ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística Informática

"Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio"

TESIS DE GRADO

Previo a la obtención del Título de

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentada por:

Daryna Marycruz Calderón Orozco

GUAYAQUIL - ECUADOR

AÑO 2005

AGRADECIMIENTO

A Dios.
A mi mamá por su apoyo, dedicación y sus sabios consejos.
A mis tíos Vinicio y Wilson por su ayuda incondicional y sus consejos.
A mi abuelita Amalia por su inmenso cariño y por la ayuda que siempre nos ha brindado a mi mamá y a mi.
A todas las personas que colaboraron en la realización de este trabajo, especialmente a la ayuda del Ing. Juan Alvarado Director de Tesis.

DEDICATORIA

A Dios.
A mi madre.
A mis tíos Vinicio y Wilson.
A mi abuelita Rosa Amalia.
Y a todos mis familiares.

TRIBUNAL DE GRADUACIÓN

Mat. Washington Armas
DIRECTOR DEL ICM

Ing. Juan Alvarado Ortega
DIRECTOR DE TESIS

Mat. Jhonny Bustamante R.
VOCAL

Ing. Oscar Mendoza Macias
VOCAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL).

_____

Daryna Marycruz Calderón Orozco

RESUMEN

El presente trabajo es desarrollado con el fin de encontrar las causas por las cuales empleados de una empresa de servicios realiza sobretiempo, para esto se utilizó el Método del Árbol de Decisión, de las técnicas de este método se aplicó el Algoritmo C4.5 utilizando el programa WEKA. Este programa implementa el Algoritmo J4.8, el cual es una extensión del Algoritmo C4.5.

El primer capítulo presenta una revisión rápida de los conceptos más importantes de horas normales y horas extras o sobretiempo, según el código de trabajo.

El segundo capítulo se explica los diferentes métodos para extraer patrones de datos utilizando el proceso de descubrimiento de conocimiento en base de datos.

El tercer capítulo presenta el modelo propuesto para extraer información del comportamiento del sobretiempo, para lo cual se utilizo el Algoritmo J4.8 implementado por el programa WEKA.

Finalmente, en el cuarto capítulo están las conclusiones y recomendaciones del trabajo desarrollado.

ÍNDICE GERERAL

Pág.

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

INTRODUCCIÓN

Debido a la crisis económica en la que se encuentra el país muchas personas se encuentran con la necesidad de buscar la forma de sustentar sus gastos, ya sea teniendo trabajos adicionales o haciendo sobretiempo en las empresas en las que laboran, siendo esto último, la causa de dolor de cabeza y preocupación de los dueños o administradores de las mismas, por tener que pagar gastos imprevistos.

Analizar las causas por las cuales, empleados de una empresa de servicios realizan sobretiempo es el objetivo principal de esta tesis, para ello se diseñará e implementará un modelo que ayude a los dueños o administradores de la misma a tomar las decisiones que considere pertinentes, enfocándose en el grupo que más sobretiempo realiza.

Para lograr esto, se aplicaran los diferentes pasos del proceso de KDD (*descubrimiento del conocimiento en bases de datos*), el cual va ha permitir identificar o descubrir patrones de sobretiempo dentro de los datos.

De las técnicas existentes en minería de datos se aplicará los árboles de decisión mediante el uso de WEKA. Este programa implementa numerosos algoritmos de aprendizaje y múltiples herramientas para transformar las

bases de datos y realizar un exhaustivo análisis. De los algoritmos que WEKA implementa, trabajaré con el Algoritmo J4.8, el cual es una extensión del Algoritmo C4.5.

CAPÍTULO I

I. INTRODUCCIÓN: HORAS NORMALES Y HORAS EXTRAS.

En el presente capítulo se realizará una revisión rápida de algunos conceptos necesarios, por considerarse de importancia para el desarrollo de esta tesis.

Dichos conceptos están, según el Código de Trabajo emitido en el año 2004 por el Congreso Nacional.

1.1 De la duración máxima de la jornada de trabajo, de los descansos obligatorios y de las vacaciones

Según el Código de Trabajo, el Capítulo V consta de tres Parágrafos de los cuales mencionares el primero que es de nuestro interés.

Parágrafo 1ro.

De las jornadas y descansos

Art. 47.- De la jornada máxima.- La jornada máxima de trabajo será de *ocho horas diarias*, de manera que no exceda de *cuarenta horas semanales*, salvo disposición de la ley en contrario.

El tiempo máximo de trabajo efectivo en el subsuelo será de seis horas diarias y solamente por concepto de horas suplementarias, extraordinarias o de recuperación, podrá prolongarse por una hora más, con la remuneración y los recargos correspondientes.

Art. 48.- Jornada especial.- Las comisiones sectoriales y las comisiones de trabajo determinarán las industrias en que no sea

permitido el trabajo durante la jornada completa, y fijarán el número de horas de labor.

