UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

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

Obtención de patrones mediante generación de reglas a partir de los datos almacenados del proceso de Diagnóstico a las Organizaciones Productivas de la UCI

Autor:
Laynier Antonio Piedra Diéguez

Tutores:

Msc. Maidel Beatriz Ginarte Durán
Ing. Norge Sánchez Tumbarell

Marzo de 2013

Año 55 de la Revolución

DECLARACIÓN DE AUTORÍA

Yo Laynier Antonio Piedra Diéguez, declaro ser el único autor del trabajo titulado: *Obtención de patrones mediante generación de reglas a partir de los datos almacenados del proceso de Diagnóstico a las Organizaciones Productivas de la UCI* y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma con carácter exclusivo.

Para que así conste firmo la presente a los ___ días del mes de _____ del año ____.

_____
Laynier Piedra Diéguez
(Autor)

_____ _____
Msc. Maidel Beatriz Ginarte Durán Ing. Norge Sánchez Tumbarell
(Tutor) (Tutor)

Síntesis del Tutor:

Maidel B Ginarte Durán. Graduada de Ingeniería en Ciencias Informáticas en 2007. Máster en calidad de software. Profesor Asistente. Especialista de calidad de software. Auditora de la bolsa de la Oficina Nacional de Normalización.Trabaja en el centro Calisoft desde 2007. Ha participado en la totalidad de los Diagnósticos a las organizaciones productivas de la UCI. Ha impartido las asignaturas de Teleinformática, Máquinas Computadoras, Programación Web y Práctica Profesional.

Norges Sánchez Tumbarell. Graduado de Ingeniería en Ciencias Informáticas en 2007. Profesor Instructor. Jefe de Proyecto de Pregrado del centro Cenia. Trabaja en el centro CENIA desde el 2007. Tiene varios años de experiencia como Arquitecto de Bases de Datos. Ha impartido la asignatura de Matemáticas y cursos optativos de Bases de Datos.

Dedicatoria

A mi familia.

A todos los que iniciaron este sueño en 2002 y aún queriéndolo, no pudieron continuar, a sus padres como a los mios. Yo lo estoy terminando por todos ellos.

Agradecimientos

Agradezco a mi esposa, por su amor, su espera, su ayuda idónea y oportuna, y su tutoría en este trabajo.

Agradezco a mis padres por la paciencia y la impaciencia que me han tenido. Agradezco a todos los que sabían de mi esfuerzo y me dieron ánimo y apoyo.

Agradezco a mi niño por ser el más bueno e inteligente del mundo, de tal forma que me permitió dedicarme a esto.

Agradezco a Lucmary porque unas veces me sostuvo y otras luchó por mí hasta que logró incluirme aquí de nuevo. Agradezco a Alier por corregirme y aleccionarme siempre, a Marzo por darme el tiempo y el espacio sin pedirme nada a cambio, a Espinosa por ayudarme sin tener la obligación de hacerlo. A mis profesores de P4 y Física. A los que estudiaron conmigo: Enrique, Rayner, Nilmar, Danny y Yuneisy.

Agradezco a todo el que me ayudó en algo, por pequeño que hay sido, fue importante, y a los que me ayudaron sin yo saberlo es más especial mi agradecimiento.

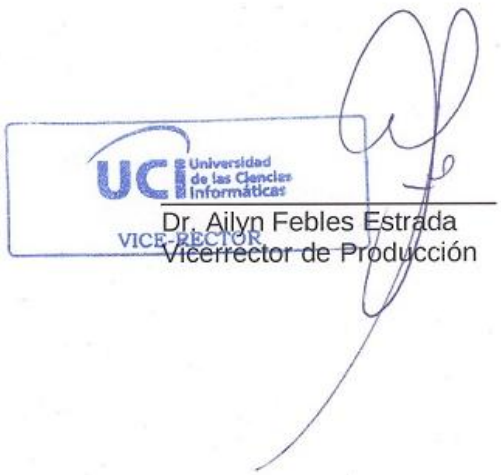
Por último pero fundamentalmente, agradezco a Dios por todo y todos hasta este momento y en lo adelante para siempre.

Aval de la tesis referida a la aplicación de técnicas de minería de datos a los datos recogidos del diagnóstico

Universidad de las Ciencias Informáticas (UCI), La Habana, 21/02/2013

La tesis del estudiante del Curso por Encuentros (CPE) Laynier Antonio Piedra Diéguez, aborda la obtención de patrones a través de la aplicación de técnicas de minería de datos sobre los datos obtenidos del proceso de diagnóstico. El documento constituye un material de consulta para investigaciones futuras. Estos datos han sido recogidos por personal calificado para la tarea y son por tanto confiables y útiles para el análisis que se realizó con ellos.

Teniendo en cuenta los resultados obtenidos durante su desarrollo, considero que la tesis contiene un alto valor científico. Estos resultados del proceso de minería de datos aportan nuevos elementos a la información que se tenía sobre el desarrollo productivo de la UCI. Son informaciones capaces de generar más conocimiento y apoyar la toma de decisiones a nivel gerencial.



Dr. Ailyn Febles Estrada
Vicerrector de Producción

Resumen:
La minería de datos se encarga de encontrar o predecir patrones y tendencias a través de información oculta en grandes volúmenes de datos, los cuales no pueden ser analizados por los métodos tradicionales. El diagnóstico que se realiza anualmente a las organizaciones productivas en la Universidad de las Ciencias Informáticases un proceso generador de datos que aún no ha sido tratado mediante técnicas de minería. En el presente trabajo se aplicaránalgunas de estas técnicas, específicamente las relacionadas a la generación de reglas a partir de los datos almacenados mediante el proceso de diagnóstico.

Palabras claves:
Minería, datos, diagnóstico.

Abstract:

Data mining is about finding or predicting patterns and tendencies trough information that is hidden in big amounts of data, which can not be analized by traditional methods. Diagnosis that is made annually to productive organizations in the Universityof Informatic Sciences is a data generator process that has not been treated with data mining techniques. This work has as objective the aplication of some of this techniques to data that is extracted during diagnose process.

Keywords:

Data, mining, diagnosis.

ÍNDICE

Introducción

La era en que nos encontramos está basada y dominada por la información. Todos o casi todos los ámbitos de la vida moderna dependen del uso eficaz de la información y del conocimiento de la realidad que esta genere o modifique. Desde la cultura, la política y el comercio hasta el ambiente doméstico, las nuevas tecnologías han invadido la vida del hombre actual haciendo más cómodas o más eficientes y complejas las relaciones humanas, sin embargo la generación de datos es vertiginosa, y su acumulación ha crecido a mucha más velocidad que el desarrollo y modernización de los métodos de análisis de la información. Mucho conocimiento valioso puede estarse perdiendo o dejándose de utilizar.

La aplicación de técnicas y herramientas del campo de la minería de datos constituye una alternativa importante para recuperar este conocimiento oculto y que hasta el momento yace implícito en los datos acumulados. "*Al enfrentar un ambiente más competitivo las empresas requieren detecnologías que les permitan pronosticar*" (MANEIRO 2008).

La minería de datos se ha definido por algunos como *"el descubrimiento eficiente de información valiosa, no-obvia de una gran colección de datos, cuyo objetivo es ayudar a buscar situaciones interesantes con los criterios correctos"* (MANEIRO 2008). Es vista por otros como "*una tecnología para el desarrollo y el descubrimiento de la información muchas veces oculta en los mismos datos. Información que muchas veces no se tiene en cuenta y que en determinados casos puede ser valiosa y crítica para un mejor conocimiento del negocio y aportar mayor base a la toma de decisiones en cualquier tipo de escenario"* (GALVIS and MARTÍNEZ 2004). Esta información valiosa puede generar el conocimiento necesario para el éxito productivo a partir de una correcta toma de decisiones.

La información que genera un proceso de minería puede confirmar ideas o producir nuevas afirmaciones sobre un campo que se estudie. La digitalización e informatización de la vida moderna genera cantidades enormes de datos para acumular, mediante lectura de tarjetas, perfiles de clientes, descripciones de productos y procesos, encuestas, internet, y muchos más. "*En la actualidad, alrededor del mundo, se ha estimado que el crecimientode los datos almacenados en las bases de datos se duplica cada 20 meses, mientras que las técnicas de análisis de información no han tenido un desarrollo equivalente."* (MÉNDEZ 2006).

En la Universidad de las Ciencias Informáticas (UCI) se realizan diagnósticos a la actividad productiva de forma anual para el apoyo a la toma de decisiones. Durante estos diagnósticos se aplican técnicas de recopilación de información. Una de estas técnicas consiste en realizar encuestas de acuerdo a las necesidades de la Alta Gerencia de la Universidad. Son encuestados la totalidad de los proyectos productivos tanto de servicio como de desarrollo. El centro que realiza esta actividad es el Centro Nacional de Calidad de Software (CALISOFT), del Ministerio de la Informática y las Comunicaciones de Cuba.

En la actualidad el centro Calisoft cuenta con muchos datos derivados de varios procesos de diagnóstico aplicados durante 5 años a distintas áreas productivas de la UCI, por lo que resulta importante extraer de los datos generados por estos procesos toda la información valiosa que contenga. El conocimiento que se pueda formular de la información obtenida puede ser importante para la actividad productiva, ya que en estos datos acumulados puede encontrarse información no trivial que apoye a la toma de decisiones a nivel gerencial.

No se ha realizado hasta ahora un análisis profundo a estos datos más allá de la obtención de métricas y la elaboración de indicadores durante la aplicación del proceso de diagnóstico. Los patrones que siguen las colecciones de datos son importantes para analizarse, pues determinan características y conductas de los productos y/o los centros que los producen.

Existen tendencias positivas o negativas que atender ya que pueden dar al traste con el éxito o fracaso de una meta determinada. Estos datos no generan por sí solos nueva información valiosa sino que a través de un proceso más profundo de análisis se puede obtener información que apoye la toma de decisiones para mejorar la producción en la Universidad. No basta tener acumulada y organizada esta información resultante de los diagnósticos, sino que es necesario extraer nuevo conocimiento hasta ahora oculto en esos datos que contribuya a la toma de decisiones a nivel gerencial, aportando nuevos elementos o confirmando los que ya se conocían.

Problemaa resolver:

¿Qué patrones se pueden obtener de los datos recogidos en los diagnósticos realizados a los proyectos productivos de la UCI para contribuir a la toma de decisiones?

Objetivo general:

Obtener patrones mediante la generación de reglas a partir de los datos almacenados del proceso de diagnóstico para contribuir a la toma de decisiones.

Objeto de estudio:

Proceso de diagnóstico a las organizaciones productivas de la UCI.

Campo de acción:

Datos recogidos durante el proceso de diágnóstico sobre el programa de mejora en los años 2010 y 2012.

Objetivos Específicos:

- Definir los conceptos necesarios para la investigación.
- Establecer la herramienta y metodología a utilizar en el proyecto.
- Analizar los diagnósticos realizados.
- Aplicar técnicas de minería de datos basadas en reglas a los datos recogidos en los diagnósticos.
- Validar los resultados obtenidos.

Tareas de la investigación:

- Realización de búsquedas bibliográficas sobre conceptos de tipos de conocimiento, minería de datos, metodologías, algoritmos, herramientas.
- Análisis sobre los tipos de conocimiento, minería de datos, metodologías, algoritmos, herramientas.
- Selección de las metodologías y herramientas de minería de datos a usar.
- Análisis de la actividad productiva en la UCI.
- Estudio del proceso de diagnóstico realizado en la UCI.
- Selección de los algoritmos de minería de datos a usar.
- Aplicación de las técnicas de minería de datos seleccionadas a los conjuntos de datos seleccionados para el estudio.
- Realización de encuestas y entrevistas a personas especializadas en la realización del proceso de diagnóstico y otras áreas del conocimiento relacionadas para la validación de los resultados.
- Validación de la solución mediante encuesta a la Alta Gerencia sobre el aporte de dicha propuesta a la toma de decisiones.

El trabajo estará dividido en tres capítulos, resumidos a continuación:

Capítulo 1:

Se analizan varios conceptos relacionados con el dominio del problema. Se describe la situación actual de la problemática descrita y las actuales tendencias. Varias ideas son

valoradas para asumirlas en el trabajo. Se escogen las herramientas que se pueden utilizar y la metodología adecuada, describiendo la misma después de compararla con otras tendencias. Por último se analiza el escenario específico al que se refiere el objetivo general del trabajo y las ventajas de actuar en consecuencia con este objetivo.

Capítulo 2:

En este capítulo se analiza el negocio y se preparan los datos para su análisis según propone la metodología seleccionada. Se analizan y seleccionan las técnicas de minería que se van a utilizar. Varios modelos de minería de datos son construidos.

Capítulo 3:

Se valida la propuesta de solución, mediante métodos que permitan medir la utilidad de los resultados y su aporte a la toma de decisiones. Se utilizan gráficos, encuestas y entrevistas a la Alta Gerencia y un comité de expertos que facilitan una valoración del trabajo.

Capítulo 1 Fundamentación teórica

.1.1. Introducción

En este capítulo se presenta el marco teórico de la investigación abordando conceptos de diferentes autores acerca de los términos que se utilizan. Se presenta la principal tendencia en cuanto al análisis de grandes volúmenes de datos en la actualidad. La minería de datos como un recurso fundamental del Knowledge Discovery in Databases o Descubrimiento de Conocimiento en Bases de Datos (KDD), es definida y mostradas sus herramientas y aplicaciones.

Para la realización eficiente de un proyecto de minería de datos es importante la utilización de una metodología, en función de lo cual se comparan las más usadas en la actualidad describiéndose cada una y detallándose la metodología escogida para el desarrollo de esta investigación.

Se aborda también la aplicación el proceso de diagnóstico a las organizaciones productivas en la Universidad de las Ciencias Informáticas, y los datos recogidos por este proceso.

Se muestran las conclusiones parciales según los objetivos del capítulo.

.1.2. Gestión del conocimiento

.1.2.1. Datos, información y conocimiento

Se puede entender como dato al "*Antecedente necesario para llegar al conocimiento exacto de algo o para deducir las consecuencias legítimas de un hecho." (LENGUA 2011)*

"Los datos están constituidos por los registros de los hechos, acontecimientos, transacciones, etc. Pueden ser series de números o de caracteres que por sí mismos no constituyen información. Los datos se pueden considerar la materia prima para obtener la información." (OVIEDO 2006).

Se entenderá por tanto a los datos como símbolos, ya sean letras, números, o cualquier representación que signifique una cantidad, una medida, una descripción, pero que por sí mismos no pueden comunicar una información de tal manera que tampoco influyan en las decisiones de quien los reciba.

Según Idalberto Chiavenato, información *"es un conjunto de datos con un significado, o sea, que reduce la incertidumbre o que aumenta el conocimiento de algo. En verdad, la información es un mensaje con significado en un determinado contexto, disponible para uso inmediato y que proporciona orientación a las acciones por el hecho de reducir el margen de incertidumbre con respecto a nuestras decisiones"* (CHIAVENATO 2006). Según Czinkota y Kotabe la información "*consiste en datos seleccionados y ordenados con un propósito específico*" (CZINKOTA and KOTABE 2001). Para Ferrell y Hirt, la información *"comprende los datos y conocimientos que se usan en la toma de decisiones"* (FERRELL and HIRT. 2004).

El "*conocimiento significa entonces apropiarnos de las propiedades y relaciones de las cosas, entender lo que son y lo que no son"* (MUÑANTE 2004). Se puede ver como una actividad cognitiva que relaciona los datos y la información que estos representan a partir de la realidad.

Las definiciones de Chiavenato y Muñante son las que mejor describen la relación entre datos, información y conocimiento. Tomando elementos de ambas para esta investigación, se entenderá que los datos relacionados coherentemente van a constituir una información, la cual producirá conocimiento a partir de la acción del individuo y la apropiación que haga de ésta para un propósito específico.

Las propiedades manifestadas en los datos constituyen una información que se traduce enconocimiento mediante la percepción cognitiva y asimilación de ésta por parte de los individuos.

Es comprensible que en volúmenes considerables de datos parte de la información quede implícita y fuera del alcance de los métodos analíticos tradicionales. En muchas ocasiones se pierde la posibilidad de obtener conocimiento valioso para la toma de decisiones, sin embargo existen técnicas y herramientas que permiten recuperar esta información. La minería de datos es uno de los campos de la informática que se encarga de esta función.

.1.3. KDD (Knowledge Discovery in Databases)

Actualmente, el volumen de datos almacenados excede por mucho nuestra capacidad de analizar, reducir y utilizar estos datos sin el uso de técnicas de análisis automatizadas. Esta cantidad de datos crece cada día.

Según el profesor de la Universidad de Harvard Pardis Sabeti, existen actualmente grandes conjuntos de datos que se quieren explotar, los cuales contienen muchas relaciones que se desean entender, sin embargo son conjuntos tan grandes que nos es posible hacerlo mediante el ojo humano, un algoritmo de análisis de datos es una forma de resolver este problema (JORGE 2011).

Al proceso completo de extracción de información, la preparación de los datos y la interpretación de los resultados se le denomina Knowledge Discovery in Databases (KDD), definido más formalmente como "*el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles*" (FAYYAD *et al.* 1996). Se trata de encontrar patrones, tendencias y relaciones al interpretar grandes cantidades de datos. Para esto se apoya en técnicas de aprendizaje automático, estadística, representación del conocimiento, razonamiento basado en casos, inteligencia artificial, visualización de datos, etc. En el KDD son comunes los problemas de inducción de reglas, clasificación y *clustering* (agrupamiento), reconocimiento de patrones, modelado predictivo y otros.

KDD tiene la capacidad de recuperar información nueva y significativa a partir de los datos existentes.

.

Figura 1.KDD (LÓPEZ and HERRERO 2006).

KDD es un proceso interactivo e iterativo, que involucra numerosos pasos, la minería de datos es una parte del proceso.

.1.4. Origen de la minería de datos

La Minería de datos puede verse informalmente como un conjunto de técnicas, herramientas y resultados de investigaciones que resultan en el aprovechamiento de información oculta en grandes cantidades de datos.

En la década de los 60 los estadísticos ya manejaban términos como *data fishing*, *data mining* y *data archaeology*, sin embargo no es hasta los años 80 cuando los especialistas comienzan a consolidar el término de Minería de Datos. En 1991 G. Piatesky-Shapiro y Frawley enunciaron la data mining como *la extracción no trivial de información implícita, desconocida previamente, y potencialmente útil desde los* datos (PIATESKY-SHAPIRO and FRAWLEY 1991) otros conceptos son:

- *"…el proceso de extracción de información previamente desconocida, válida y procesable desde grandes bases de datos para luego ser utilizada en la toma de decisiones"*(CABENA *et al.* 1997).

- *"…la exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos"* (BERRY and LINOFF 1997).

- *"…un conjunto de técnicas agrupadas con el fin de crear mecanismos adecuados de dirección, entre ellas puede citarse la estadística, el reconocimiento de patrones, la clasificación y la predicción." (RODRÍGUEZ and PÉREZ 2002).*

.1.5. Aplicaciones de la minería de datos.

La minería de datos es utilizada por cientos de empresas en la actualidad, y sus beneficios son importantísimos para la toma de decisiones. Es utilizada en el marketing para la identificación de patrones de compra de los clientes, segmentación de clientes, consistente en la agrupación de los clientes concaracterísticas similares, por ejemplo demográficas. La minería de datos es útil también para predecir respuestas a campañas de promociones por correo electrónico y análisis de cestas de la compra, o sea descubrir relaciones entre productos.

En el sector de las compañías de seguros y la salud privada para predecir patrones de comportamiento, fraudes, venta de pólizas, etc.

En los bancos es útil para identificar clientes, detectar patrones de uso fraudulento en tarjetas, predecir clientes e identificar reglas de mercadeo.

