,32,44,52,70,87,92,9,25,59,90→Avanzado con un soporte de 0.909091 y una confianza de 1. La combinación de dichas aptitudes e intereses son las características que debe tener un estudiante para ser avanzado en la carrera de Ingeniería en Ciencias Informáticas con una confianza de 1 y un soporte de 0,91 %.

Leyenda:

Las siguientes preguntas pertenecen a la categoría Medicina y Ciencias de la Salud dentro de los intereses de los estudiantes:

15-¿Convences fácilmente a otras personas sobre la validez de tus argumentos?

32-¿Participarías en una campaña de prevención de la enfermedad del Dengue?

44-¿Te gustaría hacer un curso de primeros auxilios?

52-¿Te resultaría interesante el estudio de las ciencias biológicas?

70-¿Te gustaría investigar sobre una nueva vacuna?

87-Ante una emergencia epidémica ¿Participarías en una campaña brindando tu ayuda?

92-Ante un llamado solidario ¿Te ofrecerías para cuidar un enfermo?

Las siguientes preguntas pertenecen a la categoría Ingeniería y Computación dentro de las aptitudes de los estudiantes:

9-¿Eres exigente y crítico con tu equipo de trabajo?

25- ¿Planificas detalladamente tus trabajos antes de empezar?

59-¿Crees que tus ideas son importantes, y haces todo lo posible para ponerlas en práctica?

90-¿Trabajar con objetos te resulta más gratificante que trabajar con personas?

3.5. Conclusiones del capítulo

Las pruebas que se llevaron a cabo con la técnica de validación cruzada, en las que se midió el grado de solidez del modelo planteado a través de la métrica de precisión arrojó un promedio de aciertos de un 89 %. Tras realizar las mismas pruebas, pero con un modelo más reducido de reglas, se puede afirmar que los segundos modelos planteados, con 30 reglas, se ajustan a los resultados del modelo completo, ya que no es muy significativo el número de aciertos en la clasificación. Se estableció que se puede reducir el modelo de reglas, para una mejor interpretación de los resultados.

CONCLUSIONES

Como resultado de la presente investigación se obtuvo un modelo para extraer reglas de asociación a partir del test de orientación vocacional CHASIDE correlacionando las habilidades, intereses y competencias del estudiante de la carrera de Ingeniería en Ciencias Informáticas con el desempeño académico. El modelo responde a la necesidad de la universidad de mejorar los resultados docentes alcanzados, especialmente en la disminución del fracaso escolar. En base a los resultados obtenidos se arribó a las siguientes conclusiones:

1. Se estableció un marco conceptual de referencia para soportar los fundamentos teóricos de la investigación y dominar cada aspecto de la propuesta, lo que permitió definir los software de predominio de cómputo y sus elementos como parte de los software de inteligencia artificial. Se introduce la minería de datos como proceso para extraer conocimiento del test CHASIDE en forma de reglas de asociación, que también se utilizan para expresar el conocimiento previo. Esto se respalda mediante el uso de la metodología la cual describe detalladamente los pasos a realizar y se ajustan las tareas a las características del problema.
2. Se implementó una aplicación para la captura de los datos del test CHASIDE, lo que permitió su posterior procesamiento.
3. Se determinaron las técnicas de exploración de datos a utilizar mediante la metodología Catalyst utilizando la tarea descriptiva y la técnica de reglas de asociación. Logrando obtener un modelo de reglas de asociación el cual permite identificar las habilidades, intereses y competencias según el comportamiento académico del estudiante en Ciencias Informáticas.
4. Se implementa el algoritmo Apriori, una vez perfilados los datos para procesar la vista minable y extraer el modelo de reglas de asociación, el cual permitió establecer las aptitudes e intereses de los estudiantes en la carrera de Ingeniería en Ciencias Informáticas.
5. La realización de pruebas de validación cruzada evidenció la precisión del modelo propuesto, mostrando que es preciso en un 89 %. Además, el modelo propuesto facilita la clasificación en tres clases (Rendimiento normal, Buen rendimiento y Alto rendimiento) de un encuestado

teniendo en cuenta las preguntas marcadas.