Art. 49.- Jornada nocturna.- La jornada nocturna, entendiéndose por tal la que se realiza entre las 7 p.m. y las 6 a.m. del día siguiente, podrá tener la misma duración y dará derecho a igual remuneración que la diurna, aumentada en un veinticinco por ciento.

Art. 50.- Límite de jornada y descanso forzosos.- Las jornadas de trabajo obligatorio no pueden exceder de cinco en la semana, o sea de cuarenta horas hebdomadarias.

Los días sábados y domingos serán de descanso forzoso y, si en razón de las circunstancias, no pudiere interrumpirse el trabajo en tales días, se designará otro tiempo igual de la semana para el descanso, mediante acuerdo entre empleador y trabajadores.

Art. 51.- Duración del descanso.- El descanso de que trata el artículo anterior lo gozarán a la vez todos los trabajadores o por turnos si así lo exigiere la índole de las labores que realicen.

Comprenderá un mínimo de cuarenta y dos horas consecutivas.

Art. 52.- Trabajo en domingos y sábados por la tarde.- Las circunstancias por las que, accidental o permanentemente, se autorice el trabajo en los días domingos y sábados por la tarde, no podrán ser otras que éstas:

1a.- Necesidad de evitar un grave daño al establecimiento o explotación amenazado por la inminencia de un accidente; y, en general, por caso fortuito o fuerza mayor que demande atención impostergable. Cuando esto ocurra no es necesario que preceda autorización del Inspector del Trabajo, pero el empleador quedará obligado a comunicárselo dentro de las veinticuatro horas siguientes al peligro o accidente, bajo multa que será impuesta de conformidad con lo previsto en el artículo 626 de este Código, que impondrá el Inspector de Trabajo. En estos casos, el trabajo deberá limitarse al tiempo estrictamente necesario para atender al daño o peligro; y,

2a.- La condición manifiesta de que la industria, explotación o labor no pueda interrumpirse por la naturaleza de las

necesidades que satisfacen, por razones de carácter técnico o porque su interrupción irrogue perjuicios al interés público.

Art. 53.- Descanso semanal remunerado.- El descanso semanal forzoso será pagado con la cantidad equivalente a la remuneración íntegra, o sea de dos días, de acuerdo con la naturaleza de la labor o industria.

En caso de trabajadores a destajo, dicho pago se hará tomando como base el promedio de la remuneración devengada de lunes a viernes; y, en ningún caso, será inferior a la remuneración mínima.

Art. 54.- Pérdida de la remuneración.- El trabajador que faltare injustificadamente a media jornada continua de trabajo en el curso de la semana, tendrá derecho a la remuneración de seis días, y el trabajador que faltare injustificadamente a una jornada completa de trabajo en la semana, sólo tendrá derecho a la remuneración de cinco jornadas.

Tanto en el primer caso como en el segundo, el trabajador no perderá la remuneración si la falta estuvo autorizada por el

empleador o por la ley, o si se debiere a enfermedad, calamidad doméstica o fuerza mayor debidamente comprobada, y no excediere de los máximos permitidos.

La jornada completa de falta puede integrarse con medias jornadas en días distintos.

No podrá el empleador imponer indemnización al trabajador por concepto de faltas.

Art. 55.- Remuneración por horas suplementarias y extraordinarias.- Por convenio escrito entre las partes, la jornada de trabajo podrá exceder del límite fijado en los artículos 47 y 49, siempre que se proceda con autorización del Inspector del Trabajo y se observen las siguientes prescripciones:

1a.- Las horas suplementarias no podrán exceder de cuatro en un día, ni de doce en la semana;

2a.- Si tuviere lugar *durante el día o hasta las doce de la noche*, el empleador pagará la remuneración correspondiente a cada una de las horas suplementarias con más un *cincuenta*

por ciento de recargo. Si dichas horas estuvieren comprendidas entre las *doce de la noche y las seis de la mañana*, el trabajador tendrá derecho a un *ciento por ciento de recargo*. Para calcularlo se tomará como base la remuneración que corresponda a la hora de trabajo diurno;

3a.- En el trabajo a destajo se tomarán en cuenta para el recargo de la remuneración las unidades de obra ejecutadas durante las horas excedentes de las ocho obligatorias; en tal caso, se aumentará la remuneración correspondiente a cada unidad en un *cincuenta por ciento o en un ciento por ciento*, respectivamente, de acuerdo con la regla anterior. Para calcular este recargo, se tomará como base el valor de la unidad de la obra realizada durante el trabajo diurno; y,

4a.- El trabajo que se ejecutare el sábado o el domingo deberá ser pagado con el 100% de recargo.

Art. 56.- Prohibición.- Ni aun por contrato podrá estipularse mayor duración de trabajo diario que la establecida en el artículo que antecede.