En la medicina se pueden emplear técnicas de minería de datos para la identificación de terapias médicas satisfactorias para diferentes enfermedades. La asociación de síntomas y clasificación diferencial de patologías. El estudio de factores genéticos, precedentes, hábitos, alimenticios y de riesgo para la salud en distintas patologías. Estudios epidemiológicos, identificación de terapias médicas, etc. (LÓPEZ and HERRERO 2006).

Se puede apreciar también un amplio espectro de utilización en áreas como la biología, la industria farmacéutica, la educación y muchos más. En casi todas las áreas del conocimiento humano donde se almacenan grandes cantidades de datos, la Minería de datos constituye un apoyo fundamental para la toma de decisiones.

.1.6. Herramientas de minería de datos

En la actualidad se utilizan varias herramientas informáticas para la minería de datos, a continuación se describen las más conocidas y utilizadas.

WEKA: *"… es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Los algoritmos pueden ser aplicados directamente a un juego de datos o llamados desde tu propio código en java. Weka contiene algoritmos de pre-procesamiento, clasificación, regresión, agrupamiento, reglas de asociación y visualización. Es también adecuado para el desarrollo de nuevos esquemas de aprendizaje automático. Weka es un software libre bajo la Licencia Pública General GNU." (GROUP 2012).*

WEKA es el acrónimo de *Waikato Environment for Knowledge Analysis*, permite a través de la interfaz gráfica el acceso a esta gran colección de técnicas de análisis de datos. Solo requiere que los datos a analizar se encuentren en un cierto formato o se conviertan al mismo; es un formato conocido como ARFF (*Attribute-Relation File Format*). La licencia específica de Weka es GPL, por lo que es de libre distribución y difusión. Tiene la ventaja adicional de ser independiente de que como está desarrollado en Java es independiente de la arquitectura, pudiendo correr en cualquier plataforma que tenga una máquina virtual Java disponible.

RapidMiner: Es la herramienta más ampliamente usada en el mundo para la minería. Funciona como una aplicación independiente para el análisis de datos o como motor de minería invocado desde aplicaciones desarroladas por los usuarios. Es de las más populares, sin embargo sus versiones Enterprise pueden ser de código cerrado. La versión *Community*

Edition no contiene soporte para varias características del software como se aclara en el sitio web oficial que comercializa el producto. (DORTMUND 2001).

R: "*es un lenguaje y un entorno para computación y gráficos estadísticos, un conjunto integrado de servicios de software para la manipulación de datos, cálculo y representación gráfica".* (R-PROYECT 2012) R está orientado a la estadística pero es útil también en tareas de clasificación y agrupamiento, modelado gráfico, etc.

Knime: *(Konstanz Información Miner) Es una plataforma de código abierto para el procesamiento, análisis, exploración e integración de datos. Es utilizado por los profesionales de la industria y el mundo académico en más de 60 países.*(BERTHOLD 2012).

KNIME se desarroló con *Eclipse,* está por tanto programado en Java. Permite la creación de modelos estadísticos y de minería de datos y existe la posibilidad de llamar desde ella a Weka y/o de incorporar código R.

En esta encuesta realizada a través de internet en el año 2012 a 1480 usuarios, se preguntó que herramienta analítica/de minería de datos, utilizó durante los 12 meses anteriores.

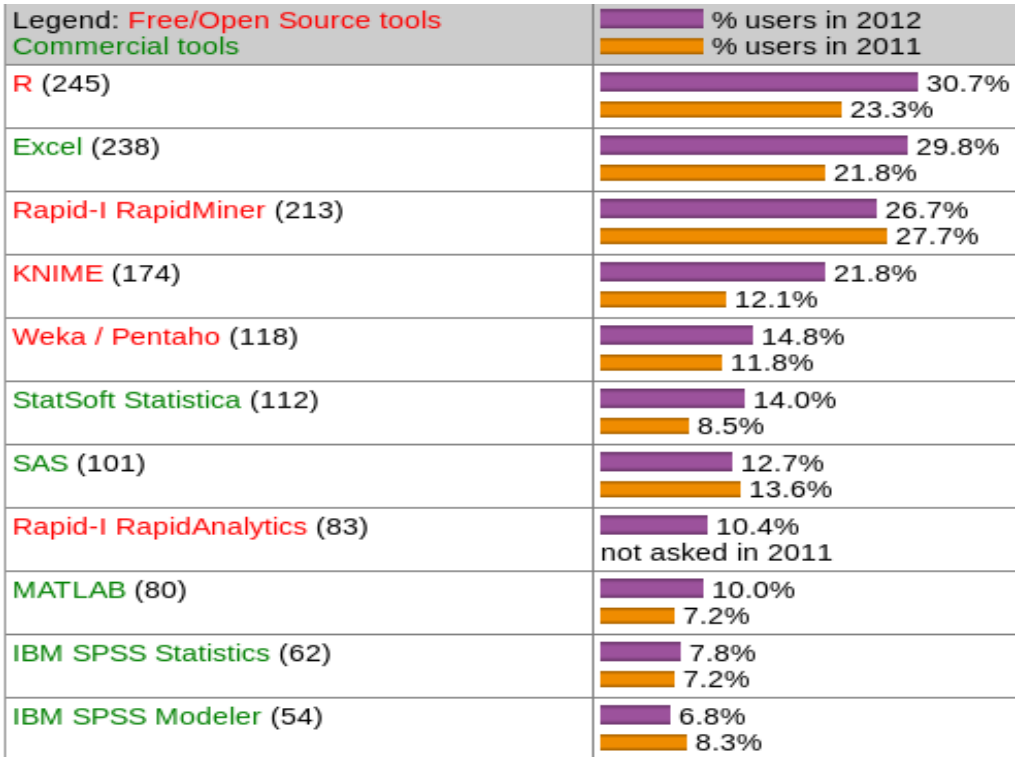


Figura 2. Encuesta sobre el uso de herramientas de minería de datos (KDNUGGETS 2010).

Se puede apreciar en la tabla el avance o retroceso de la preferencia de los usuarios por las diferentes herramientas entre 2011 y 2012.

Para esta investigación se utilizarán los algoritmos implementados en la herramienta WEKA, pero que son técnicas de minería comunes en casi todos los entornos mencionados anteriormente.

.1.7. Metodologías para la realización de proyectos de minería de datos

Aunque parece factible la aplicación del proceso de KDD durante la realización de proyectos de minería de datos, existe una preferencia mayoritaria por parte de los usuarios para utilizar metodologías entre las cuales se encuentran: CRoss-Industry Standard Process for Data Mining. (CRIPS – DM) y Sample, Explore, Modify, Model, Assess (SEMMA). Esto se muestra en la siguiente figura.

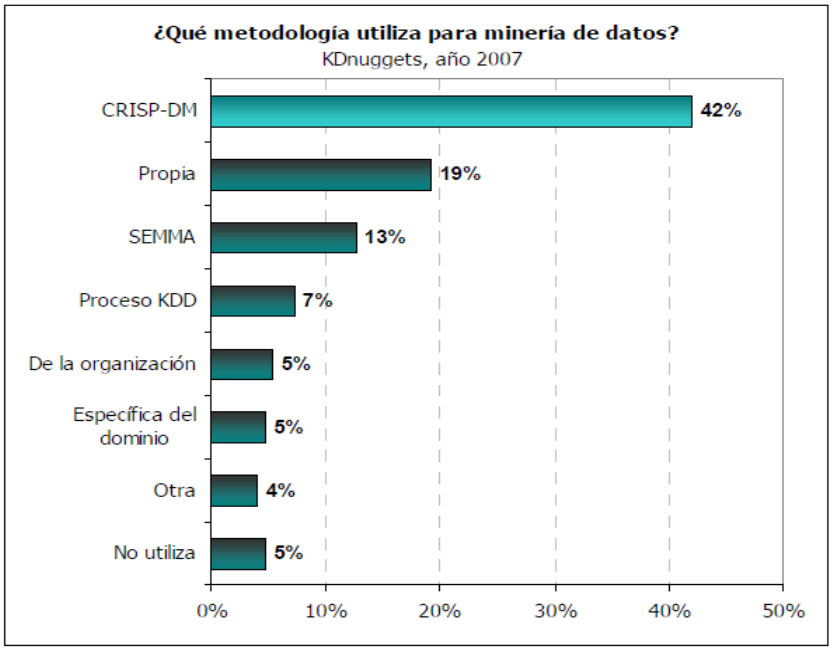


Figura 3. Encuesta sobre el uso de metodologías de minería de datos (KDNUGGETS 2010).

.1.7.1.SEMMA:

SEMMA, creada por el SAS Institute, se define como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos" (MOINE *et al.* 2011).

La compañía insiste no obstante en que SEMMA no es una metodología "*SEMMA no es una metodología de minería de datos, sino más bien una organización lógica de la herramienta*

funcional conjunto de SAS Enterprise Miner para la realización de las tareas principales de la minería de datos."

"El nombre de esta terminología (SEMMA) es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración). La metodología SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS." (MOINE *et al.* 2011).

"A partir de una muestra estadísticamente representativa de los datos, SEMMA hace que sea fácil de aplicar técnicas estadísticas de exploración y visualización, seleccionar y transformar las variables predictivas más importantes, el modelo de las variables para predecir los resultados y confirmar la exactitud de un modelo." (SAS 2012).

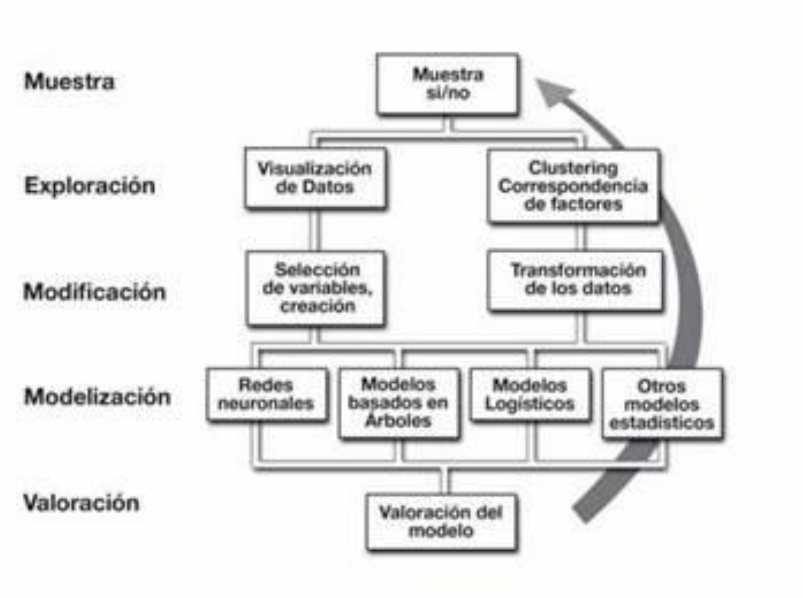


Figura 4. Fases de SEMMA(ESPAÑOLES 2006).

SEMMA plantea la exploración como una búsqueda de tendencias y anomalías no previstas con el fin de obtener la comprensión y las ideas y una ayuda para refinar el proceso de descubrimiento a través de técnicas estadísticas como el análisis factorial, análisis de correspondencia, y la agrupación.

El paso correspondiente a la modificación se refiere a la creación, selección y transformación de las variables para centrar el proceso de selección del modelo. Ya sea la eliminación de variables que puedan incluir resultados poco significativos a la solución o la introducción de nuevas variables necesarias para el siguiente paso. También permite, dado que la minería es un proceso iterativo, la modificación de los resultados que se vayan obteniendo en cada ejecución del minado.

La modelación consistiría en permitir que el software buscara automáticamente una combinación de datos que prediga confiablemente el resultado deseado. Podrían ser técnicas como redes neuronales, en árboles, modelos logísticos, y otros modelos estadísticos.
La evaluación propone realizarla sobre porciones reservadas de la muestra o conjuntos de datos con resultados conocidos para la variable predicha en el modelo. "*Mediante la evaluación de los resultados obtenidos de cada etapa del proceso de SEMMA, puede determinar cómo modelar nuevas preguntas planteadas por los resultados anteriores, y por lo tanto proceder de nuevo a la fase de exploración para el refinamiento adicional de los datos."* (SAS 2012).

.1.7.2.CRISP-DM:

"CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining." (MOINE *et al.* 2011 2011).

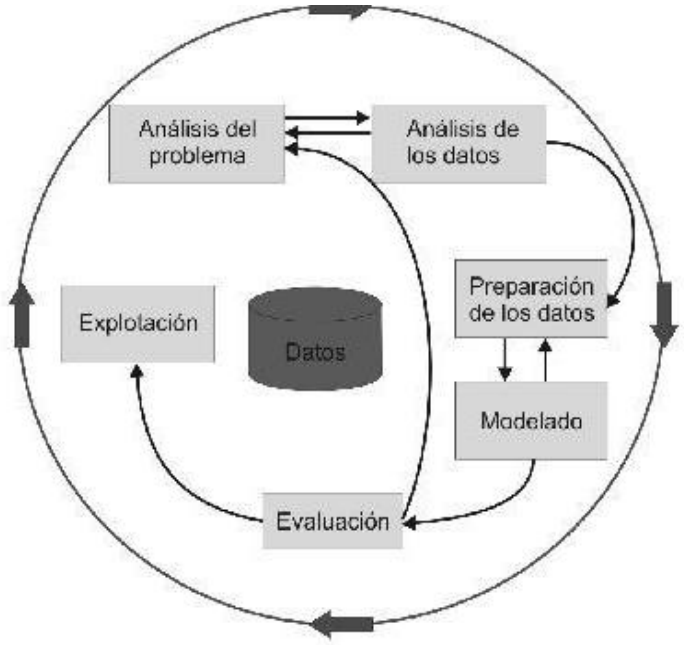


Figura 5. Fases de la metodología CRISP-DM (CRISP-DM 2007).

"La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos."(CRISP-DM 2007)

Este modelo consiste en 6 fases relacionadas de forma cíclica pero algunas son bidireccionales como se puede apreciar en la figura anterior, lo cual significa que la salida de cada fase puede volver a procesarse en la fase anterior. No es rígida, las fases se descomponen en varias tareas generales de segundo nivel que a su vez se descomponen en tareas específicas, aunque no dice cómo realizarlas. O sea que establece un conjunto de tareas y actividades en cada fase del proyecto, sin especificar cómo llevarlas a cabo. Aun así es más detallado en el proceso de minería que SEMMA y KDD que proveen solo una guía general para el trabajo.

.1.8. Metodología CRISP-DM

La metodología que se ha escogido para la realización de este proyecto de minería de datos es CRISP-DM, debido a que es la más ampliamente utilizada en el mundo, se encuentra debidamente documentada y detallada. Es una metodología para el desarrollo de proyectos de minería de datos y no está enfocada al trabajo con software específico de empresa alguna como es el caso de SEMMA, que es la segunda más utilizada.
A continuación se describe en detalle, puede ser sin embargo que por la naturaleza de los datos con los que se trabaje en el proyecto entre otros factores, se determine profundizar más o menos en algún nivel, fase o tarea de CRISP.

.1.8.1. Niveles jerárquicos de CRISP-DM

Esta metodología sigue una jerarquía de cuatro niveles enunciados en el acápite anterior. *"En el **nivel superior**, el proceso de minería de datos es organizado en un número de fases; cada fase consiste de varias tareas genéricas de **segundo nivel**.*

*El **tercer nivel**, el nivel de tarea especializado, es el lugar para describir como las acciones en las tareas genéricas deberían ser realizadas en ciertas situaciones específicas. La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos.*

*El **cuarto nivel**, la instancia de proceso, es un registro de las acciones, decisiones, y de los resultados de una minería de datos real contratada." (CRISP-DM 2007)*

Los autores de CRISP-DM especifican que en la práctica, puede suceder que muchas tareas se repitan o se realicen en un orden diferente, sin que esto sea contraproducente con el modelo.

La metodología plantea, contextualmente dentro de la minería cuatro dimensiones diferentes para tratar un problema de minería de datos. Estas son el dominio de aplicación, el tipo de problemas de minería de datos, el aspecto técnico y la herramienta(s) y/o técnica(s) a utilizar durante el proyecto. "*Un contexto específico de minería de datos es un valor concreto para una o más de estas dimensiones." (CRISP-DM 2007)*

.1.8.2.Fases del modelo de referencia CRISP-DM

- **Comprensión del negocio:** *Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos. (CRISP-DM 2007)*
- **Comprensión de los datos:** *La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta. (CRISP-DM 2007)*
- **Preparación de datos:** *La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [los datos que serán provistos en las herramientas de modelado] de los datos en brutos iniciales. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan. (CRISP-DM 2007)*
- **Modelado:** *En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario. (CRISP-DM 2007)*
- **Evaluación:** *Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente*

considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida. (CRISP-DM 2007)

- **Desarrollo:** *el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. (CRISP-DM 2007)*

.1.8.3.Tareas generales de cada fase de la metodología CRISP-DM

Es importante aclarar que cada una de estas tareas genéricas tiene objetos de salida que a su vez constituyen tareas más específicas para esa fase del proyecto.

Fase de Comprensión del Negocio:

1. Tarea Determinar los objetivos de negocio
2. Tarea Evaluar la situación
3. Tarea Determinar los objetivos de la minería de datos
4. Tarea Producir el plan del proyecto

Fase Comprensión de los datos:

1. Tarea Recolectar datos iniciales
2. Tarea Describir los datos
3. Tarea Explorar los datos
4. Tarea Verificar la calidad de los datos

Fase Preparación de datos:

1. Tarea Selección de datos
2. Tarea Limpiar datos
3. Tarea Construir datos
4. Tarea Integrar datos
5. Tarea Formatear datos

Fase Modelado:

1. Tarea Escoger la técnica de modelado
2. Tarea Generar la prueba de diseño
3. Tarea Construir el modelo
4. Tarea Evaluar el modelo

Fase Evaluación:

1. Tarea Evaluar los resultados
2. Tarea Revisar el proceso
3. Tarea Determinar los próximos pasos

Fase Desarrollo:

1. Tarea Desarrollar el plan
2. Tarea Planear la supervisión y el mantenimiento
3. Tarea Producir el informe final
4. Tarea Revisar el proyecto

Estas son según los creadores de la metodología, las tareas generales de cada fase. La metodología contiene además la "guía de usuario" que más que fases, tareas, y salidas, contiene el asesoramiento más detallado sobre cómo realizar proyectos de minería de datos. También explica los informes a producir durante y al final del proyecto. La metodología finaliza con un apéndice que incluye un glosario con terminología importante y una caracterización de los tipos de problemas de minería de datos.

.1.9. Diagnóstico a las Organizaciones Productivas en la Universidad de las Ciencias Informáticas (UCI).

La UCI cuenta con una estructura organizativa muy dinámica, que ha ido madurando con el paso del tiempo. En la UCI se ha combinado la formación de profesionales con la investigación y el desarrollo. *En el año 2007 se realizó un estudio a la actividad productiva de la universidad, donde se aplicaron varias encuestas (DURÁN 2012).* Este estudio no constituyó un proceso, sin embargo *era necesario realizar un análisis del entorno en cuanto a la actividad productiva involucrando a todas las áreas, los directivos, profesores y estudiantes vinculados a la producción en vista a realizar una proyección futura (DURÁN 2012).*

Al estar interactuando en un mercado internacional tan competitivo como los es el de software, existe la necesidad de un proceso que capture la información de los proyectos productivos y nutra a la Alta Gerencia del estado de los mismos para apoyar la toma de decisiones. *"La medición en las empresas es sumamente importante, ya que es una manera sencilla e ilustrativa, de mostrar los resultados de la gestión empresarial. En base a ello, se pueden establecer y diseñar estrategias que ayuden a desarrolar los aspectos positivos y contrarrestar los aspectos negativos resultantes." (PERESSON 2007).*

Debido a la importancia que reviste el contar con información oportuna y facilidades para su análisis en el momento de la toma de decisiones (CLAUDIA ETNA CARIGNANO 2005), se

aprueba en 2009 el *Diagnóstico a las Organizaciones Productivas en la Universidad de las Ciencias Informáticas*.