RECOMENDACIONES

La investigación realizada no cubre todos los aspectos posibles del análisis en las demás carreras existentes en el país en el contexto universitario. Por este motivo se recomienda lo siguiente:

1. Incluir a la herramienta desarrollada del test CHASIDE a la web universitaria y pre-universitaria para explotar sus potencialidades.
2. Definir otros patrones que no sean los resultados académicos dentro del contexto universitario que resulten de interés institucional para explorar otras correlaciones.
3. Introducir en los análisis de clasificación otros elementos que no sean el soporte y la confianza.
4. Extender el modelo a otras carreras que estén dentro de las clasificaciones del test CHASIDE.

REFERENCIAS BIBLIOGRÁFICAS

[1] Di Doménico, C. & Vilanova, A. Orientación vocacional: origen, evolución y estado actual. *Orientación y sociedad*, vol. 2, pp. 47–58, 2000. Consultado el: 2017-06-08.

[2] De Fruyt, F. & Mervielde, I. The five-factor model of personality and Holland's RIASEC interest types. vol. 23, *Personality and Individual Differences*, no. 1, pp. 87–103, Julio 1997. Consultado el: 2016-11-10, doi:10.1016/S0191-8869(97)00004-4.

[3] Maura, V. G. El servicio de orientación vocacional-profesional (SOVP) de la Universidad de La Habana: una estrategia educativa para la elección y desarrollo profesional responsable del estudiante. vol. 6, *Pedagogía Universitaria*, no. 4, 2013. Consultado el: 2016-11-09.

[4] Forward, A. & Lethbridge, T. C. A taxonomy of software types to facilitate search and evidence-based software engineering. En *Proceedings of the 2008 conference of the center for advanced studies on collaborative research: meeting of minds*, pp. 14. ACM, 2008. Consultado el: 2016-11-18.

[5] Shacham, M. & Brauner, N. Application of stepwise regression for dynamic parameter estimation. *Computers & Chemical Engineering*, vol. 69, pp. 26–38, 2014.

[6] Hong, X. & Chen, S. Elastic net orthogonal forward regression. *Neurocomputing*, vol. 148, pp. 551–560, 2015.

[7] Uriel, E. Regresión lineal múltiple: estimación y propiedades. *Universidad de Valencia Versión*, pp. 09–2013, 2013. Consultado el: 2017-06-11.

[8] Billionnet, C.; Sherrill, D.; Annesi-Maesano, I. et al. Estimating the health effects of exposure to multi-pollutant mixture. vol. 22, *Annals of epidemiology*, no. 2, pp. 126–141, 2012.

[9] Berhane, K.; Chang, C.-C.; McConnell, R.; Gauderman, W. J.; Avol, E.; Rapapport, E.; Urman, R.; Lurmann, F. & Gilliland, F. Association of changes in air quality with bronchitic symptoms in children in California, 1993-2012. vol. 315, *Jama*, no. 14, pp. 1491–1501, 2016.

[10] Sepúlveda, J. F. D. & Correa, J. C. Comparación entre árboles de regresión CART y regresión lineal. vol. 6, *Comunicaciones en Estadística*, no. 2, pp. 175–195, 2013. Consultado el: 2017-06-11.

[11] Jie, Y.; Houjin, H.; Xun, M.; Kebin, L.; Xuesong, Y. & Jie, X. Relationship between pulmonary function and indoor air pollution from coal combustion among adult residents in an inner-city area of southwest China. vol. 47, *Brazilian Journal of Medical and Biological Research*, no. 11, pp. 982–989, 2014.

[12] Toti, G.; Vilalta, R.; Lindner, P.; Lefer, B.; Macias, C. & Price, D. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. *Artificial Intelligence in Medicine*, vol. 74, pp. 44–52, 2016.

[13] Sommerville, I. *Software engineering*. Pearson, 2016.

[14] Siegmund, J.; Siegmund, N. & Apel, S. How reviewers think about internal and external validity in empirical software engineering. En *Software Engineering*, pp. 83–84, 2016.

[15] Forward, A. & Lethbridge, T. C. A taxonomy of software types to facilitate search and evidence-based software engineering. En *Proceedings of the 2008 conference of the center for advanced studies on collaborative research: meeting of minds*, pp. 14. ACM, 2008.