.1.9.1. Caracterización del Diagnóstico.

El Diagnóstico se realiza anualmente en la UCI, es aplicado por un grupo de especialistas de calidad de software que interactúan con la parte "Cliente" del proceso, la cual está formada por la dirección de la universidad y los "Diagnosticados" que son las organizaciones productivas a las que se les realiza dicho proceso.

Este proceso fue madurando y afirmándose con el paso de los años, adquiriendo elementos en cada edición que permitieran recuperar información cada vez más útil a la toma de decisiones.

Es en el diagnóstico 2008 se comienza a trabajar en una primera versión del proceso y es en el año 2009 que se aprueba el "—IPP-1000: 2009 Diagnóstico a las organizaciones productivas", este aportó plantillas para la documentación y una estructura organizativa de la información.(DURÁN 2012).

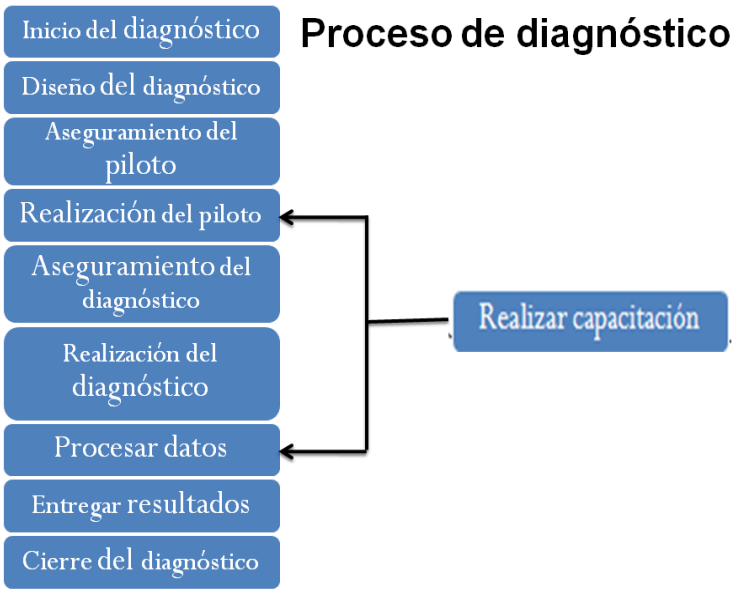


Figura 6. Proceso de Diagnóstico*a las Organizaciones Productivas en la Universidad de las Ciencias Informáticas*. (DURÁN 2012).

El Diagnóstico está formado por 10 subprocesos. Del subproceso *Procesar los datos*, se obtienen numerosos datos que tras ser debidamente revisados, se incluirán en el *Libro del*

Diagnóstico. *El objetivo de este subproceso es realizar el procesamiento de los datos para obtener la información en vista a elaborar los indicadores. El Grupo de diagnóstico depura los datos recogidos con el objetivo de encontrar anomalías en ellos de tal forma que los datos resultantes sean útiles o significativos* (DURÁN 2012).

.1.9.2.Naturaleza de los datosrecogidos

Aunque el producto final del diagnóstico es el Libro del Diagnóstico, durante la realización se generan muchos datos de las encuestas realizadas a cada proyecto diagnosticado. Las encuestas pueden estar enfocadas a un aspecto específico de la actividad productiva como pueden ser el éxito o fracaso de los proyectos, la correcta implementación de normas y estándares para la calidad del software, etc.

Los datos recogidos pueden ser almacenados en formato de documentos o bases de datos.*Se almacenan los documentos que se usan y se generan al aplicar alguna técnica para recopilar la información o para generar los resultados, como hojas de cálculo programadas, encuestas, bases de datos.*(DURÁN 2012).Estos son los datos que se analizan y procesan para elaborar indicadores y gráficas de tal manera que puedan ofrecer información útil a los intereses del diagnóstico.

Durante los diagnósticos se elabora un expediente que contendrá los datos que se recojan en cada plantilla. *El expediente de proyecto es una herramienta que agrupa y organiza todos los artefactos que se generan.*(DURÁN 2012). Este expediente resultará un repositorio estructurado que mediante una herramienta informática organizará y gestionará cada una de las partes que lo contienen.

Figura 7. Estructura del repositorio. (DURÁN 2012).

De esta manera pueden ser consultados, actualizados y servir como fuente de elaboración para las informaciones que requieran los especialistas, durante y después de la aplicación del Diagnóstico.

.1.9.3.Diagnósticos realizados entre 2010 y 2012

En el año 2010 el objetivo general del diagnóstico fue "*determinar el nivel de implementación, basado en las prácticas principales, de las áreas de procesos del nivel 2 de CMMI en los proyectos."* (CALISOFT 2010). Durante este diagnóstico no solo se encuestaron los proyectos productivos internos de la universidad sino que se extendió a las facultades regionales. *Tiene como alcance a los Centros productivos que desarrolan software o brindan servicio en la UCI y las Mini-UCI.* (CALISOFT 2010).

Según el Libro del Diagnóstico del 2010, en total se encuestaron 19 centros.

En 2011 tuvo como propósito *determinar fortalezas y debilidades de dicha actividad, cuyo conocimiento permitirá a la organización plantear y cumplir sus objetivos estratégicos con mayor eficacia.* (CALISOFT 2010) El objetivo general de este diagnóstico fue similar al del año precedente en cuanto a la diagnosticar la implementación y el avance del programa de mejora en la producción. Sus objetivos estrictamente fueron:

- *Determinar el avance del nivel de capacitación e implementación de las áreas de proceso del nivel 2 de CMMI por proyectos o líneas de producción.*

- *Identificar la problemática existente asociada a la iniciativa de mejora de proceso y la gestión de los riesgos en la organización.*

El Diagnóstico 2011 y el correspondiente a 2012 alcanzaron los 14 centros productivos de la UCI.

En 2012 estuvieron orientados a determinar el avance del nivel de capacitación e implementación de las áreas de proceso del nivel 2 de CMMI por proyectos o líneas de producción fundamentalmente, aunque también a valorar la organización al iniciar la mejora de proceso de software y detectar las fortalezas y las debilidades para acometer la mejora de proceso de software e identificar los riesgos de la estrategia de mejora.

Entre 2010, 2011 y 2012 se encuestaron, según los libros del diagnóstico, 178, 162 y 145 proyectos respectivamente de las 14 áreas productivas existentes, acerca del programa de

mejora, específicamente de la implementación de las áreas de proceso de CMMI como el modelo de calidad para las organizaciones productivas de la universidad.

La calidad de un proyecto, influye en los resultados del mismo a la llegada de su terminación. Toda organización desea tener éxito en todos los aspectos del desarrollo de un producto y se auxilian de las normas, estándares y modelos para asegurar en alguna medida un resultado productivo satisfactorio *"las empresas de este sector pertenecen a una nueva generación de emprendedores con novedosos modelos de negocios, nuevas estrategias de cooperación y competencia y originales sistemas de innovación."* (BÁRCENA and PRADO 2011)—*El hecho de que la industria de software se organice a partir de estándares técnicos que garantizan la interconectividad de los sistemas, la caracteriza como "industria de red* (TIGRE and MARQUES 2009).

Según el SEI, Instituto de Ingeniería de software de la Universidad Carnegie Mellon, refiriéndose a la implantación de CMMI a nivel mundial, *el porcentaje de las organizaciones con nivel 3 de madurez está en constante aumento* (INSTITUTE 2010). Lo cual muestra el interés mundial por avanzar en la implementación de los niveles del estándar. En la UCI actualmente se ha implementado hasta el nivel 2 de CMMI. Este modelo *"CMMI tiene la intención de proporcionar una guía para mejorar los procesos."* (IEEE 2004). Es natural pensar que la mejora de los procesos produzca mejores productos resultantes de estos. Sería además importante probarlo en una muestra real y significativa de estos proyectos,conocer las características que engloban los proyectos certificados y hasta que punto la implementación y el conocimiento de las áreas de procesos de CMMI ha influido en los resultados, también qué se puede esperar de los proyectos aún por certificarse en comparación al resto que lo ha podido lograr.

El resultado natural esperado no siempre corresponde al resultado real de un proyecto. Pero conocer los aspectos similares de los que han tenido éxito o no en un área específica, puede ser clave para el ulterior desarrollo de otros proyectos bajo el mismo modelo o estándar.

La aplicación de técnicas de minería a los datos recogidos como parte del proceso de Diagnóstico sobre la implantación del Programa de Mejora puede descubrir algunos de estos factores antes mencionados.

.1.10. Conclusiones parciales

- En cuanto a datos, información y conocimiento los autores coinciden en varios puntos para definirlos, sin embargo para la investigación se definió un nuevo concepto que los relacione.
- WEKA es la herramienta a utilizar para la minería.
- La metodología CRISP-DM es la más adecuada para guiar la investigación.
- Los diagnósticos a las organizaciones productivas son productores de una gran cantidad de datos.
- Mediante minería de datos, los datos resultantes del proceso de diagnóstico pueden arrojar información que confirme o agregue a la visión que tiene la Alta Gerencia sobre el desarrollo de software en los proyectos productivos de la universidad.

Capítulo 2 Fundamentación de la propuesta de solución.

2.1 Introducción

En este capítulo se analiza el problema planteado y los pasos a tener en cuenta para modelar la solución siguiendo a rasgos generales la metodología CRISP-DM. Algunas de las tareas que propone dicha metodología se realizaron en el capítulo anterior de este trabajo.

Se describe el volumen de datos, y los arreglos y limpieza que es necesario hacer en ellos para una mejor modelación a través de los algoritmos seleccionados, los cuales se detallan, describiendo su finalidad fundamental dentro de la minería de datos. Luego se describe la herramienta informática a utilizar para la presentación de los resultados y se dan las conclusiones parciales del capítulo.

2.2 Fase Análisis del Problema

El diagnóstico es un proceso generador de datos, los cuales se analizan y con la ayuda de la herramienta Excel se representan adecuadamente para brindar una información coherente a la alta gerencia sobre el estado en que se encuentran los proyectos productivos de la universidad.

La información con que se confecciona el Libro del Diagnóstico muchas veces no pasa de la mera elaboración de gráficas y estadísticas, únicas o comparativas con procesos anteriores, de acuerdo a los datos obtenidos de las entrevistas hechas durante el diagnóstico. Cualquier información nueva relativa a patrones o tendencias que un algoritmo de minería de datos descubra, es importante e interesante a los involucrados en este proceso pero durante la realización de varios diagnósticos no se ha incluido la minería de datos como una parte del proceso de análisis.

Los resultados podrían apoyar las ideas que se tienen respecto al desarrollo del software en los proyectos productivos o aportar nueva información hasta ahora oculta en este volumen de datos.

Se dispone de WEKA 3.6.5 del año 2011, con una gran variedad de algoritmos de predicción, análisis de asociaciones, agrupación y visualización mediante gráficos de dispersión para una mejor comprensión de la salida de los algoritmos y/o la obtención de información de patrones y tendencias básicas que se observen entre las parejas de datos analizados.

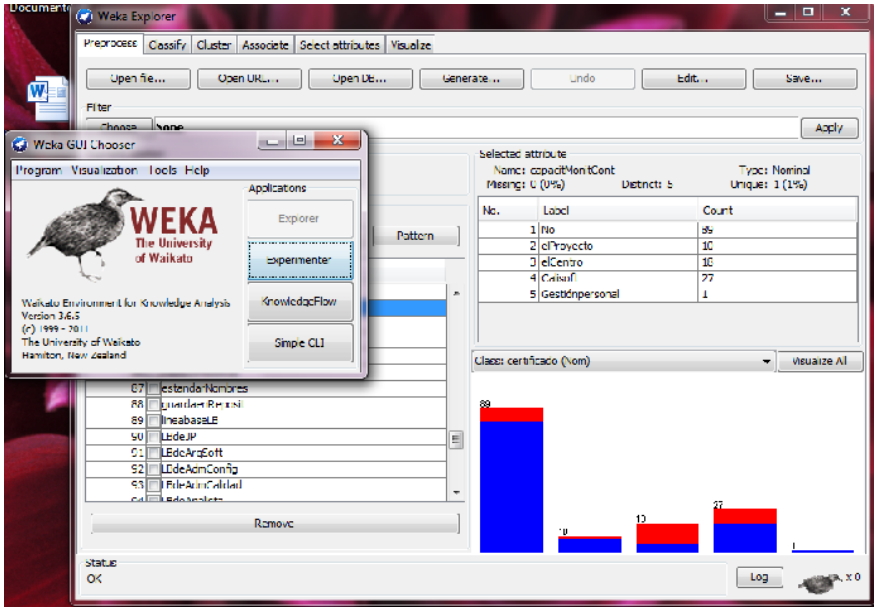


Figura 8. Interfaz de WEKA 3.6.5. (Fuente de elaboración: propia)

WEKA es utilizado por un número importante de usuarios para la minería de datos. Se puede apreciar que entre los internautas encuestados en la gráfica de la figura 2 hay un aumento de 2011 a 2012 con respecto a cantidad de usuarios que utilizan esta herramienta.

2.3 Fases deComprensión y preparación de los Datos

Los datos adquiridos corresponden a un libro de Excel elaborado durante el proceso de diagnóstico, donde a partir de los datos recopilados se construyen indicadores sobre el grado de implementación de las buenas prácticas de cada área de proceso de CMMI.

Se decidió tomar los datos relacionados con el año 2010 por ser el año en que se comenzó a trabajar en la implementación del modelo y por tanto en los centros se comenzó a organizar el trabajo según un estandar internacional de calidad. La elección de 2012 como el otro conjunto de datos a analizar se debe a que ya se logró la certificación de 3 de los centros que producen software en la universidad, y es de esperar una madurez en las prácticas de mejora.

Específicamente los datos recogidos en las entrevistas a los proyectos consisten en 29754 datos durante el diagnóstico del año 2010 y la cifra de 26523 datos durante 2012.

Los datos se recogieron mediante entrevista por lo que el factor humano influye en la calidad de los mismos, ya que influencias circunstanciales de tipo ambiental, físico o sicológico podrían afectar los datos. Con base en esto se decidió prescindir de los datos de mayor

subjetividad en la colección. Preguntas como "*Califique de 1 a 5 la necesidad (real o posible) de las siguientes…*" o "*¿Cree que…?*" carecen de un valor objetivo para este trabajo ya que se considera que pertenecen a la opinión subjetiva del entrevistado y no a la situación real del proyecto.

Dentro de los datos considerados para analizar mediante minería, se retiraron además de la selección los relacionados con las preguntas "*¿Por qué?*" tanto por el grado de subjetividad como por la poca relevancia que tendría durante el proceso de minado y el ruido innecesario que añadiría a los datos analizados.

Otra tarea que se llevó a cabo durante la verificación de la calidad y limpieza de los datos fue la de excluir los valores numéricos de la colección de datos. Esta decisión se tomó luego de verificar que la mayoría de estos valores correspondían a las preguntas subjetivas y constituían una pequeña minoría del total, por lo que resultarían poco significativos en los resultados, siendo engorroso enfrascarse en la discretización de estos valores reales-necesaria para muchas técnicas de minado-, además de por la característica del diagnóstico que mide estado actual de los proyectos y no su evolución en un periodo de tiempo, lo cual hace que se carezca de una variable temporal continua, factor imprescindible para aplicar potentes técnicas predictivas de minado que contiene la herramienta.

Finalmente se obtuvo para 2010 una colección de 22912 datos. La cantidad final de datos para analizar en 2012 alcanzó 16644 datos. En lo adelante el nomenclador de cada dato será llamado *atributo* y al valor en sí se le llamará *instancia* u ocurrencia del atributo. Por ejemplo el atributo *centro* tendría ocurrencias como CEIGE, CENIA, etc. Esto se hace para facilitar la comprensión de la salida de los algoritmos y sobre todo por el formato de archivo que comprende la herramienta WEKA para analizar los datos.

WEKA acepta nativamente un formato de archivo denominado arff, que significa*Attribute-Relation File Format.* La estructura de este formato se puede dividir en tres partes:

- Cabecera, que es donde se define el nombre de la relación y cuyo formato es`@relation <nombre-de-la-relación>`.
- Declaraciónes de atributos, que se expresan de la forma `@attribute <nombre-del-atributo><tipo>`.
- Sección de datos. Declaramos los datos que componen la relación separando entre comas losatributos y con saltos de línea las relaciones.`@Data` es el encabezado de esta sección.

Al ser miles de datos cualquier error puede ocurrir y es muy complejo de encontrar, por esto es muy importante al tomar los datos del excel que los nombres de los atributos no deben contener espacios, en caso imprescindible ponerlos entre comillas. Tampoco deben existir comas en los atributos o instancias, ya que provocaría que el software lo entendiera como dos ocurrencias distintas y se corrieran los valores, trayendo como consecuencia que existieran instancias sin atributos.

Figura 9. Fragmento de archivo final2010.arff (Fuente de elaboración: propia)

Los datos pueden ser de tipo String, numeric, interger, date. Para este trabajo se logró homogeneizar los datos al tipo nominal. De las 4 interfaces de la herramienta, se utilizará el modo explorador ya que es más sencillo obtener acceso a las técnicas de minería desde ella. Las restantes interfaces corresponden a una consola para trabajar con códigos, un modo experimentador para realizar clasificaciones avanzadas y compararlas mediante métodos estadísticos. Funciona de modo similar a una herramienta *Case* y da una idea del funcionamiento interno de WEKA.

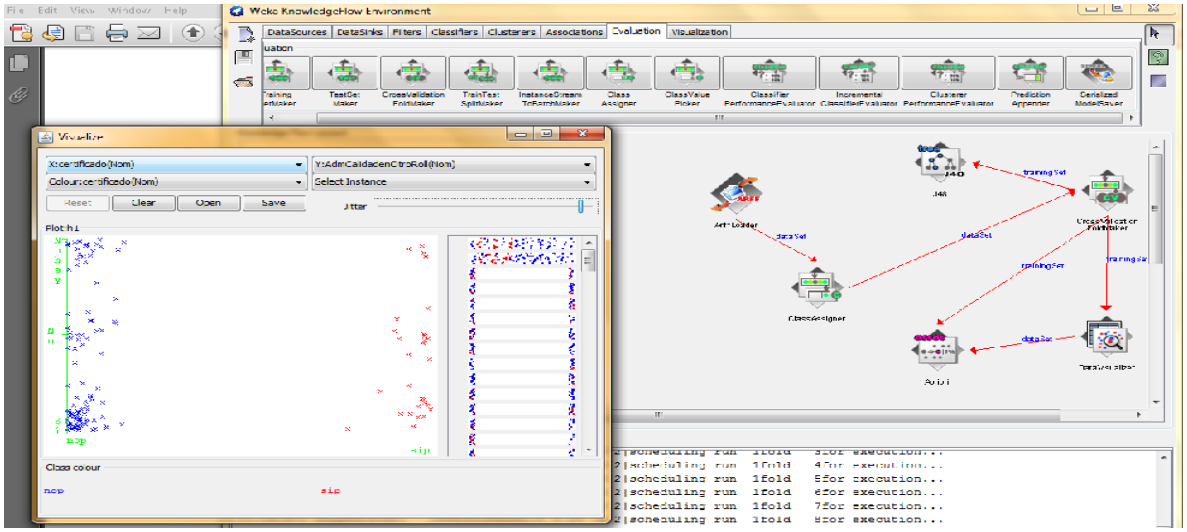


Figura 10. Interfaz *Knowledge flow* (Fuente de elaboración: propia).

La interfaz Explorer facilita el acceso mediante pestañas y botones, a las diferentes zonas de preprocesamiento de los datos, donde será posible aplicar filtros y editar los atributos e instancias del archivo arff cargado, aunque estos cambios no afectan el archivo sino solamenta la carga realizada. Contiene secciones para la clasificación, asociación, agrupamiento y visualización gráfica.

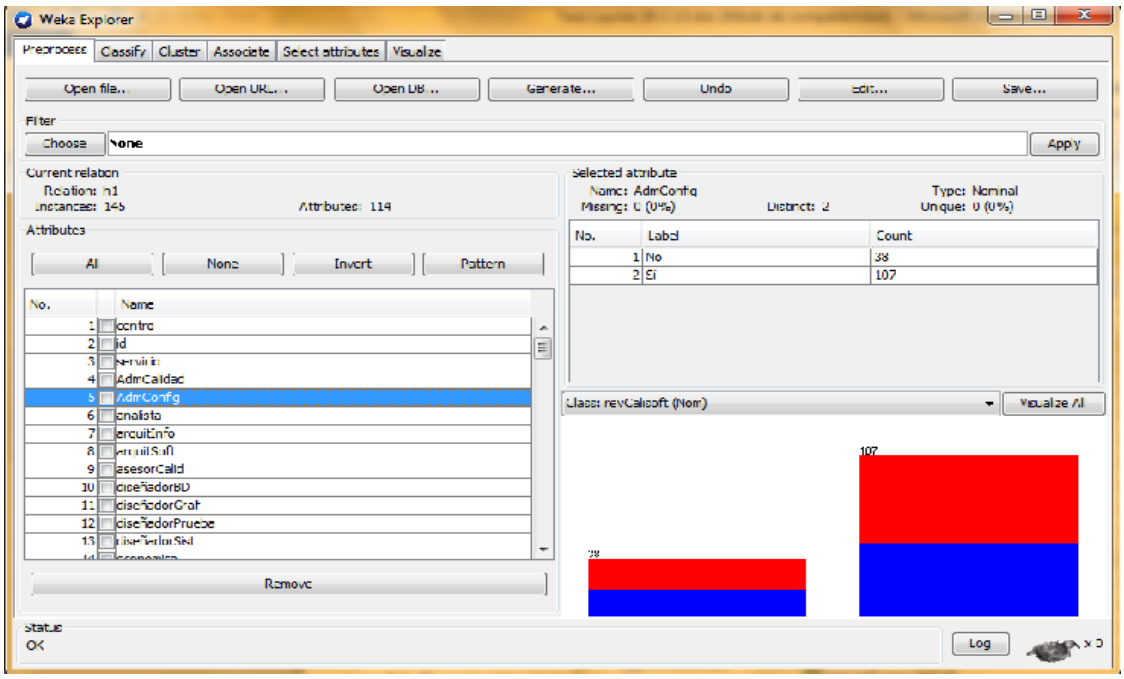


Figura 11. Interfaz *Explorer* (Fuente de elaboración: propia).

2.4 Descripción de los algoritmos a utilizar

Los algoritmos de minería de datos contenidos en la herramienta están agrupados en diferentes pestañas de la interfaz. Los del tipo predictivo son los algoritmos de clasificación, mientras que los de asociación, agrupamiento, selección de atributos y visualización, son más del tipo descriptivo.

Los algoritmos de clasificación necesitan un atributo llamado *clase,* alrededor del cual se desarrola el algoritmo, ya sea para predecir el valor de este atributo mediante técnicas de aprendizaje automático, o para establecer qué relaciones y en qué forma se relacionan los demás atributos con la clase. En este caso para la certificación de nivel 2 de CMMI se escogieron los centros a certificar, por lo tanto se contará además con un atributo *certificado* del cual se conoce su valor y será tomado como clase del conjunto.

Los algoritmos descriptivos, son llamados también de descubrimiento del conocimiento y no necesitan la presencia de un atributo clase para su funcionamiento, sino que descubren patrones y tendencias partir de las asociaciones que forman los datos entre sí y su comportamiento ante diferentes condiciones.

Utilizaremos en este caso tanto técnicas predictivas como descriptivas, pero siempre mediante algoritmos generadores de reglas.

Aplicaremos a estos datos varias técnicas de minería, mediante las cuales se puede obtener:

- Nueva información hasta ahora desconocida.
- Ninguna información útil.
- Información que confirma o ratifica lo que ya se conocía

Los algoritmos que se utilizarán serán:

- Selección de atributos mediante CfsSubsetEval y BestFirst.
- Clasificación mediante OneR.
- Clasificación PRISM.
- Reglas de Asociación mediante Apriori.

CfsSubsetEval y BestFirst

El algoritmoCfsSubsetEval es un evaluadorque calcula la correlación de la clase con cada atributo,y elimina atributos que tienen una correlación muy altacomo atributos redundantes. Va saleccionando los atributos con más relación con la clase pero menos interrelación entre ellos.

Este evaluador utiliza un método de búsqueda propio de la Inteligencia Artificial, el cual combina en sí las ventajas de otros dos métodos de búsqueda. Su nombre en español *MejorPrimero* combina la ventaja delos métodos de búsqueda "Primero en Profundidad" y "Primero en anchura". Del uno toma la ventaja de que permite encontrar una solución sin tener que expandirse completamente por todas las ramas del árbol de búsqueda que forma y del otro toma la ventaja de que no queda atrapada en callejones sin salida. En resúmen sigue un único camino cada vez, y lo cambia cuando alguna ruta parece más prometedora que la que se está siguiendo en ese momento.

El proceso de selección de atributos trata de seleccionar el subconjunto más pequeño de atributos tal que no se afecte significativamente el porcentaje de acierto en la clasificación. Un atributo se considera relevante si no es irrelevante o redundante. Un atributo es irrelevante si no afecta de ninguna forma el concepto meta y es redundante si no añade nada nuevo al concepto meta (ANGULO 2010).

La selección de atributos mayormente se utilizará en este trabajo para prestar mayor atención a las reglas que contengan alguno de estos atributos.

Algoritmo OneR

Es un clasificador simple que busca los atributos que mejor explican el atributo clase y escoje de ellos el que menor error tenga, retornándolo en forma de una regla.

Este clasificador construye un clasificador consistente en usar una única variable en el antecedente, es decir, se genera una regla que clasifica a un objeto en base a un solo atributo. Se generan todas las reglas del tipo "`Si variable = valor Entonces clase = categoría`" para una única variable. Este algoritmo se fundamenta en la tesis de que reglas de clasificación muy sencillas trabajan bien en la mayoría de las bases de datos empleadas. También suele usarse como algoritmo base para realizar comparaciones (Holte 1993).

Reglas de Asociación mediante Apriori

Estas reglas pueden ayudarnos a analizar los datos y atomar buenas decisiones dentro del ámbito del problema,son utilizadas para representare identificar dependencias entre atributos en una colección de datos.

El algoritmo Apriori *permite identificar las posibles correlaciones o interdependencias entre distintas acciones o sucesos; pudiendo reconocer cómo la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros.(AGRAWAL and SRIKANT 1994).*

Las reglas de asociación contituyen una técnica excelente ya que pueden predecir cualquier atributo y mostrarnos cualquier combinación de éstos. Las reglas no suelen utilizarse todas juntas sino que diferentes grupos muestran distintas regularidades del conjunto de datos.
Normalmente se basan en descubrir combinaciones de pares atributo-valor que ocurren con frecuencia en un grupo de datos.

Solo son seleccionadas las reglas que sean “interesantes”, esto es que cubran la mayor cantidad de instancias y que tengan la mayor precisión.

Algoritmo PRISM

PRISM es un algoritmo de inducción de reglas. Tiene la ventaja de que identifica en cada paso una regla que cubra algunas instancias para luego eliminar todos los ejemplos cubiertos por esta regla, pareciéndose más a un conjunto de reglas que a un árbol de decisión. Este algoritmo solo busca reglas perfectas o correctas de forma que cualquier regla con exactitud menor al 100% es considerada incorrecta. *La principal ventajas del algoritmo Prism es que opera de una manera muysimilar a la de un algoritmo divide y vencerás de tipo arriba-abajo o top-down (MORALES 2003).* Aunque pareciera una desventaja es interesante a esta investigación el hecho de que *las instancias cubiertas por una regla son eliminadas del conjunto deinstancias, de forma que las reglas subsecuentes actúan sobre las instanciasno cubiertas (MORALES 2003)*. Esto no será desventajoso ya que se utiliza otro algoritmo generador de reglas en el minado para cubrir cualquier regla que haya quedado fuera.

2.5 Resultados

Durante el 2010 se trabajó para preparar los centros que se iban a certificar. En este sentido se han diferenciado los proyectos de estos centros, y se ha adicionado un atributo *certificado* para distinguirlos, el cual tomará valor *sip* cuando se trate de un proyecto perteneciente a un centro en preparación de certificarse y *nop* cuando el centro no estaba trabajando en pos de esa meta.

La variable certificado será usada como variable clase en varias técnicas de minería. Sabiendo que finalmente en julio de 2011 se consiguió certificar los centros CEIGE, CEDIN Y CESIM con nivel 2 de CMMI, cuando se trate de los modelos correspondientes al analisis del

diagnóstico del año 2012, el atributo certificado identificará los proyectos pertenecientes a estos centros antes mencionados cuando tome el valor *sip.* Tomará el valor *nop* para el resto. WEKA identificó los valores de *sip* en rojo y *nop* en negro cuando se muestran en los gráficos de dispersión.

Se modeló primeramente el juego de datos correspondiente al diagnóstico 2010. Los modelos que se presentan corresponderán a este diagnóstico hasta que se especifique el año 2012, a partir de ese momento los modelos corresponderán a los datos del diagnóstico del año 2012.

Se suprimieron los atributos *id* y *centro* mediante el filtro *remove* de WEKA durante el preprocesado de los datos, una vez cargados en la herramienta. Esto impedirá que creen resultados innecesarios, y tributará a la confidencialidad de la información.

2.5.1 Modelación de los datos pertenecientes al diagnóstico realizado en 2010. Selección de atributos con CfsSubsetEval y BestFirst

Objetivo de la técnica: Determinar los atributos que más influían en el estado de los proyectos con respecto a la certificación.

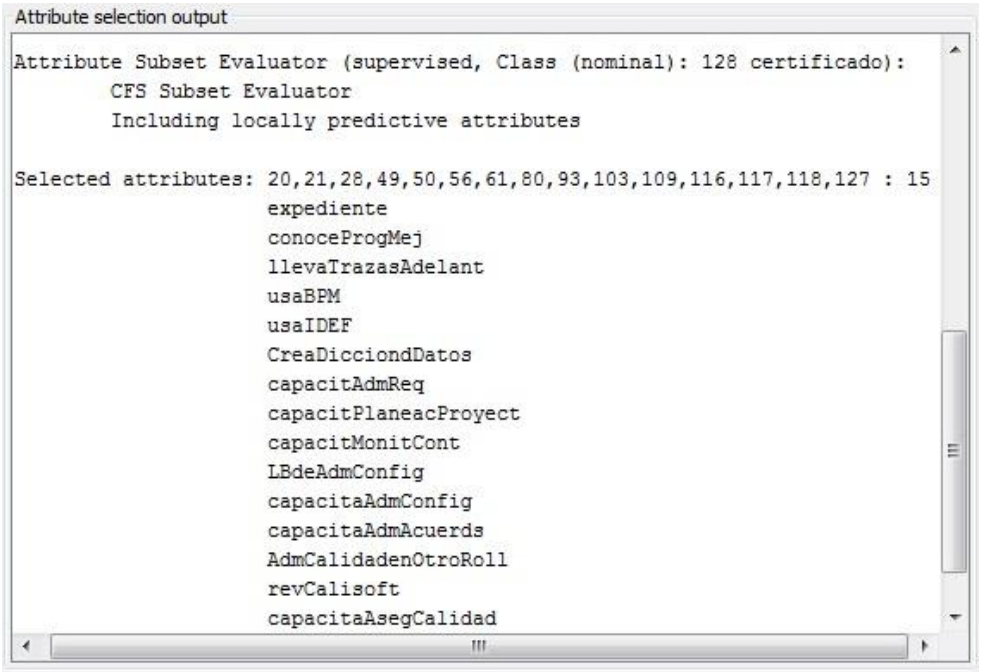


Figura 12. Salida de las técnicas de selección de atributos. (Fuente de elaboración: propia).

Los atributos que mejor se relacionaron con la clase corresponden a las preguntas:

- ¿Qué expediente de proyecto se usa en su proyecto?

- ¿Conoce el Programa de Mejora que se está llevando a cabo en la UCI?
- Llevar una trazabilidad desde los requisitoshacia los productos en que se van convirtiendo [trazabilidad hacia adelante].
- Marque cuáles técnicas de modelado de negocio utilizan en su proyecto: [BPM].
- Marque cuáles técnicas de modelado de negocio utilizan en su proyecto: [IDEF].
- Marque cuáles técnicas de análisis de requisitos utilizan en su proyecto [Crear diccionario de datos].
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Administración de requisitos?
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Planeación de Proyecto?
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Monitoreo y Control de Proyecto?
- ¿En caso de generar líneas base quién las aprueba para ser liberadas? [Administrador de la configuración].
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Administración de acuerdos con los proveedores?
- ¿El administrador de la calidad desempeña otro rol dentro del proyecto?
- ¿Su proyecto ha sido revisado o auditado por Calisoft?
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Aseguramiento de la calidad de los procesos y productos?

A continuación se observará mediante gráficas de dispersión la relación de algunos de éstos atributos con la clase del juego de datos, en los casos donde esta relación arroje patrones interesantes.

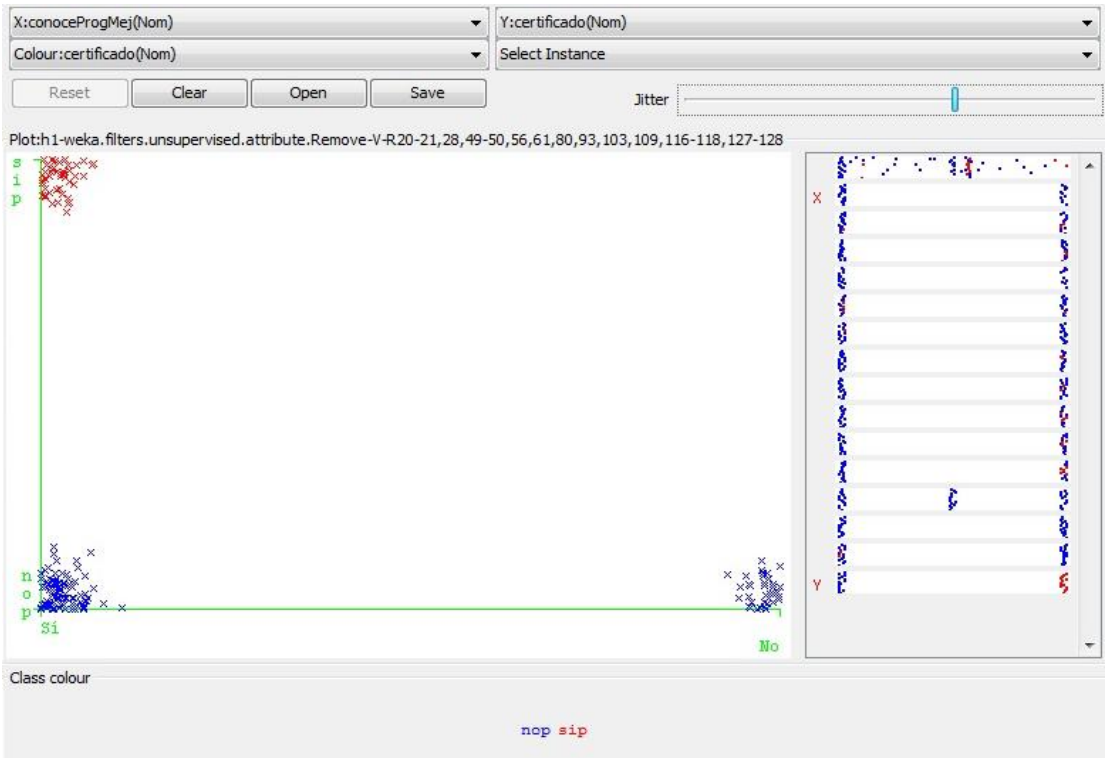


Figura 13. Gráfica de dispersión conoceProgMej Vs certificado. (Fuente de elaboración: propia).

Guías para la interpretación:

El eje x corresponde a si los proyectos conocían o no del programa de mejora.

El eje y muestra si se estaba trabajando o no por certificar esos proyectos.

Un resultado ideal sería que todos los proyectos conocieran del programa de mejora aunque no se estuviera trabajando con ellos para certificarlos.

Interpretación:

Se puede apreciar como-aunque no estaban siendo preparados para certificarse-, una cantidad considerable de proyectos no conocían del programa de mejora que se estaba llevando en la UCI en ese momento.

A continuación se observa el gráfico que relaciona los proyectos que los cuales las Líneas Base eran aprobadas por el Administrador de la Configuración, con el atributo que especifica si los proyectos tenían conocimiento sobre el programa de mejora que se estaba llevando a cabo en la universidad.

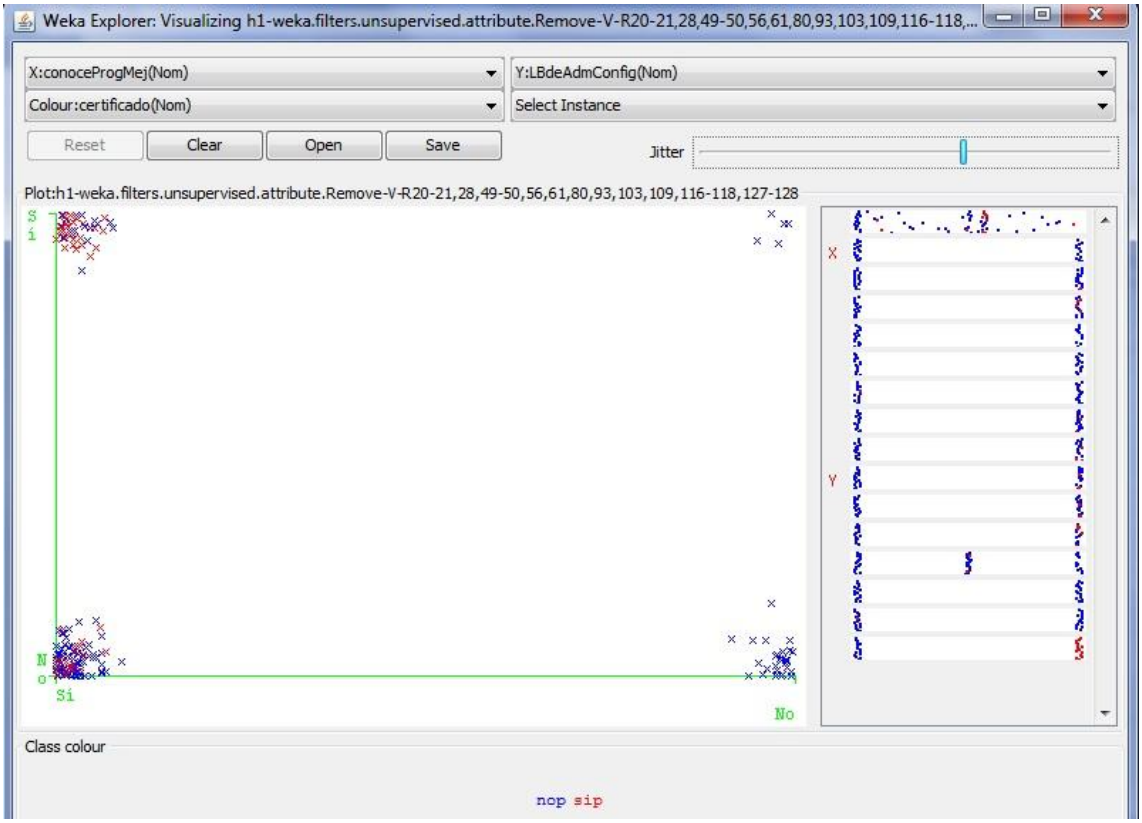


Figura 14. Gráfica de dispersión conoceProgMej Vs LBdeAdmConfig (Fuente de elaboración: propia).

Guías para la interpretación:

El eje x corresponde a si los proyectos conocían o no del programa de mejora.