[16] Šmite, D.; Wohlin, C.; Galviņa, Z. & Prikladnicki, R. An empirically based terminology and taxonomy for global software engineering. vol. 19, *Empirical Software Engineering*, no. 1, pp. 105–153, 2014. Consultado el: 2016-11-04.

[17] Antony, P.; Manujesh, P. & Jnanesh, N. Data mining and machine learning approaches on engineering materials—A review. En *Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference on*, pp. 69–73. IEEE, 2016.

[18] Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C. & Wirth, R. CRISP-DM 1.0 Step-by-step data mining guide. 2000.

[19] Pyle, D. *Business modeling and data mining*. Morgan Kaufmann, 2003.

[20] Mariscal, G.; Marbán, Ó. & Fernández, C. A survey of data mining and knowledge discovery process models and methodologies. vol. 25, *The Knowledge Engineering Review*, no. 02, pp. 137–166, 2010.

[21] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. et al. Knowledge Discovery and Data Mining: Towards a Unifying Framework. En *KDD*, vol. 96, pp. 82–88, 1996.

[22] Moine, J. M. *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo*. Tesis de doctorado, Facultad de Informática, 2013. Consultado el: 2017-02-16.

[23] Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. The KDD process for extracting useful knowledge from volumes of data. vol. 39, *Communications of the ACM*, no. 11, pp. 27–34, 1996. Consultado el: 2017-02-16.

[24] Moyle, S. & Jorge, A. RAMSYS-A methodology for supporting rapid remote collaborative data mining projects. En *Proc. ECML/PKDD Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pp. 20–31, 2001. Consultado el: 2017-06-11.

[25] Institute, S. SEMMA Data Mining Methodology. Technical report, 2005.

[26] Cios, K. J. & Kurgan, L. A. Trends in data mining and knowledge discovery. En *Advanced techniques in knowledge discovery and data mining*, pp. 1–26. Springer, 2005.

[27] Martínez de Pisón Ascacíbar, F. J. Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado. 2003.

[28] Printed in the United States of America. Text Mining for Clementine 12.0 User's Guide.

[29] Cerrito, P. B. *Introduction to data mining using SAS Enterprise Miner*. SAS Publishing, 2006. Consultado el: 2017-05-24.

[30] Pereira, R. T. Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos. vol. 8, *Revista Guillermo de Ockham*, no. 1, 2010. Consultado el: 2017-05-24.

[31] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. The WEKA data mining software: an update. vol. 11, *ACM SIGKDD explorations newsletter*, no. 1, pp. 10–18, 2009. Consultado el: 2017-05-24.

[32] Rodrıguez, D. M. Algoritmos Evolutivos para la extracción de Reglas de Asociación Cuantitativas. 2014. Consultado el: 2017-06-11.

[33] Zaki, M. J.; Meira Jr, W. & Meira, W. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014. Consultado el: 2017-06-11.

[34] Borgelt, C. An Implementation of the FP-growth Algorithm. En *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pp. 1–5. ACM, 2005. Consultado el: 2017-06-11.

[35] Agrawal, R.; Srikant, R. & others. Fast algorithms for mining association rules. En *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, pp. 487–499, 1994. Consultado el: 2017-05-09.

[36] Srikant, R. & Agrawal, R. Mining quantitative association rules in large relational tables. En *Acm Sigmod Record*, vol. 25, pp. 1–12. ACM, 1996. Consultado el: 2017-06-11.

[37] Yacuzzi, E. El estudio de caso como metodología de investigación: teoría, mecanismos causales, validación. Technical report, Serie Documentos de Trabajo, Universidad del CEMA: Área: negocios, 2005. Consultado el: 2017-06-02.

[38] Kozak, A. & Kozak, R. Does cross validation provide additional information in the evaluation of regression models? vol. 33, *Canadian Journal of Forest Research*, no. 6, pp. 976–987, 2003. Consultado el: 2017-06-02.

A. Evaluación del test CHASIDE

Funcionamiento del Test

Detallamos a continuación el funcionamiento de este test.

1.- Por cada pregunta contestada afirmativamente se marcará con una cruz el número correspondiente a la actividad seleccionada.