El eje y corresponde a si el Administrador de la Configuración era quien aprobaba las Líneas Base de estos proyectos.

Un resultado ideal sería que en todos los proyectos que conocieran del programa de mejora, el Administrador de la Configuración aprobara las Líneas Base.

Interpretación:

Aunque la inmensa mayoría conocía del programa de mejora y dentro de esta mayoría se encontraba la totalidad de los proyectos para certificarse, las Líneas Base que se generaban no eran aprobadas por el Administrador de la Configuración, que según se especifica para el área de proceso de *Gestion de la Configuración*es quien debe establecer las Líneas Base del proyecto.

En esta gráfica se puede observar si los proyectos en los cuales se trabajaba por lograr la evaluación, el Administrador de la Calidad era cargado con otras responsabilidades adicionales a las suyas.

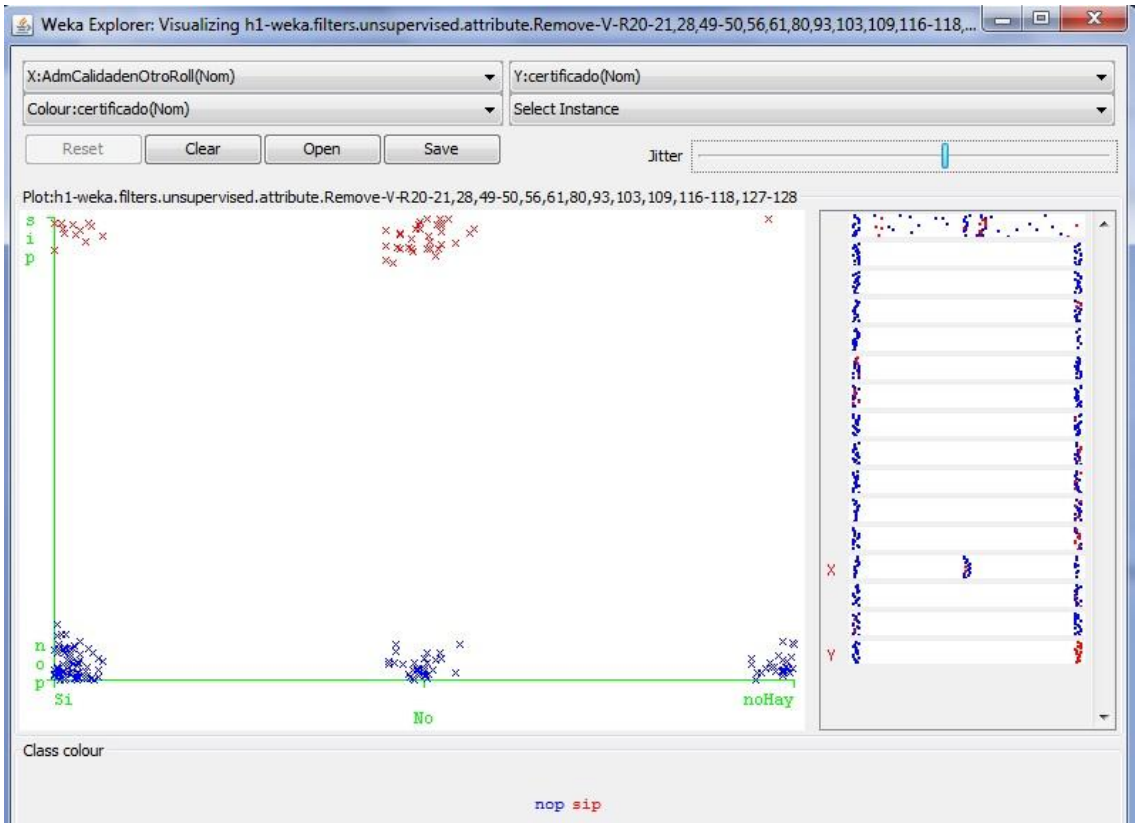


Figura 15. Gráfica de dispersión AdmCalidadOtroRol Vs certificado (Fuente de elaboración: propia).

Guías para la interpretación:

El eje x corresponde a si los proyectos empleaban al Administrador de la Calidad en otro rol además del suyo.

El eje y muestra si se estaba trabajando o no por certificar esos proyectos.

Un resultado ideal sería que en todos los proyectos donde se estuvo trabajando en pos de la certificación, el Administrador de la Calidad no ocupara otro rol más que el suyo.

Interpretación:

Cerca de la mitad de todos los proyectos en los cuales existía el rol administrador de la calidad, lo sobrecargaban con otras responsabilidades dentro del proyecto. En esta situación estaba el 25,5 % de los proyectos por certificarse.

Es importante señalar además que el algoritmo seleccionó las capacitaciones en 6 de las 7 áreas de proceso de CMMI como fundamentales en el resultado de la clase, siendo éstos un tercio de todos los atributos seleccionados.

2.5.2 Clasificación con OneR

Objetivo de esta técnica: Determinar el atributo que más influye en la certificación conviertiéndolo en una regla que explique la relación.

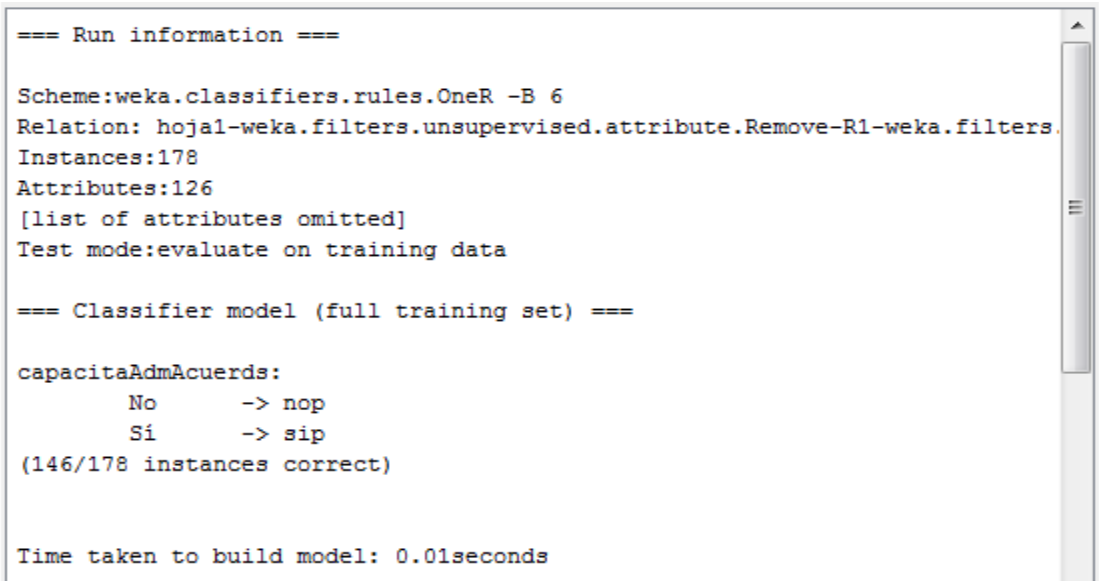


Figura 16. Salida del algoritmo OneR (Fuente de elaboración: propia).

Guía para la interpretación:

En el extremo superior izquierdo se muestra el atributo seleccionado por el algoritmo. A continuación se muestran los valores que puede tomar y como influye en la certificación.

Interpretación:

El algoritmo OneR muestra que la mejor relación con la certificación correspondía a un atributo de capacitación específicamente al referido a la capacitación en Administración de Acuerdos con Proveedores. La regla plantea quea los proyectos que no se escogieron para certificar, no se les dio capacitación en esta área.

2.5.3 Clasificación PRISM

Objetivo de la técnica: Obtener reglas que expliquen un atributo específico a partir de condiciones generadas por otros atributos.

La selección de atributos mostró a todos los atributos de capacitación, excepto el referente al área de Medición y Análisis, entre los 15 seleccionados. De estos atributos el referente a la capacitación en Administación de Acuerdos, fue seleccionado por el algoritmo OneR, y convertido en la mejor regla para clasificar la clase.

A continuación se utilizará PRISM para extraer patrones de la correlación entre este atributo y el juego de datos.

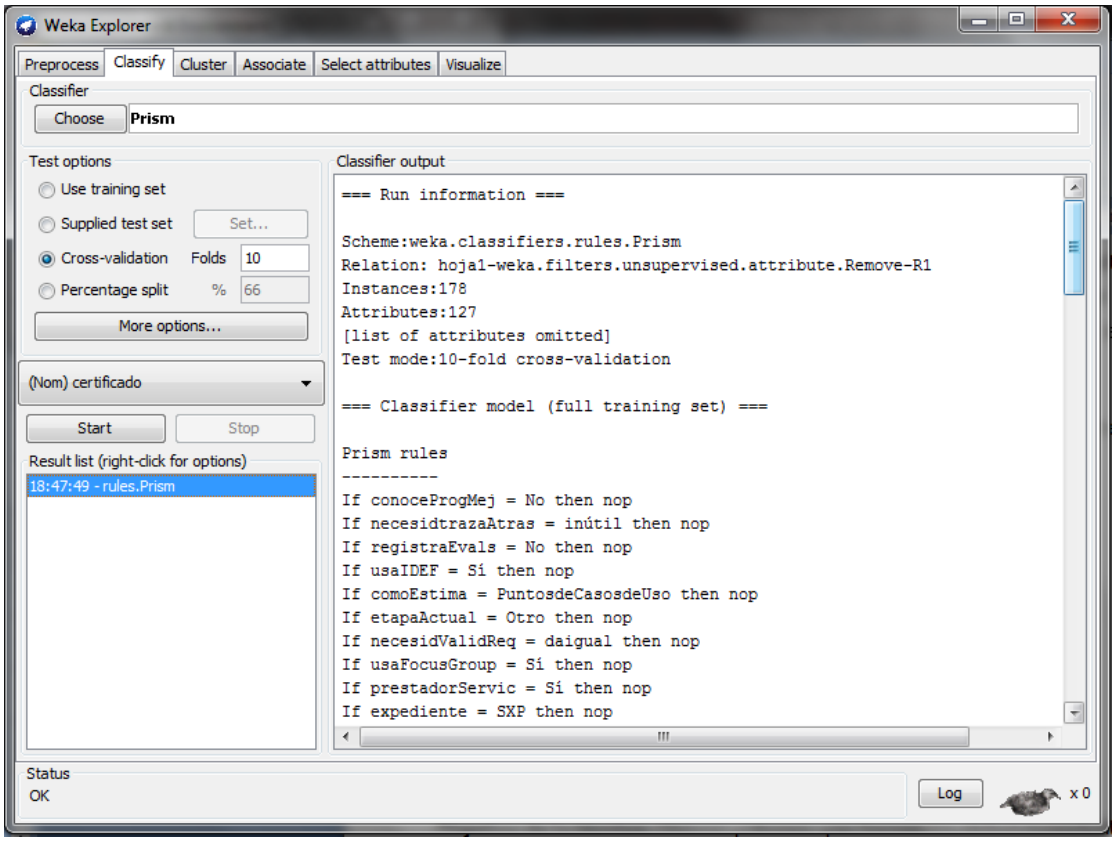


Figura 17. Salida del algoritmo PRISM (Fuente de elaboración: propia).

Tabla 1. Conjunto de reglas generadas por weka para la clase capacitaAdmAcuerds(Fuente de elaboración: propia).

Reglas de PRISM: capacitaAdmAcuerds	Reglas de PRISM: capacitaAdmAcuerds
If AdmCalidad = No then No If elaboraPlanCyMR = No then No If comoEstima = PuntosdeCasosdeUso then No If comoEstima = cocomo then No If comoEstima = Propia then No If necesidValidReq = daigual then No If necesidTrazaAdelant = inútil then No	If certificado = sip and gestorConoc = Sí then Sí If certificado = sip and necesidIdentReqInterfaz = daigual then Sí If certificado = sip and necesidFirmarRequisit = útil then Sí If etapaActual = Despliegue

If evalCambios = No and necesidtrazaAtras = Importante then No If capacitMonitCont = No and necesidDetectIncosist = Importantethen No If comoEstima = Experienciadelequipo and entendtoClientAnalist = Sí then No If comoEstima = Ninguna and necesidadObtenRequisit = Importante then No If comoEstima = Ninguna and usaAnalyDiseñ = No then No If economico = Sí and diseñadorPrueba = No then No If comoEstima = Ninguna and necesidIdentReqInterfaz = Importante and necesidtrazaAtras = Imprescindible then No If comoEstima = EstimaciónenelMicrosoftOfficeProject then Sí If comoEstima = Albet then Sí If capacitMedAnal = Sí and identifReqInterf = No then Sí If certificado = sip and necesidValidReq = útil then Sí If jefeOperac = Sí and AdmConfig = Sí then Sí If capacitMedAnal = Sí and etapaActual = ModelacióndelNegocio then Sí If certificado = sip and comoEstima = Experienciadelequipo then Sí	and usaPrototiposCaptura = Sí then Sí If capacitMedAnal = Sí and necesidDetectIncosist = útil then Sí If capacitMonitCont = Sí and necesidadObtenRequisit = útil then Sí If capacitMonitCont = Sí and etapaActual = AnálisisyDiseño then Sí If etapaActual = Pruebasinternas and arquitInfo = Sí then Sí If certificado = sip and etapaActual = EstudioPreliminar then Sí If necesidtrazaAtras = Imprescindible and revisionesAG = No then Sí If etapaActual = Pruebasinternas and usaVotoAcumulativo = Sí then Sí If capacitAdmReq = Sí and necesidespecifReqenDoc = útil and ArquiTecnicAplic = No then Sí If capacitAdmReq = Sí and etapaActual = EstudioPreliminar and diseñadorBD = Sí then Sí If etapaActual = Pruebasinternas and usaSoport = No and modelaNegocio = No then Sí If usaCuestionario = Sí and necesidIdentReqInterfaz = Imprescindible and arquitInfo = Sí then Sí If etapaActual = Requisitos and necesidtrazaAtras = daigual and planificador = Sí then Sí

Guía para la interpretación:

Cada regla consta de dos partes: A la izquierda de la palabra *then* se encuentra la condición en forma de uno o más pares de atributo-valor y a la derecha se encuentra la conclusión en forma del valor que toma el atributo para el cual se han generado las reglas.

Interpretación:

Pueden apreciarse patrones como: los proyectos que estiman mediante Cocomo, por experiencia -y no estaba en la preparación para la evaluación de nivel 2 de CMMI-, mediante un método propio o no estiman en absoluto, son proyectos que no han recibido esta

capacitación, lo que es igual a decir que no se capacitaban en las demás áreas. Esto es posible inferirloya que en la segunda columna del grupo de reglas se observa que solo han recibido capacitación en Administración de Acuerdos con Proveedores (AC) los proyectos que han recibido capacitación en las demás áreas.

Se observa además que se capacitaron solo los proyectos en los que se estaba trabajando para lograr la evaluación.

En los proyectos que recibían capacitación en AC se destacan roles como el Gestor de Conocimiento, Arquitecto de Información y el Planificador. El rol Económico era una responsabilidad poco que se usaba solo en los proyectos fuera de la evaluación de CMMI, por lo que también correspondía a proyectos sin capacitación en el área de procesos que se vincula a este rol.

2.5.4 Reglas de asociación

Objetivo de la técnica: Obtener patrones que forman los atributos mediante reglas.

1. jefeOperac=No prestadorServic=No 172 ==> jefeProyecto=Sí 172 conf:(1)
   1.1. jefeOperac=No usaFocusGroup=No 170 ==> jefeProyecto=Sí 170 conf:(1)
   1.2. usaFocusGroup=No 172 ==> jefeProyecto=Sí 172 conf:(1)
   1.3. prestadorServic=No 174 ==> jefeProyecto=Sí 174 conf:(1)
   1.4. guardaenReposit=Sí 173 ==> jefeProyecto=Sí 173 conf:(1)
2. prestadorServic=No 174 ==> jefeOperac=No 172 conf:(0.99)
   2.1. usaFocusGroup=No 172 ==> jefeOperac=No 170 conf:(0.99)
   2.2. usaFocusGroup=No 172 ==> jefeOperac=No 170 conf:(0.99)

3. analista=Sí controlProdyEntreg=Sí 166 ==> guardaenReposit=Sí 165 conf:(0.99)
   3.1. prestadorServic=No usaFocusGroup=No controlProdyEntreg=Sí 163 ==> guardaenReposit=Sí 162 conf:(0.99)
   3.2. usaNormas=Ninguno guardaenReposit=Sí 161 ==> prestadorServic=No 160 conf:(0.99)
   3.3. prestadorServic=No usaFocusGroup=No controlProdyEntreg=Sí 163 ==> guardaenReposit=Sí 162 conf:(0.99)

4. analista=Sí usaFocusGroup=No 163 ==> controlProdyEntreg=Sí 160 conf:(0.98)
   4.1. analista=Sí usaFocusGroup=No 163 ==> tieneReposit=Sí 160 conf:(0.98)
   4.2. usaVotoAcumulativo=No 167 ==> usaFocusGroup=No 162 conf:(0.97)

4.3. usaIDEF=No 165 ==> usaFocusGroup=No 161 conf:(0.98)

Guía para la interpretación:

Cada regla tiene una parte a la izquierda de la flecha y una parte a su derecha.
A la izquierda puede existir más de una pareja atributo-valor seguido por el número de casos que cubre. A la derecha habrá una pareja atributo-valor que es la consecuencia de la regla, seguida por su número de casos.
Por último se muestra el porciento de confiabilidad de dicha regla.

Interpretación:

Las reglas de asociación seleccionadas se obtuvieron con un nivel de confianza mínimo de 95%, lo cual implica que son reglas que se cumplen para casi la totalidad de los atributos que cubre cada una.

Las relaciones obvias o triviales se suprimieron ya que no representaban patrones útiles e impedían la subida de nuevas reglas. Las reglas se presentan en 4 grupos por cada corrida del algoritmo en las cuales se manipuló convenientemente el nivel de confianza, la cantidad de atributos y el número de reglas para una mejor generación de éstas.

Podemos inferir de estas asociaciones en el caso del:

- Primer y Segundo grupo:

  Al ser reglas con confianza 0.99 o 1 puede asegurarse que involucran la totalidad de los atributos que asocian. Podría asegurarse entonces que los proyectos tenían todos un Jefe de Proyecto, pero no tenían Prestador de servicios, ni Jefe de Operaciones. La técnica grupal de captura de requisitos FocusGroup no era utilizada.

- Tercer grupo:

  El control de los documentos internos y entregables se llevaba y almacenaba en un repositorio de datos.

- Cuarto grupo:

  Los que no utilizan IDEF comotécnica de modelado de negocio o el Voto acumulativo como técnica de priorización de requisitos tampoco usan FocusGroup, pero como FocusGroup es raramente utilizada, este no puede considerarse como un patrón en el juego de datos.

2.5.5 Modelación de los datos pertenecientes al diagnóstico realizado en 2012. Selección de atributos con CfsSubsetEval y BestFirst

Objetivo de esta técnica: Determinar los atributos que más influyen en la certificación.

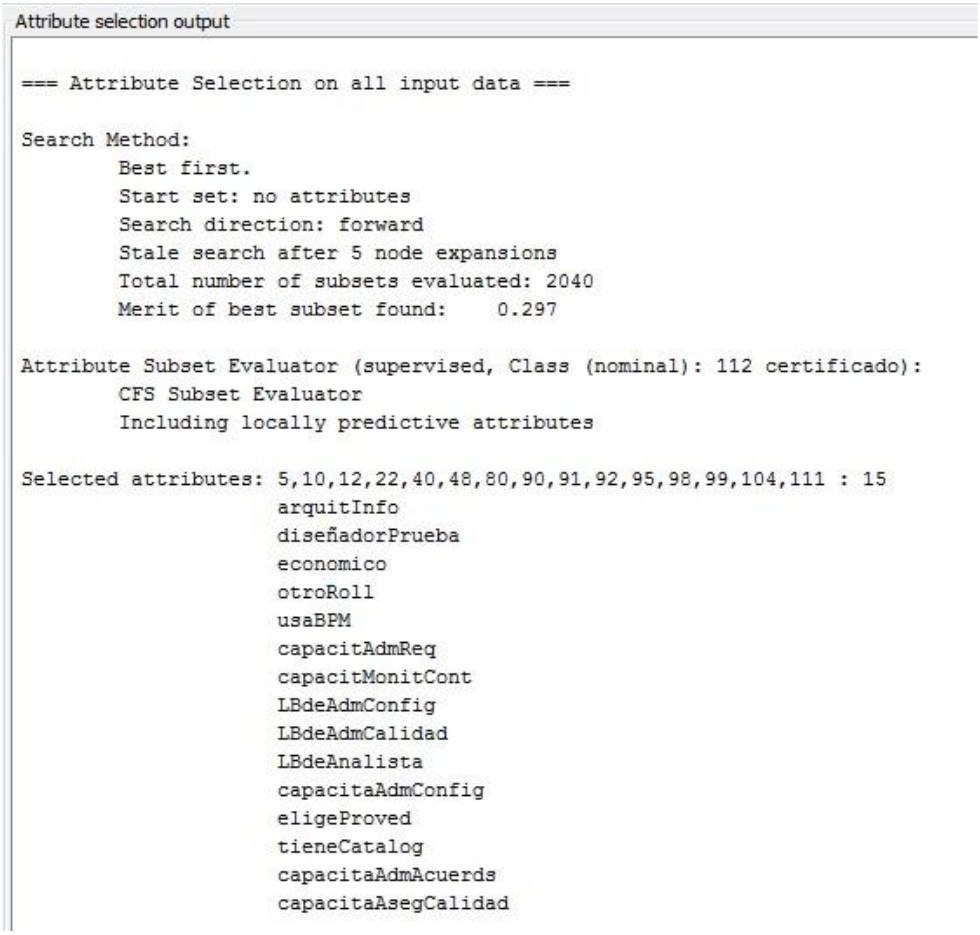


Figura 18. Salida de la técnica de selección de atributos (Fuente de elaboración: propia).

Los atributos escogidos corresponden a las preguntas:

- ¿Qué roles se desempeñan en el proyecto? [Arquitecto de Información]
- ¿Qué roles se desempeñan en el proyecto? [Diseñador de Pruebas].
- ¿Qué roles se desempeñan en el proyecto? [Económico].
- ¿Qué roles se desempeñan en el proyecto? [Otro].
- Marque cuáles técnicas de modelado de negocio utilizan en su proyecto: [BPM].
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Administración de requisitos?
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Monitoreo y Control de Proyecto?

- ¿En caso de generar líneas base quién las aprueba para ser liberadas? [Administrador de la configuración].
- ¿En caso de generar líneas base quién las aprueba para ser liberadas? [Administrador de la calidad].
- ¿En caso de generar líneas base quién las aprueba para ser liberadas? [Analista].
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Administración de la Configuración?
- ¿Se eligen a los proveedores basado en algún criterio para su selección?
- ¿El proyecto cuenta con algún catálogo de los posibles proveedores que pudieran brindarle productos o servicios?
- ¿Cómo obtuvieron la capacitación recibida? [administración de acuerdos con proveedores: Calisoft, el Centro, No han recibido]
- ¿Miembros de su proyecto han recibido capacitación sobre el proceso Aseguramiento de la calidad de los procesos y productos?

Interpretación:
Se pueden apreciar varios aspectos en la salida. De los 15 atributos mejor relacionados con la certificación, la 3ra parte la constituyen los atributos relacionados con las capacitaciones.

De los atributos relacionados al rol que aprueba las líneas base de los proyectos, la mitad fue escogida por el algoritmo. Se puede apreciar que de estos atributos los que mejor se relacionan con el atributo clase son aquellos proyectos en los que las líneas base son aprobadas por el analista, el administrador de la calidad y el administrador de la configuración, exclusiva o conjuntamente, contrastando esto con que en casi el 70% de los proyectos, el jefe de proyecto participa o individualmente aprueba las líneas base.

Se debe notar además que los atributos que identifican los roles que aprueban las líneas base, pertenecen a las áreas señaladas por los atributos de capacitación seleccionados.

2.5.6 Clasificación con OneR

Objetivo de esta técnica: Determinar el atributo que más influye en la certificación convirtiéndolo en una regla que explique la relación.

```
Scheme:weka.classifiers.rules.OneR -B 6
Relation: h1-weka.filters.unsupervised.attribute.Remove-R1-2
Instances:145
Attributes:112
[list of attributes omitted]
Test mode:10-fold cross-validation

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

capacitaAdmAcuerds:
        No      -> nop
        elProyecto      -> nop
        Calisoft        -> nop
        elCentro        -> sip
(125/145 instances correct)


Time taken to build model: 0.01seconds
```

Figura 19. Salida del algoritmo OneR (Fuente de elaboración: propia).

Guía para la interpretación:

En el extremo superior izquierdo se muestra el atributo seleccionado por el algoritmo. A continuación se muestran los valores que puede tomar y como influye en la certificación.

Interpretación:

El algoritmo tuvo una precisión del 86.21% clasificando incorrectamente 20 instancias, aproximadamente un 13.80% del total. Los algoritmos clasifican incorrectamente las instancias que se encuentran en la frontera de decisión, por lo que no corresponden a una u otra clasificación. Lo cual hace que las instancias clasificadas correctamente sean absolutamente confiables para deducir un patrón o clasificar ejemplos posteriores.

OneR ratifica que la capacitación que se recibe en los proyectos resulta esencial para la correcta implementación de las buenas prácticas de CMMI.

Gráficamente se puede observar la distribución de este atributo con respecto a sus valores:

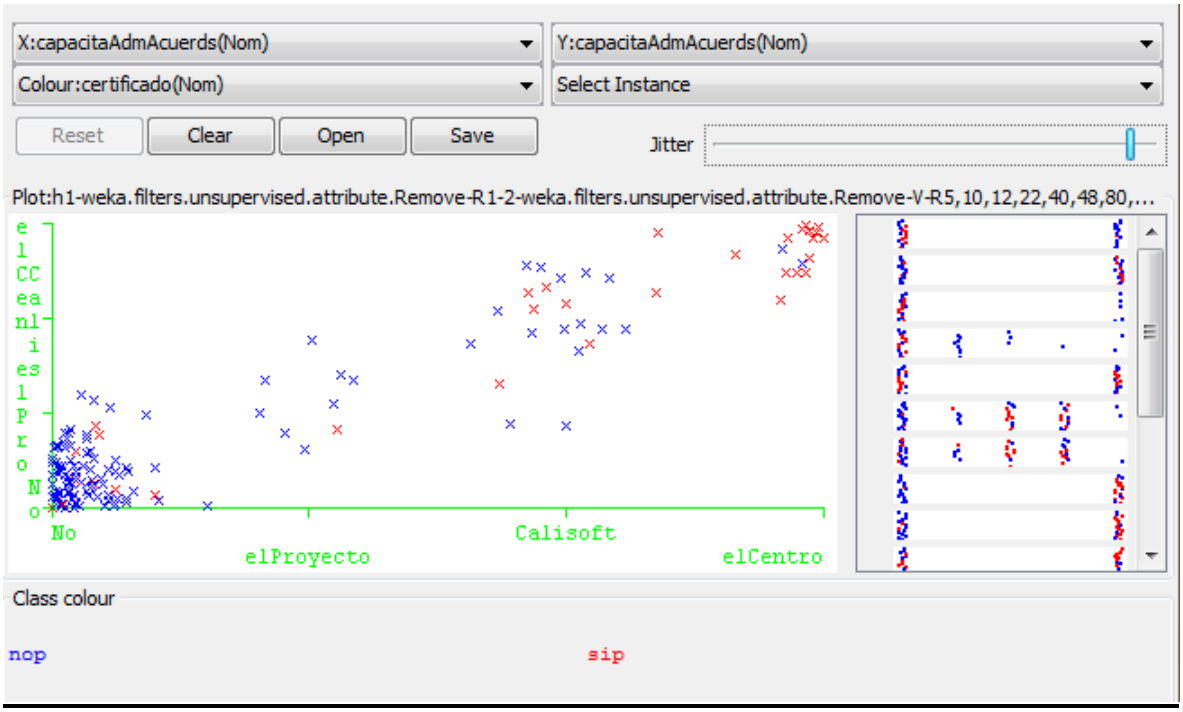


Figura 20. Gráfica de dispersión CapacitaAdmAcuerds (Fuente de elaboración: propia).

Guía para la interpretación de la gráfica:

Ambos ejes de la gráfica representan los valores que puede tomar el atributo correspondiente al origen de la capacitación en Administración de Acuerdos con Proveedores. En este caso se muestran los proyectos, evaluados y no evaluados que han recibido esta capacitación a través de Calisoft, del centro, del mismo proyecto, o no la han recibido hasta el momento.

Un resultado ideal sería que todos los proyectos hubieran recibido su capacitación por Calisoft o por el centro al que pertenecieran.

Interpretación:

La mayoría de los proyectos no han recibido capacitación sobre el proceso Administración de Acuerdos con Proveedores. Se puede apreciar también que los proyectos certificados no han recibido la capacitación por una única vía sin que esto signifique que la recibieron por varios canales. Aún más existe alrededor de una decena de proyectos que a más de un año de certificación de nivel 2 de CMMI no han recibido esta capacitación por ninguna vía.

2.5.7 Clasificación con Prism

Objetivo de la técnica: Obtener reglas que expliquen un atributo específico a partir de condiciones generadas por otros atributos.

De la selección de atributos se encontró destacada la coincidencia entre los roles y áreas de proceso, en cuanto al establecimiento de las líneas base y la capacitación. La generación de reglas estará encaminada a la investigación de algunos de estos atributos.

Tabla 2. Conjunto de reglas generadas por weka para las clases LBdeAdmCalidad y capacitaAsegCalidad (Fuente de elaboración: propia).

Reglas de PRISM: LBdeAdmCalidad.	Reglas de PRISM: capacitaAsegCalidad.
1. If expediente = migracion then No 2. If expediente = propio then No 3. If comoEstima = Empírico then No 4. If comoEstima = COCOMOII then No 5. If comoEstima = analogía and usaCuestionario = Sí then No 6. If documNecesids = No and aplicaListaVerifCalidad = No then No 7. If capacitPlaneacProyect = gestionPersonal then Sí 8. If comoEstima = experiencia then Sí 9. If comoEstima = ninguno then Sí 10. If LBdeAnalista = Sí and comoEstima = Sí then Sí 11. If AdmCalidadenOtroRol = No and capacitAdmReq = Calisoft then Sí	1. If expediente = propio then No 2. If expediente = migracion then No 3. If expediente = ExpServicio then No 4. If capacitAdmReq = Gestiónpersonal then No 5. If comoEstima = UCI then No 6. If comoEstima = COCOMOII then No 7. If expediente = xp then No 8. If comoEstima = expertos then No 9. If capacitMonitCont = Gestiónpersonal then No 10. If capacitaAdmConfig = No and tieneCatalog = No then No 11. If capacitaAdmAcuerds = No and arquitSoft = Sí then No 12. If capacitaAdmAcuerds = No and llevaTrazasAdelant = No then No 13. If capacitaAdmConfig = No and llevaTrazasAtras = No then No 14. If capacitaAdmAcuerds = No and capacitPlaneacProyect = elCentro then No 15. If capacitMedAnal = elProyecto then Sí 16. If comoEstima = analogia then Sí 17. If capacitaAdmAcuerds = Calisoft and capacitaAdmConfig = Calisoft then Sí 18. If capacitaAdmConfig = Calisoft and AdmCalidad = No then Sí 19. If capacitMonitCont = elCentro and ubicaReqSubsist = Sí then Sí

Guía para la interpretación:

Cada regla consta de dos partes: A la izquierda de la palabra *then* se encuentra la condición en forma de uno o más pares de atributo-valor y a la derecha se encuentra la conclusión en forma del valor que toma el atributo para el cual se han generado las reglas.

Interpretación:

Se pueden observar las reglas generadas por PRISM para el atributo relacionado con la pregunta *"¿En caso de generar Líneas Base éstas son aprobadas por el Administrador de la Calidad?"*, y las reglas referentes a la capacitación recibida en el área a la que pertenece este rol: Aseguramiento de la Calidad. Atributos clase respectivamente: *LBdeAdmCalidad* y *capacitaAsegCalidad.*

Es apreciable que las Líneas Base (LB) no las aprueba del Administrador de la Calidad cuando utilizan como expediente de proyecto uno propio o el de migración. Esto se cumple también cuando utilizan como estimación un método empírico, COCOMO, o estiman por analogía. Sin embargo se puede ver en el lado derecho de la tabla como los proyectos que utilizan estos expedientes y métodos no han recibido capacitación en el Aseguramiento de la Calidad.

Los que se han capacitado a través de gestiones personales, estiman por experiencia o no utilizan ningún método de estimación le han otorgado al Administrador de la Calidad la responsabilidad de aprobar las líneas.

Tabla 3. Conjunto de reglas generadas por weka para las clases LBdeAdmConfig y capacitaAdmConfig (Fuente de elaboración: propia).

Reglas de PRISM:LBdeAdmConfig.	Reglas de PRISM:capacitaAdmConfig.
1. If expediente = SXP then No 2. If expediente = migracion then No 3. If expediente = ExpServicio then No 4. If capacitAdmReq = Gestiónpersonal then No 5. If comoEstima = Empírico then No 6. If comoEstima = ninguno then No 7. If comoEstima = COCOMOII then No 8. If capacitMonitCont = Gestiónpersonal then No 9. If comoEstima = analogia then Sí 10. If comoEstima = UCI then Sí 11. If comoEstima = experiencia then Sí 12. If AdmCalidadenOtroRol = No and CreaDicciondDatos = Sí and revisionesHitoCliente = Sí then Sí	1. If expediente = migracion then No 2. If capacitAdmReq = Gestiónpersonal then No 3. If comoEstima = UCI then No 4. If comoEstima = COCOMOII then No 5. If expediente = xp then No 6. If comoEstima = expertos then No 7. If capacitMonitCont = Gestiónpersonal then No 8. If estandarNombres = No and capacitaAdmAcuerds = No then No 9. If capacitAdmReq = No and asesorCalid = Sí then No 10. If AdmCalidadenOtroRol = Nohay and capacitAdmReq = No then No 11. If capacitPlaneacProyect = No and identifRiesgos = No then No 12. If capacitaAdmAcuerds = Calisoft and capacitaAsegCalidad = Sí then Calisoft 13. If capacitaAdmAcuerds = elCentro 14. and capacitMedAnal = elCentro then elCentro

	15. If capacitAdmReq = elCentro and asesorCalid = Sí then elCentro 16. If capacitPlaneacProyect = elCentro and usaBPM = Sí then elCentro

Interpretación:

En este caso se analiza si es el Administrador de la Configuración (AC), quien aprueba las LB dentro de los proyectos, en contraste de si han recibido esos mismos proyectos capacitación en este aspecto. Es observable que los que se han capacitado mediante gestión personal en otras dos áreas de proceso diferentes no asocian esta responsabilidad al AC, sin embargo quienes se capacitaron de esta forma no se han capacitado en absoluto para el área en cuestión.

Por otra parte los proyectos que estiman por analogía, mediante método UCI o por experiencia coinciden en que es el AC quien aprueba las líneas. Los proyectos reciben las capacitaciones por una misma área: a través de Calisoft, o su propio centro, incluso el proyecto. No se ha establecido por tanto un organismo fijo para capacitar los proyectos, ni actúan varios al mismo tiempo sobre los capacitados para hacer más efectiva esta actividad.

Tabla 4. Conjunto de reglas generadas por weka para las clases LBdeAnalista y capacitAdmReq (Fuente de elaboración: propia).

Reglas de PRISM:LBdeAnalista.	Reglas de PRISM:capacitAdmReq.
1. If comoEstima = analogia then No 2. If expediente = migracion then No 3. If expediente = ExpServicio then No 4. If especifReqenDoc = No then No 5. If capacitAdmReq = Gestiónpersonal then No 6. If comoEstima = ninguno then No 7. If comoEstima = COCOMOII then No 8. If comoEstima = UCI and AdmCalidad = Sí then No 9. If comoEstima = UCI and AdmCalidad = No then Sí 10. If certificado = sip and hayDesviaciones = No then Sí	1. If expediente = propio then No 2. If comoEstima = UCI then No 3. If comoEstima = COCOMOII then No 4. If expediente = xp then No 5. If expediente = ninguno then No 6. If expediente = exp2.0 then No 7. If comoEstima = expertos then No 8. If capacitMonitCont = No and AdmCalidadenOtroRol = Nohay then No 9. If gestorConoc = Sí and pruebaEltos = No then elCentro 10. If comoEstima = Empírico and AdmCalidad = Sí then elCentro 11. If comoEstima = AlBET and servicio = No then elCentro 12. If capacitaAdmAcuerds = Calisoft and expediente = 3.3-PM then Calisoft 13. If capacitPlaneacProyect = Calisoft and comoEstima = No then Calisoft

Interpretación:

En la tabla 4 observamos que ocurre como en la tabla 1: los proyectos donde el analista es quien aprueba las LB son proyectos que no han recibido capacitación en el área que especifica las responsabilidades y habilidades de ese rol. Igual ocurre que la capacitación, en los casos en que existe, proviene de un mismo origen. No existe nunca más de un nivel capacitando los proyectos.

2.5.8 Reglas de asociación.

Objetivo de la técnica: Obtener patrones que forman los atributos mediante reglas.

1. jefeProyecto=Sí planeaENI=No revisionesAG=Sí 131 ==> controlProdyEntreg=Sí 131 conf:(1)

2. revisionesAG=Sí 136 ==> gestorConoc=No 131 conf:(0.96)

2.1. gestorConoc=No 139 ==> ingProcesos=No 133 conf:(0.96)

2.2. gestorConoc=No 139 ==> usaIDEF=No 133 conf:(0.96)

2.3. gestorConoc=No 139 ==> revisionesAG=Sí 131 conf:(0.94)

2.4. usaIDEF=No 139 ==> ingProcesos=No 133 conf:(0.96)

3. analizaRevs=Sí 124 ==> sigueNCdeRevs=Sí 123 conf:(0.99)

3.1. revInternas=Sí 126 ==> sigueNCdeRevs=Sí 123 conf:(0.98)

3.2. sigueNCdeRevs=Sí 127 ==> revInternas=Sí 123 conf:(0.97)

3.3. sigueNCdeRevs=Sí 127 ==> analizaRevs=Sí 123 conf:(0.97)

4. usaFocusGroup=No 135 ==> ingProcesos=No 130 conf:(0.96)

4.1. especifReqenDoc=Sí 133 ==> ingProcesos=No 128 conf:(0.96)

4.2. especifReqenDoc=Sí 133 ==> usaIDEF=No 128 conf:(0.96)

4.3. ctrlCambios=Sí 132 ==> ingProcesos=No 127 conf:(0.96)

4.4. analista=Sí 130 ==> especifReqenDoc=Sí 125 conf:(0.96)

5. capacitaAsegCalidad=No 96 ==> capacitaAdmAcuerds=No 93 conf:(0.97)

5.1. capacitaAdmAcuerds=No 97 ==> capacitaAsegCalidad=No 93 conf:(0.96)

5.2. capacitaAsegCalidad=No certificado=nop 87 ==> capacitaAdmAcuerds=No 86 conf:(0.99)

5.3. capacitMonitCont=No capacitaAdmAcuerds=No 83 ==> capacitaAsegCalidad=No 81 conf:(0.98)

5.4. capacitaAdmConfig=No 84 ==> capacitaAdmAcuerds=No 81 conf:(0.96)

5.5. capacitaAdmConfig=No 84 ==> capacitaAsegCalidad=No 81 conf:(0.96)

Guía para la interpretación:
Cada regla tiene una parte a la izquierda de la flecha y una parte a su derecha.
A la izquierda puede existir más de una pareja atributo-valor seguido por el número de casos que cubre. A la derecha habrá una pareja atributo-valor que es la consecuencia de la regla, seguida por su número de casos.
Por último se muestra el porciento de confiabilidad de dicha regla.