2.- Cada número marcado vale un punto. Súmalos verticalmente y coloca el resultado en los casilleros vacíos debajo de cada columna.

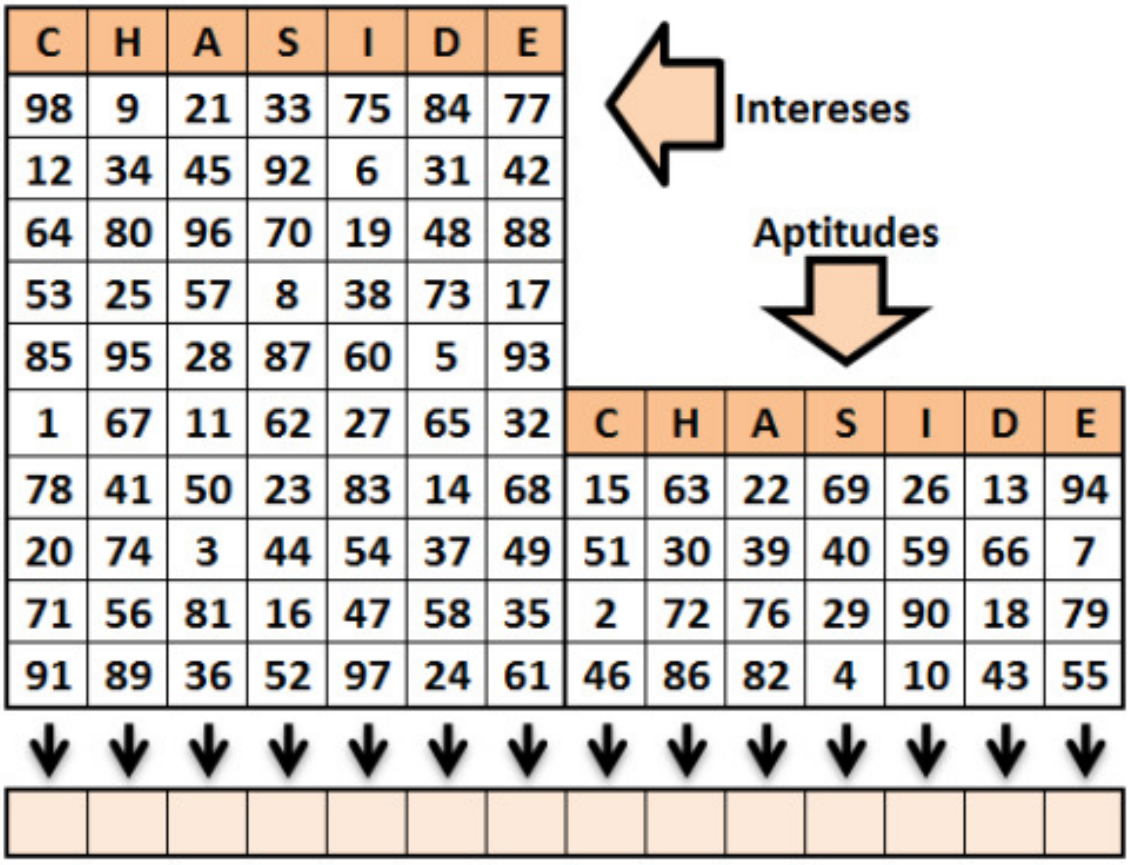


Figura A.1: *Matriz de evaluación del test CHASIDE*

Fuente: http://esocreo.wordpress.com/2009/03/27/test-de-orientacion-vocacional/

El test CHASIDE evalúa las áreas:

1. Administrativas y Contables (C)
2. Humanísticas y Sociales (H)
3. Artísticas (A)
4. Medicina y Ciencias de la Salud (S)
5. Ingeniería y Computación (I)
6. Defensa y Seguridad (D)
7. Ciencias Exactas y Agrarias (E)

Preguntas del test CHASIDE:

1. ¿Aceptarías trabajar escribiendo artículos en la sección económica de un diario?
2. ¿Te gustaría dirigir un proyecto de urbanización en tu provincia?
3. ¿A una frustración siempre opones un pensamiento positivo?
4. ¿Te dedicarías a socorrer a personas accidentadas o atacadas por asaltantes?
5. Cuándo eras chico ¿te interesaba saber como estaban construidos tus juguetes?
6. ¿Te interesan más los misterios de la naturaleza que los secretos de la tecnología?
7. ¿Escuchas atentamente los problemas que te plantean tus amigos?
8. ¿Te ofrecerías para explicarle a tus compañeros un tema que ellos no entendieron?
9. ¿Eres exigente y crítico con tu equipo de trabajo?
10. ¿Te atrae armar rompecabezas?
11. ¿Sabes la diferencia conceptual entre macroeconomía y microeconomía?
12. Usar uniforme, ¿Te hace sentir distinto, importante?
13. ¿Participarías como profesional en un espectáculo de acrobacia aérea?
14. ¿Organizas tu dinero de modo que te alcance para lo que debes hacer con él?
15. ¿Convences fácilmente a otras personas sobre la validez de tus argumentos?
16. ¿Estás informado sobre los últimos descubrimientos científicos?
17. Ante una situación de emergencia, ¿Actuas rápidamente?
18. Cuando tenés que resolver un problema matemático ¿Perseveras hasta encontrar la solución?
19. Si te convocara tu club preferido para planificar, organizar y dirigir un campo de deportes ¿Aceptarías?
20. ¿Eres el que pone un toque de alegría en las fiestas?
21. ¿Crees que los detalles son tan importantes como el todo?

22. ¿Te sentirías a gusto trabajando en un ámbito hospitalario?
23. ¿Te gustaría participar para mantener el orden en grandes desordenes o catástrofes?
24. ¿Pasarías varias horas leyendo un libro de tu interés?
25. ¿Planificas detalladamente tus trabajos antes de empezar?
26. ¿Entablas una relación casi personal con tu computadora?
27. ¿Disfrutarías modelar arcilla?
28. ¿Ayudas a no videntes habitualmente a cruzar la calle?
29. ¿Consideras importante que desde la escuela primaria se fomente la actitud crítica y la participación activa?
30. ¿Te parece bien que las mujeres formen parte de las fuerzas armadas bajo las mismas condiciones que los hombres?
31. ¿Te gustaría crear nuevas técnicas para descubrir las patologías de algunas enfermedades a través del microscopio?
32. ¿Participarías en una campaña de prevención de la enfermedad del Dengue?
33. ¿Te interesan los temas relacionados al pasado y el progreso de la humanidad?
34. ¿Te incluirías en un proyecto de investigación de los movimientos sísmicos y sus consecuencias?
35. Fuera de los horarios escolares ¿dedicas algún tiempo a actividades corporales?
36. ¿Te interesan las actividades de mucha acción y de reacción rápida en situaciones imprevistas de peligro?
37. ¿Te ofrecerías para colaborar como voluntarios en los gabinetes especiales de la NASA?
38. ¿Te ofrecerías para organizar la despedida de soltero de uno de tus amigos?
39. ¿Te gusta más el trabajo manual que el trabajo intelectual?
40. ¿Estarías dispuesto a renunciar a un momento placentero por ofrecer tu servicio como profesional?
41. ¿Participarías en una investigación sobre la violencia en el fútbol?
42. ¿Te gustaría trabajar en un laboratorio mientras estudias?
43. ¿Arriesgarías tu vida para salvar la vida de otra persona que no conoces?
44. ¿Te gustaría hacer un curso de primeros auxilios?
45. ¿Tolerarías empezar tantas veces como fuere necesario hasta obtener el logro deseado?
46. ¿Distribuís tu horario adecuadamente para poder hacer todo lo planeado?
47. ¿Harías un curso para aprender a fabricar los instrumentos y/o piezas de las máquinas o aparatos que utilizas?

48. ¿Elegirías una profesión en la que tuvieras que estar algunos meses alejados de tu familia, por ejemplo marino?
49. ¿Te radicarías en una zona agrícola-ganadera para desarrollar tus actividades como profesional?
50. Cuando estas en un grupo de trabajo, ¿Te agrada producir ideas originales y que sean tenidas en cuenta?
51. ¿Te resulta fácil coordinar un grupo de trabajo?
52. ¿Te resultaría interesante el estudio de las ciencias biológicas?
53. Si una empresa solicita un profesional como gerente de comercialización, ¿Te sentirías a gusto desempeñando ese rol?
54. ¿Te incluirías en un proyecto nacional de desarrollo de la principal fuente de recursos de tu provincia?
55. ¿Tienes interés por saber cuáles son las causas que determinan ciertos fenómenos, aunque saberlo no incida en tu vida?
56. ¿Descubrirte algún filósofo o escritor que haya expresado tus mismas ideas con antelación?
57. ¿Desearías que te regalen algún instrumento musical para tu cumpleaños?
58. ¿Aceptarías colaborar con el cumplimiento de las normas en lugares públicos?
59. ¿Crees que tus ideas son importantes, y haces todo lo posible para ponerlas en práctica?
60. Cuando se descompone un artefacto en tu casa ¿Te dispones prontamente a repararlo?
61. ¿Formarías parte de un equipo de trabajo orientado a la preservación de la flora y la fauna en extinción?
62. ¿Acostumbras a leer revistas relacionadas con los últimos avances científicos y tecnológicos en el área de salud?
63. ¿Preservar las raíces culturales de nuestro país te parece importante y necesario?
64. ¿Te gustaría realizar una investigación que contribuyera a hacer más justa la distribución de la riqueza?
65. ¿Te gustaría realizar tareas auxiliares en una nave, como izado y arriado de velas, pintura y conservación de cascos, arreglo de desperfectos, de motores, etc?
66. ¿Crees que el país debe poseer la más alta tecnología armamentista, a cualquier precio?
67. La libertad y la justicia ¿Son valores importantes en tu vida?
68. ¿Aceptarías hacer una práctica rentada en industria de productos alimenticios, en el sector de control de calidad?
69. ¿Consideras que la salud pública debe ser prioritaria, gratuita y eficiente para todos?
70. ¿Te gustaría investigar sobre una nueva vacuna?

71. En un equipo de trabajo ¿Preferís el rol de coordinador?
72. En una discusión entre amigos ¿Te ofreces como mediador?
73. ¿Estás de acuerdo con la formación de un cuerpo de soldados profesionales?
74. ¿Lucharías por una causa justa hasta las últimas consecuencias?
75. ¿Te gustaría investigar científicamente sobre cultivos agrícolas?
76. ¿Harías un nuevo diseño de una prenda pasada de moda, ante una reunión imprevista?
77. ¿Visitarías un observatorio astronómico para conocer en acción el funcionamiento de los aparatos?
78. ¿Dirigirías el área de importación y exportación de una empresa?
79. ¿Te inhibís al entrar a un lugar nuevo con gente desconocida?
80. ¿Te gratificaría trabajar con niños?
81. ¿Harías el afiche para una campaña contra el SIDA?
82. ¿Dirigirías un grupo de teatro independiente?
83. ¿Enviarías tu currículum a una empresa automotriz que solicita gente para su área de producción?
84. ¿Participarías en un grupo de defensa internacional dentro de alguna fuerza armada?
85. ¿Te costearías tus estudios trabajando en una auditoría?
86. ¿Eres de los que defienden causas perdidas?
87. Ante una emergencia epidémica ¿Participarías en una campaña brindando tu ayuda?
88. ¿Saber responder que significa ADN y ARN?
89. ¿Elegirías una carrera cuyo instrumento de trabajo fuere la utilización de un idioma extranjero?
90. ¿Trabajar con objetos te resulta más gratificante que trabajar con personas?
91. ¿Te resultaría gratificante ser asesor contable en una empresa reconocida?
92. Ante un llamado solidario ¿Te ofrecerías para cuidar un enfermo?
93. ¿Te atrae investigar sobre los misterios del universo, por ejemplo, agujeros negros?
94. El trabajo individual ¿Te resulta más rápido y efectivo que el trabajo grupal?
95. ¿Dedicarías parte de tu tiempo para ayudar a personas de zonas cadenciadas?
96. Cuando elegís tu ropa o decoras un ambiente ¿Tienes en cuenta la combinación de los colores, las telas o el estilo de los muebles?
97. ¿Te gustaría trabajar como profesional dirigiendo la construcción de una empresa hidroeléctrica?
98. ¿Sabes lo que es el PBI?