De las reglas de asociación que se manifiestan en este conjunto de datos, se han desechado las que corresponden a relaciones obvias dentro de los datos por ejemplo:*"tieneReposit=Sí 136 ==> guardaenReposit=Sí 135 conf:(0.99)"*, la cual se puede interpretar como: *Los proyectos que tienen un repositorio de datos, guardan la información en este repositorio*.

Así mismo se volvió sobre el preprocesamiento de los datos para eliminar los atributos que provocaban repeticiones innecesarias de una regla o asociaciones obvias, lo cual es consistente con multidireccionalidad y flexibilidad dela metodología CRISP-DM, la cual permite retornar de la fase de modelado a la de preparación de los datos.

La implementación de Apriori en WEKA permite configurar ciertos parámetros como la confianza que se puede tener en cada regla, o cuántas reglas se desean observar en cada

corrida del algoritmo. Para una mayor generación de reglas se configuró la confianza al límite del 80% y la cantidad de reglas a 20.

Se han seleccionado en cada corrida las reglas con mayor nivel de confianza y que contengan la mayor cantidad de instancias dentro de la regla. Se han separado en 5 grupos según los patrones de información que brindan en conjunto.

Interpretación:

Del primer grupo:

Los proyectos que tienen Jefe de Proyecto y no dejan de hacer el Plan de Proyecto y realizan revisiones de avance con la Alta Gerencia, llevan un control de los productos internos y entregables del proyecto.

Del segundo grupo:

Los proyectos que realizan revisiones de avance con la Alta Gerencia, no tienen el rolGestor de Conocimiento en el proyecto, no tienen Ingeniero de Procesos, ni usan IDEF como su técnica de modelado de negocio en el proyecto.

Del tercer grupo:

En los proyectos donde se analizan los resultados de las auditorías y revisiones realizadas, se les da seguimiento a las no conformidades detectadas. En estos proyectos se realizan también revisiones internas.

Del cuarto grupo:

Los proyectos que cuentan con analista controlan y documentan los cambios y especifican los requisitos de software a través del documento “Especificación de Requisitos de Software” u otro que cumpla la misma función, pero no los capturan a través de FocusGroup, no tienen Ingeniero de Procesos, ni usan IDEF como su técnica de modelado de negocio en el proyecto.

Del quinto grupo:

En los proyectos pertenecientes a centros sin certificar que no han recibido capacitación en el área de Aseguramiento de la Calidad, tampoco han recibido capacitación en Administración de Acuerdos con Proveedores. Los proyectos que no han recibido capacitación en Aseguramiento de la Calidad, tampoco han recibido capacitación en las áreas de Monitoreo y Control de Proyecto y Administración de Acuerdos con Proveedores.

2.6 Conclusiones parciales

- Tanto los algoritmos predictivos como los descriptivos generaron reglas que permitieron identificar patrones y correlaciones entre los atributos.
- De 2010 a 2012 se mantuvo la capacitación como la actividad más afectada por la evaluación de CMMI.
- Los proyectos fuera del Programa de Mejora no se capacitaron.
- La falta de capacitación afecta la asignación de roles y tareas.
- Los roles poco comunes y las técnicas de recopilación de datos poco usadas en 2010 no desaparecieron en 2012, pero su uso y utilidad no aumentaron tampoco.
- Los roles y responsabilidades poco comunes pertenecen a proyectos que no se han revisado por parte de la Alta Gerencia.

Capítulo 3.Validación de los resultados.

3.1. Introducción

En el siguiente capítulo se realiza la validación de los patrones obtenidos al aplicar técnicas de minería de datos a los datos recogidos en los diagnósticos. Se muestran los resultados obtenidos al aplicar encuestas y entrevistas a un comité de expertos y a la alta gerencia.

3.2. Modelos entregables

Después de aplicar cada técnica de minería de datos se construyó un modelo por cada técnica que se entregará junto a una guía general de interpretación para facilitar la comprensión y cualquier posterior análisis de los resultados.

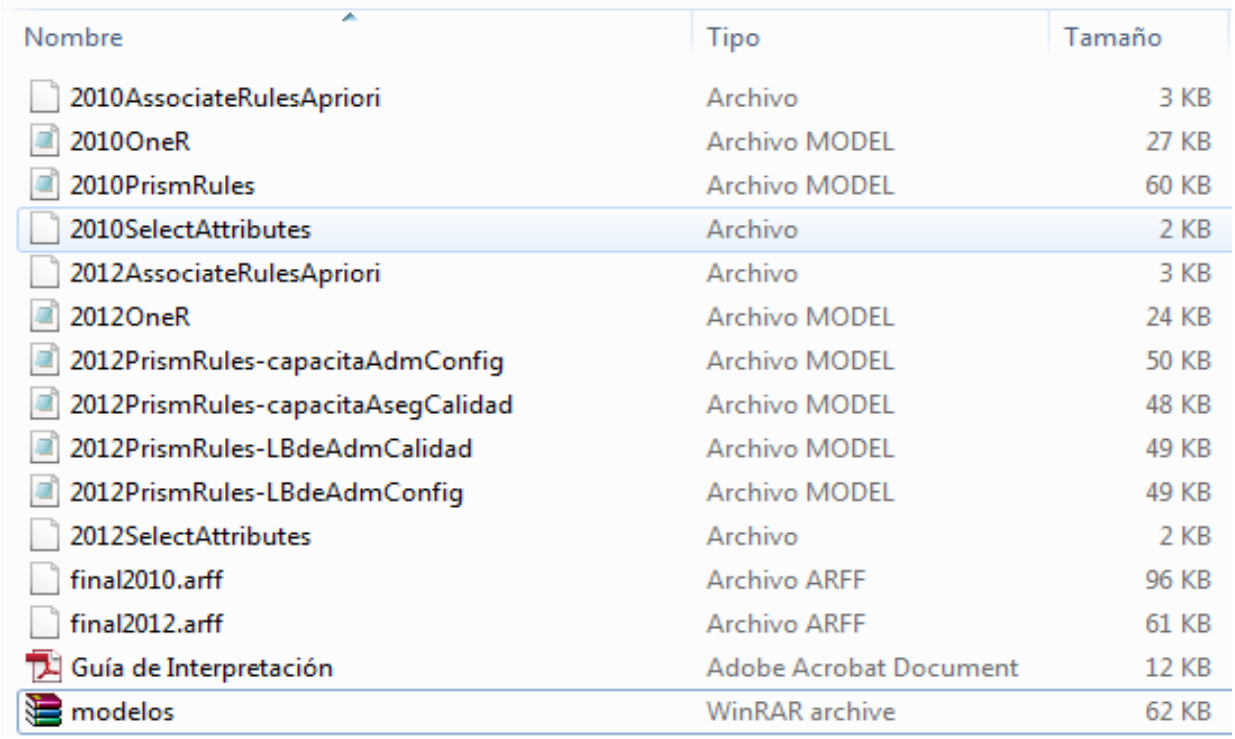


Figura 21. Modelos construidos que contienen los patrones. (Fuente de elaboración: propia).

En el caso de los modelos entrenados que pueden utilizarse para clasificar otros juegos de datos, como PRISM y OneR, en el futuro será necesario visualizarlos en la herramienta WEKA.

3.3. Método Comité de Expertos sobre validación de los resultados.

Para realizar la validación de los resultados se formó un comité de expertos de CALISOFT y otras áreas. Los expertos debían tener experiencia en la realización de auditorías,

diagnósticos, revisiones, consultorías y CMMI. No necesariamente especializados en las siete áreas de proceso de nivel 2, pero con conocimiento general sobre todas o la mayoría de ellas.

Se encuestó además a la Alta Gerencia de la universidad como principal medidor del aporte de los resultados a la toma de decisiones.
La selección de los expertos se basa en las recomendaciones de los especialistas en este método de validación que plantean hacer coincidir el interés de los expertos con el tema de estudio de tal manera que sus criterios sean significativos (ANDRANOVICH 1995).
Esta forma de validación ha sido utilizada ampliamente en numerosos estudios y ámbitos del conocimiento (HUNG *et al.* 2008).

3.3.1. Selección de los expertos.

La experiencia en el objeto de estudio es importante para tomar en cuenta los criterios de los expertos encuestados. De una planificación inicial de 11 expertos, se redujo la cantidad a 7, ya que dos se encontraban fuera de la universidad, uno no contaba con experiencia suficiente en cuanto a años de experiencia y áreas de conocimiento y otro se negó a participar.

Se puede observar en la figura 22 que la mayor parte de los expertos encuestados tienen entre 4 y 5 años de experiencia.

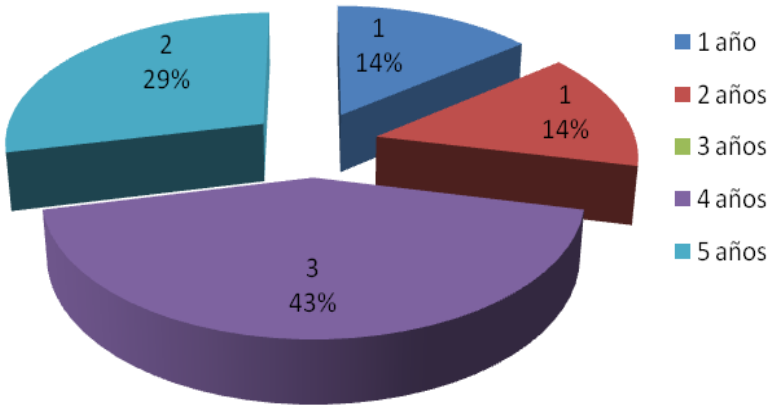


Figura 22. Muestra la composición del comité de expertos según los años de experiencia en el área de conocimiento de la calidad de software(Fuente de elaboración: propia).

En la gráfica a continuación se muestran las área de conocimiento que dominaban los expertos. Se seleccionaron especialistas con un dominio amplio y que se extendieran a casi todas las áreas que se relacionaban en los resultados.

Se puede observar en la figura 23 que los expertos dominaban casi todas las áreas de conocimiento que eran importantes para obtener criterios significativos en las encuestas.

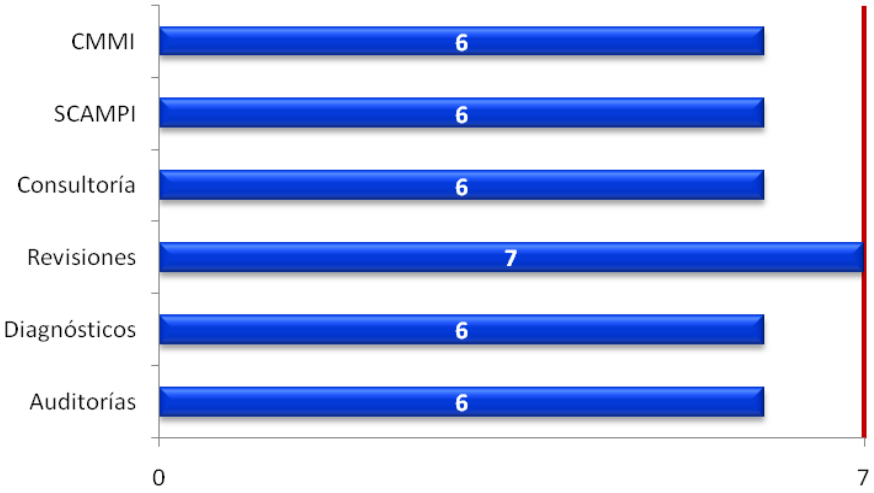


Figura 23. Muestra la composición del comité de expertos según el área de conocimiento de la calidad de software en que trabajan

El 100% de los expertos coincidió en que hasta ahora no se le habían aplicado técnicas de minería a los datos del diagnóstico. Así mismo señalaron todos que sería útil a la toma de decisiones el resultado de la aplicación de dichas técnicas sobre estos datos.

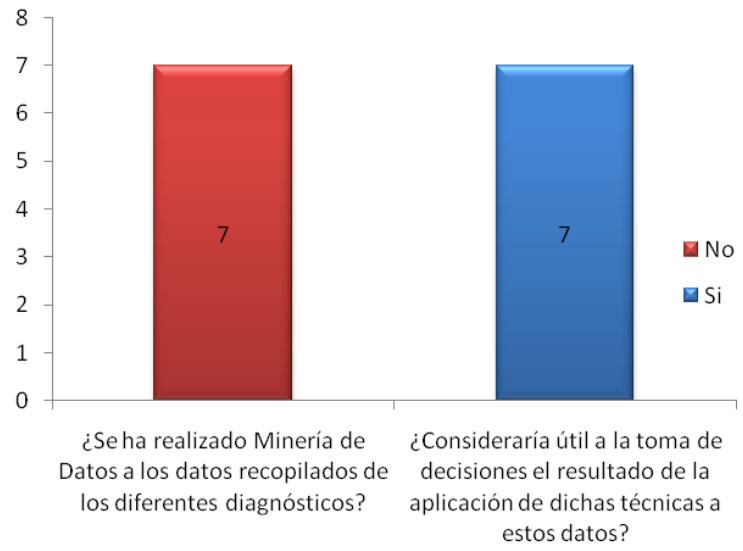


Figura 24. Muestra los resultados obtenidos en cuanto a la realización de ténicas de minería de datos a los datos del diagnóstico y la utilidad de esos resultados a la toma de decisiones (Fuente de elaboración: propia).

Antes de que los expertos respondieran la próxima pregunta se les presentó el resultado del trabajo para ser considerado en cuanto a su experiencia e interés en los mismos.

En el caso de ponderar de 1 a 5 la necesidad que le otorgarían a procesar estos datos mediante la aplicación de las técnicas seleccionadas, siendo 1 poco útil y siendo 5 muy útil, y considerando cualquier promedio superior a 4,5 como muy útil, al final se obtuvieron 4,71 puntos de ponderación como promedio. Lo consideraron útil 2 de los encuestados, los otros 5 expertos la consideraron muy útil. Las mismas cifras se presentaron para la ponderación de la necesidad, según el criterio de los expertos. Dos la consideraron necesaria y cinco la consideraron como muy necesaria.

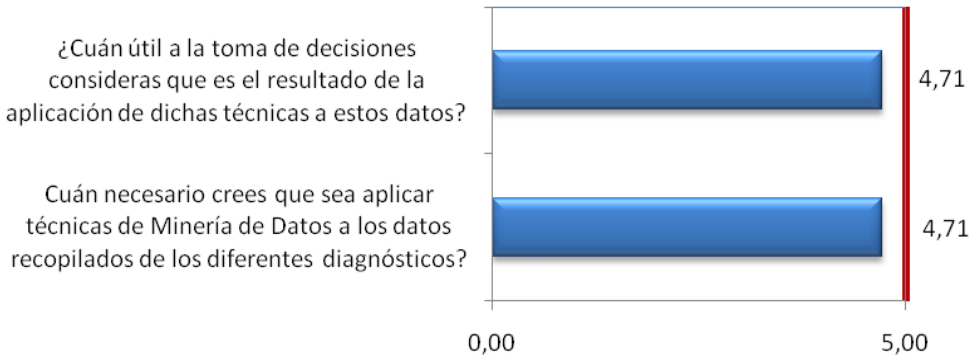


Figura 25. Muestrala ponderación de los expertos en cuanto a necesidad y utilidad de los resultados(Fuente de elaboración: propia).

A continuación los expertos tenían la opción de seleccionar si consideraban que los resultados confirmaban la información ya conocida a partir de los diagnósticos, auditorías, revisiones y otras, si brindaban nueva información o si aportaban elementos a la toma de decisiones.

Pudiendo seleccionar más de una opción, 5 expertos seleccionaron las opciones correspondientes a nueva información y aportar elementos a la toma de decisiones. Los dos expertos restantes, seleccionaron las tres opciones como válidas para estos resultados.

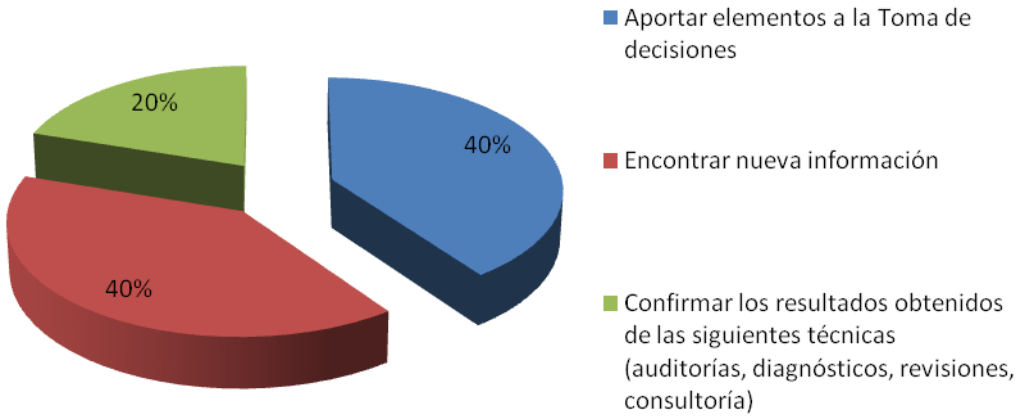


Figura 26 Aporte de los resultados según criterio de expertos (Fuente de elaboración: propia).

3.3.2. Encuesta a realizar a los expertos.

La encuesta dirigida al comité de expertos estuvo formada por 7 criterios que a continuación se listan. Ver anexo 1 Encuesta al comité de expertos.

- Años de experiencia.
- Área de conocimiento.
- Realización de la minería de datos a los datos recopilados de los diferentes diagnósticos.
- Necesidad de aplicar técnicas de minería de datos a los datos de los diagnósticos.
- Utilidad a la toma de decisiones del resultado de las técnicas de minería de dato.
- Cuán útil es a la toma de decisiones el resultado de las técnicas de minería de dato.
- Uso de los resultados obtenidos luego de aplicar alguna técnica de minería de datos.

3.3.3. Encuesta a la Alta Gerencia sobre aporte a la toma de decisiones.

La encuesta dirigida a la alta gerencia estuvo formada por 7 criterios que a continuación se listan.Ver anexo 2 encuesta a la alta gerencia.

- Necesidad de realizar acciones adicionales a las ya establecidas que reflejen el estado productivo de la UCI para la toma de decisiones
- Utilidad de los datos recopilados anualmente por el diagnóstico para apoyar la toma de decisiones
- Considera la obtención de patrones a través de la aplicación de técnicas de minería de datos como una alternativa: útil, interesante, innecesaria, se debería incorporar cada año en lo adelante.
- Utilidad de los resultados obtenidos.

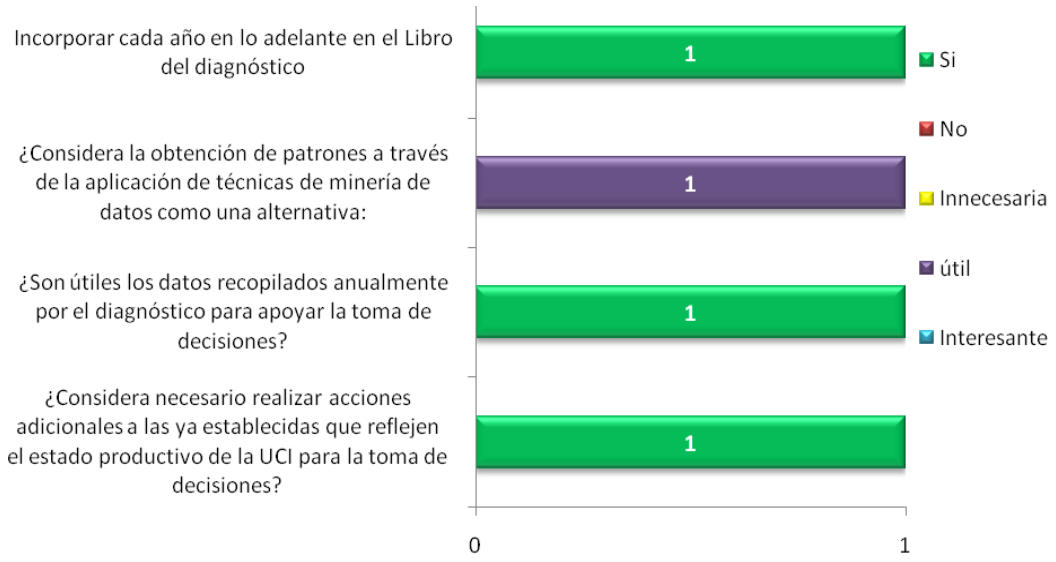


Figura 27. Muestra las respuestas de la Alta Gerencia después de analizar los resultados (Fuente de elaboración: propia).

Como puede apreciarse la Alta Gerencia juzgó conveniente incorporar esta alternativa cada año en lo adelante al analizar los datos recopilados durante el diagnóstico. Consideró además necesario este trabajo ya que consittuye una opción adicional muy útil para la toma de decisiones en la universidad.

3.3.4. Valoración del aporte del diagnóstico con mediante los resultados de la minería de datos.

Como resultado de la realización de las entrevistas a la Alta Gerencia y al comité de expertos se listan las siguientes valoraciones adicionales que aportaron los encuestados:

- *Esos resultados son muy buenos y no son posibles de obtener mediante los métodos que conocen y utilizan actualmente.*
- *Estos resultados deberían incluirse en el libro del diagnóstico ya que no tienen manera de obtenerlos manualmente y van más allá de lo que actualmente realizan.*
- *Esto es necesario para aumentar la rigurosidad en la toma de decisiones y para aumentar la comprensión de estas decisiones por todos los implicados.*
- *Esta acción adicional es necesaria para la toma de decisiones ya que las que se realizan actualmente son siempre insuficientes.*

- *Estos resultados pueden ser utilizados para la definición de las estrategias para el próximo periodo para definir indicadores para la toma de decisiones de la alta gerencia y para mejorar los procesos y redefinir roles y tareas.*

3.4. Conclusiones parciales

- El comité de expertos se conformó con especialistas con un alto grado de especialización en las áreas de conocimiento relacionadas con los resultados, así como varios años de experiencia en las mismas.
- Los resultados fueron valorados como muy útiles por los expertos.
- Según el comité de expertos y la Alta Gerencia los resultados aportan elementos a la toma de decisiones.
- La Alta Gerencia consideró que se debería incluir esta alternativa cada año en lo adelante.

Conclusiones

1. El marco teórico elaborado permitiófundamentar y elaborar teóricamente los conceptos relacionados con la investigación.
2. La herramienta WEKA y la metodologíaCRISP-DM permitieron llevar a cabo el procesamiento de los datos sobre el programa de mejora de los diagnósticos 2010 y 2012.
3. Los diagnósticos analizados generaron muchos datos a los cuales no se les realizaban posteriores análisis más profundos en busca de nueva información.
4. La aplicación de las técnicas de minería de datos seleccionadas permitieron:
    - Detectar problemas de capacitación en la mayoría de las áreas de proceso
    - Identificar roles y responsabilidades afectadas por la escasa capacitación realizada a los proyectos.
    - Encontrar patrones de información que caracterizan o caracterizaron el desarrollo productivo en los proyectos.
5. La validación de los resultados arrojó que los mismos tienen gran utilidad y son necesarios para adquirir nueva información y apoyar la toma de decisiones.

Recomendaciones

1. Realizar un análisis más amplio al juego de datos original incluyendo los valores numéricos y subjetivos.
2. Aplicar otras técnicas de minería de datos a los archivos *.arff* elaborados.

Referencias Bibliográficas

1. (CALISOFT), C. N. D. C. D. S. *Libro del Diagnóstico* CALISOFT, UCI, 2010. D-10-020.
2. AGRAWAL, R. and R. SRIKANT. *Fast Algorithms for Mining. Association Rules in Large Databases. . Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 1994. 1215:478-499.
3. ANDRANOVICH, G. *Developing community participation and consensus: The Delphi technique*. Los Ángeles, Department of Political Science, California State University, 1995.11.
4. ANGULO, J. Á. V. *CREACIÓN DE MODELOS DE PREDICCIÓN ORIENTADOS A LAS APUESTAS EN EVENTOS DEPORTIVOS.*: ESCUELA POLITÉCNICA SUPERIOR Madrid, UNIVERSIDAD CARLOS III DE MADRID, 2010. p.
5. BÁRCENA, A. and A. PRADO, Eds. *La inversión extranjera directa en América Latina y el Caribe*Santiago de Chile, Junta de Publicaciones de las Naciones Unidas, 2011. 978-92-1-121783-4
6. BERRY, M. J. A. and G. LINOFF Data Mining Techniques for Marketing, Sales, and Customer Support, 1997.
7. BERTHOLD, M. *KNIME Desktop*, 2012. [2012]. Disponible en: http://www.knime.org/knime
8. CABENA, P.; P. HADJINIAN*, et al.*, Eds. *Discovering Data Mining From concept to implementation*. 1st, Prentice Hall, 1997. 224 9780137439805
9. CLAUDIA ETNA CARIGNANO, C. L. A. *Los desafíos de la gestión de los costos en el siglo XXI. Métodos de apoyo multicriterio a las decisiones*. *XXVIII congreso argentino, de profesores universitarios de costos. 21, 22 y 23 de septiembre de 2005.* Mendoza, Argentina, 2005.
10. CRISP-DM, C. *Metodología CRISP-DM para minería de datos*, Dataprix.com, 2007.
11. CZINKOTA, M. and M. KOTABE. *Administración de Mercadotecnia*. EDITORES, I. T., Cengage Learning Latin America, 2001.
12. CHIAVENATO, I. *Introducción a la Teoría General de la Administración*. INTERAMERICANA, M.-H., McGraw-Hill, 2006.
13. DORTMUND, U. D. *Rapid - I - RapidMiner*, 2001. [2012]. Disponible en: http://rapidminer.com/
14. DURÁN, M. G. *Proceso diagnóstico a la actividad productiva en la Universidad de las Ciencias Informáticas*. Grupo de auditoría y revisiones. CALISOFT. La Habana, Universidad de las Ciencias Informáticas, 2012.80. p.

15. ESPAÑOLES, I. D. A. *ANÁLISIS AVANZADOS DE GRANDES VOLÚMENES DE DATOS EN EL SECTOR SEGUROS (2ª PARTE)*, 2006. [2012]. Disponible en: http://www.actuarios.org/espa/revista25/datamining.htm
16. FAYYAD, U. M.; G. PIATETSKY-SHAPIRO*, et al.Advanced in Knowledge Discovery and Data Mining*. *MIT Press*, Menlo Park 1996.
17. FERRELL, O. and G. HIRT., Eds. *Introducción a los Negocios en un Mundo Cambiante*. 4ta, 2004. 540 9789701039427
18. GALVIS, M. and F. MARTÍNEZ. *Confrontación de dos técnicas de Minería de Datos aplicadas a un dominio específico.*: Facultad de Ingeniería. Bogotá, Colombia, Pontificia Universidad Javeriana, 2004.125. p.
19. GROUP, M. L. *Weka 3 - Data Mining with Open Source Machine Learning Software in Java* University of Waikato, 2012. [2012]. Disponible en: http://www.cs.waikato.ac.nz/~ml/weka/
20. HUNG, H. L.; J. W. ALTSCHULD*, et al.* Methodological and conceptual issues confronting a cross-country Delphi study of educational program evaluation. *Evaluation Services Center, University of Cincinnati*, 2008.
21. IEEE, C. S. *SWEBOK. Guide to the Software Engineering Body of Knowledge*, 2004.
22. INSTITUTE, C. M. S. E. *CMMI® For Development SCAMPISM Class "A" Appraisal Results 2009 End-Year Update*, Carnegie Mellon University, 2010.
23. JORGE, M. *Harvard desarrolla algoritmo para detectar patrones ocultos en conjuntos de datos inmensos*, Alt140.com, 2011.
24. KDNUGGETS. *Gráfica comparativa entre las herraminetas de minería de datos más usadas*, 2010. [2012]. Disponible en: http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html
25. LENGUA, R. A. D. L. *DICCIONARIO DE LA LENGUA ESPAÑOLA*, 22va. Real Academia de la Lengua, 2011. [2012]. Disponible en: http://lema.rae.es/drae/
26. LÓPEZ, J. M. M. and J. G. HERRERO. *Técnicas de análisis de datos. Aplicaciones prácticas utilizando Microsoft Excel y WEKA*, Universidad Carlos III de Madrid, 2006.
27. MANEIRO, M. Y. *Minería de Datos*. Buenos Aires, Argentina, Universidad Nacional del Nordeste, 2008.33.
28. MÉNDEZ, N. D. J. C. *MINERÍA DE DATOS UNA HERRAMIENTA PARA LA TOMA DE DECISIONES*. Facultad de Ingeniería Escuela de Ingeniería en Ciencias y Sistemas. Guatemala, Universidad de San Carlos de Guatemala, 2006.96. p.
29. MOINE, J. M.; A. S. HAEDO*, et al.Estudio comparativo de metodologías para minería de datos*, 2011.

30. MORALES, C. R. *APLICACIÓN DE TÉCNICAS DE ADQUISICIÓN DECONOCIMIENTO PARA LA MEJORA DE CURSOSHIPERMEDIA ADAPTATIVOS BASADOS EN WEB*. Departamento de Ciencias de la Computación eInteligencia Artificial. Granada, UNIVERSIDAD DE GRANADAE.T.S. DE INGENIERÍA INFORMÁTICA, 2003. p.
31. MUÑANTE, J. R. D. *Modelo de gestión del conocimiento (GC) aplicado a la universidad pública en el Perú*, 2004.
32. OVIEDO, U. D. I. T. I. D. *Fundamentos de Ingeniería del Software – Tercer curso de Ingeniería Técnica Informática.* , 2006.
33. PERESSON, L. *Sistemas de gestión de la calidad con enfoque al cliente*. Valladolid, Universidad de Valladolid, 2007.116. p.
34. PIATESKY-SHAPIRO, G. and W. J. FRAWLEY, Eds. *Knowledge Discovery in Databases*, MIT Press Cambridge 1991. 0262660709
35. R-PROYECT. *The R Project for Statistical Computing*, 2012. [2012]. Disponible en: http://www.r-project.org/
36. RODRÍGUEZ, J. P. F. and A. G. PÉREZ. *Aplicación de la minería de datos en la bioinformática*. *ACIMED*, ACIMED, 2002. 10,1024-9435:69-76.
37. SAS. *SAS Enterprise Miner*, 2012. [2012]. Disponible en: http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html
38. TIGRE, P. B. and F. S. MARQUES, Eds. *Desafíos y oportunidades de la industria del software en América Latina*. 1st Bogotá, Colombia, Ediciones Mayol, 2009. 978-958-8307-56-5

Bibliografía

1. (CALISOFT), C. N. D. C. D. S. *Libro del Diagnóstico* CALISOFT, UCI, 2010. D-10-020.
2. AGRAWAL, R. and R. SRIKANT. *Fast Algorithms for Mining. Association Rules in Large Databases. . Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 1994. 1215:478-499.
3. ANDRANOVICH, G. *Developing community participation and consensus: The Delphi technique*. Los Ángeles, Department of Political Science, California State University, 1995.11.
4. ANGULO, J. Á. V. *CREACIÓN DE MODELOS DE PREDICCIÓN ORIENTADOS A LAS APUESTAS EN EVENTOS DEPORTIVOS.*: ESCUELA POLITÉCNICA SUPERIOR Madrid, UNIVERSIDAD CARLOS III DE MADRID, 2010. p.
5. BÁRCENA, A. and A. PRADO, Eds. *La inversión extranjera directa en América Latina y el Caribe*Santiago de Chile, Junta de Publicaciones de las Naciones Unidas, 2011. 978-92-1-121783-4
6. BERRY, M. J. A. and G. LINOFF Data Mining Techniques for Marketing, Sales, and Customer Support, 1997.
7. BERTHOLD, M. *KNIME Desktop*, 2012. [2012]. Disponible en: http://www.knime.org/knime
8. CABENA, P.; P. HADJINIAN*, et al.*, Eds. *Discovering Data Mining From concept to implementation*. 1st, Prentice Hall, 1997. 224 9780137439805
9. CLAUDIA ETNA CARIGNANO, C. L. A. *Los desafíos de la gestión de los costos en el siglo XXI. Métodos de apoyo multicriterio a las decisiones*. *XXVIII congreso argentino, de profesores universitarios de costos. 21, 22 y 23 de septiembre de 2005.* Mendoza, Argentina, 2005.
10. CRISP-DM, C. *Metodología CRISP-DM para minería de datos*, Dataprix.com, 2007.
11. CZINKOTA, M. and M. KOTABE. *Administración de Mercadotecnia*. EDITORES, I. T., Cengage Learning Latin America, 2001.
12. CHIAVENATO, I. *Introducción a la Teoría General de la Administración*. INTERAMERICANA, M.-H., McGraw-Hill, 2006.
13. DORTMUND, U. D. *Rapid - I - RapidMiner*, 2001. [2012]. Disponible en: http://rapidminer.com/
14. DURÁN, M. G. *Proceso diagnóstico a la actividad productiva en la Universidad de las Ciencias Informáticas*. Grupo de auditoría y revisiones. CALISOFT. La Habana, Universidad de las Ciencias Informáticas, 2012.80. p.

15. ESPAÑOLES, I. D. A. *ANÁLISIS AVANZADOS DE GRANDES VOLÚMENES DE DATOS EN EL SECTOR SEGUROS (2ª PARTE)*, 2006. [2012]. Disponible en: http://www.actuarios.org/espa/revista25/datamining.htm
16. FAYYAD, U. M.; G. PIATETSKY-SHAPIRO*, et al.Advanced in Knowledge Discovery and Data Mining*. *MIT Press*, Menlo Park 1996.
17. FERRELL, O. and G. HIRT., Eds. *Introducción a los Negocios en un Mundo Cambiante*. 4ta, 2004. 540 9789701039427
18. GALVIS, M. and F. MARTÍNEZ. *Confrontación de dos técnicas de Minería de Datos aplicadas a un dominio específico.*: Facultad de Ingeniería. Bogotá, Colombia, Pontificia Universidad Javeriana, 2004.125. p.
19. GROUP, M. L. *Weka 3 - Data Mining with Open Source Machine Learning Software in Java* University of Waikato, 2012. [2012]. Disponible en: http://www.cs.waikato.ac.nz/~ml/weka/
20. HUNG, H. L.; J. W. ALTSCHULD*, et al.* Methodological and conceptual issues confronting a cross-country Delphi study of educational program evaluation. *Evaluation Services Center, University of Cincinnati*, 2008.
21. IEEE, C. S. *SWEBOK. Guide to the Software Engineering Body of Knowledge*, 2004.
22. INSTITUTE, C. M. S. E. *CMMI® For Development SCAMPISM Class "A" Appraisal Results 2009 End-Year Update*, Carnegie Mellon University, 2010.
23. JORGE, M. *Harvard desarrolla algoritmo para detectar patrones ocultos en conjuntos de datos inmensos*, Alt140.com, 2011.
24. KDNUGGETS. *Gráfica comparativa entre las herraminetas de minería de datos más usadas*, 2010. [2012]. Disponible en: http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html
25. LENGUA, R. A. D. L. *DICCIONARIO DE LA LENGUA ESPAÑOLA*, 22va. Real Academia de la Lengua, 2011. [2012]. Disponible en: http://lema.rae.es/drae/
26. LÓPEZ, J. M. M. and J. G. HERRERO. *Técnicas de análisis de datos. Aplicaciones prácticas utilizando Microsoft Excel y WEKA*, Universidad Carlos III de Madrid, 2006.
27. MANEIRO, M. Y. *Minería de Datos*. Buenos Aires, Argentina, Universidad Nacional del Nordeste, 2008.33.
28. MÉNDEZ, N. D. J. C. *MINERÍA DE DATOS UNA HERRAMIENTA PARA LA TOMA DE DECISIONES*. Facultad de Ingeniería Escuela de Ingeniería en Ciencias y Sistemas. Guatemala, Universidad de San Carlos de Guatemala, 2006.96. p.
29. MOINE, J. M.; A. S. HAEDO*, et al.Estudio comparativo de metodologías para minería de datos*, 2011.

30. MORALES, C. R. *APLICACIÓN DE TÉCNICAS DE ADQUISICIÓN DECONOCIMIENTO PARA LA MEJORA DE CURSOSHIPERMEDIA ADAPTATIVOS BASADOS EN WEB*. Departamento de Ciencias de la Computación eInteligencia Artificial. Granada, UNIVERSIDAD DE GRANADAE.T.S. DE INGENIERÍA INFORMÁTICA, 2003. p.
31. MUÑANTE, J. R. D. *Modelo de gestión del conocimiento (GC) aplicado a la universidad pública en el Perú*, 2004.
32. OVIEDO, U. D. I. T. I. D. *Fundamentos de Ingeniería del Software – Tercer curso de Ingeniería Técnica Informática.* , 2006.
33. PERESSON, L. *Sistemas de gestión de la calidad con enfoque al cliente*. Valladolid, Universidad de Valladolid, 2007.116. p.
34. PIATESKY-SHAPIRO, G. and W. J. FRAWLEY, Eds. *Knowledge Discovery in Databases*, MIT Press Cambridge 1991. 0262660709
35. R-PROYECT. *The R Project for Statistical Computing*, 2012. [2012]. Disponible en: http://www.r-project.org/
36. RODRÍGUEZ, J. P. F. and A. G. PÉREZ. *Aplicación de la minería de datos en la bioinformática*. *ACIMED*, ACIMED, 2002. 10,1024-9435:69-76.
37. SAS. *SAS Enterprise Miner*, 2012. [2012]. Disponible en: http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html
38. TIGRE, P. B. and F. S. MARQUES, Eds. *Desafíos y oportunidades de la industria del software en América Latina*. 1st Bogotá, Colombia, Ediciones Mayol, 2009. 978-958-8307-56-5

Anexos

Anexo 1: Encuesta al comité de expertos

Área de conocimiento	Experiencia (Años)
Ej.: auditorías, diagnósticos, revisiones, consultoría, SCAMPI, CMMI (PP, PMC, PPQA, REQM, CM, SAM, MA)	

1-¿Se ha realizado Minería de datos a los datos recopilados de los diferentes diagnósticos?

Si: No:

2-¿Cuán necesario crees que sea aplicar técnicas de Minería de datos a los datos recopilados de los diferentes diagnósticos?

1	2	3	4	5

3-¿Consideraría útil a la toma de decisiones el resultado de la aplicación de dichas técnicas a estos datos?

Si: No:

4-¿Cuán útil a la toma de decisiones consideras que es el resultado de la aplicación de dichas técnicas a estos datos?

1	2	3	4	5

5-A continuación se muestra un resumen del resultado.

¿En qué cree usted que puedan ser usados los resultados obtenidos luego de aplicar alguna técnica de minería de datos?

_Confirmar los resultados obtenidos de las siguientes técnicas (auditorías, diagnósticos, revisiones, consultoría)

_Encontrar nueva información

_ Aportar elementos a la TD.

Anexo 2: Encuesta a la alta gerencia

1-¿Considera necesario realizar acciones adicionales a las ya establecidas que reflejen el estado productivo de la UCI para la toma de decisiones?

Si: No:

2- ¿Son útiles los datos recopilados anualmente por el diagnóstico para apoyar la toma de decisiones?

Si: No:

3-¿Considera la obtención de patrones a través de la aplicación de técnicas de minería de datos como una alternativa:

Innecesaria__ útil__ Interesante__

Se debería incorporar cada año en lo adelante___?

4-¿Cómo considera que pueden ser usados los resultados obtenidos?