Cuando ocurriere alguno de los casos previstos en el numeral primero del artículo 52, se podrá aumentar la jornada, debiendo el empleador dar parte del hecho al Inspector del Trabajo, dentro del mismo plazo, bajo igual sanción y con las mismas restricciones que se indican en el citado artículo.

Art. 57.- División de la jornada.- La jornada ordinaria de trabajo podrá ser dividida en dos partes, con reposo de hasta de dos horas después de las cuatro primeras horas de labor, pudiendo ser única, si a juicio del Director o Subdirector del Trabajo, así lo impusieren las circunstancias.

En caso de trabajo suplementario, las partes de cada jornada no excederán de cinco horas.

Art. 58.- Funciones de confianza.- Para los efectos de la remuneración, no se considerará como trabajo suplementario el realizado en horas que excedan de la jornada ordinaria, cuando los empleados tuvieren funciones de confianza y dirección, esto es el trabajo de quienes, en cualquier forma, representen al empleador o hagan sus veces; el de los agentes viajeros, de seguros, de comercio como vendedores y compradores, siempre

que no estén sujetos a horario fijo; y el de los guardianes o porteros residentes, siempre que exista contrato escrito ante la autoridad competente que establezca los particulares requerimientos y naturaleza de las labores.

Art. 59.- Indemnización al empleador.- Si el trabajador, sin justa causa, dejare de laborar las ocho horas de la jornada ordinaria, perderá la parte proporcional de la remuneración.

En caso de labores urgentes paralizadas por culpa del trabajador, el empleador tendrá derecho a que le indemnice el perjuicio ocasionado. Corresponde al empleador probar la culpa del trabajador.

Art. 60.- Recuperación de horas de trabajo.- Cuando por causas accidentales o imprevistas, fuerza mayor u otro motivo ajeno a la voluntad de empleadores y trabajadores se interrumpiere el trabajo, el empleador abonará la remuneración, sin perjuicio de las reglas siguientes:

1a.- El empleador tendrá derecho a recuperar el tiempo perdido aumentando hasta por tres horas las jornadas de los días subsiguientes, sin estar obligado al pago del recargo;

2a.- Dicho aumento durará hasta que las horas de exceso sean equivalentes por el número y el monto de la remuneración, a las del período de interrupción;

3a.- Si el empleador tuviere a los trabajadores en el establecimiento o fábrica hasta que se renueven las labores, perderá el derecho a la recuperación del tiempo perdido, a menos que pague el recargo sobre la remuneración correspondiente a las horas suplementarias y conformándose en todo a lo prescrito en el artículo 55, reglas 2a. y 3a.;

4a.- El trabajador que no quisiere sujetarse al trabajo suplementario devolverá al empleador lo que hubiere recibido por la remuneración correspondiente al tiempo de la interrupción; y,

5a.- La recuperación del tiempo perdido sólo podrá exigirse a los trabajadores previa autorización del Inspector del Trabajo, ante el cual el empleador elevará una solicitud detallando la fecha y causa de la interrupción, el número de horas que

duró, las remuneraciones pagadas, las modificaciones que hubieren de hacerse en el horario, así como el número y determinación de las personas a quienes se deba aplicar el recargo de tiempo.

Art. 61.- Cómputo de trabajo efectivo.- Para el efecto del cómputo de las ocho horas se considerará como tiempo de trabajo efectivo aquel en que el trabajador se halle a disposición de sus superiores o del empleador, cumpliendo órdenes suyas.

Art. 62.- Trabajo en días y horas de descanso obligatorio.- En los días y horas de descanso obligatorio, el empleador no podrá exigir al trabajador labor alguna, ni aun por concepto de trabajo a destajo, exceptuándose los casos contemplados en el artículo 52.

Art. 63.- Exhibición de horarios de labor.- En todo establecimiento de trabajo se exhibirá en lugar visible el horario de labor para los trabajadores, así como el de los servicios de turno por grupos cuando la clase de labor requiera esta forma.

Las alteraciones de horario a que dieren margen la interrupción y recuperación del trabajo serán publicadas en la misma forma.

El trabajador tendrá derecho a conocer desde la víspera las horas fijas en que comenzará y terminará su turno, cuando se trate de servicios por reemplazos en una labor continua, quedándole también el derecho de exigir remuneración por las horas de espera, en caso de omitirse dichos avisos.

Art. 64.- Reglamento interno.- Las fábricas y todos los establecimientos de trabajo colectivo elevarán a la Dirección General del Trabajo o a las subdirecciones del trabajo en sus respectivas jurisdicciones, copia legalizada del horario y del reglamento interno para su aprobación.

Sin tal aprobación, los reglamentos no surtirán efecto en todo lo que perjudiquen a los trabajadores, especialmente en lo que se refiere a sanciones.

El Director General del Trabajo y los subdirectores del trabajo reformarán, de oficio, en cualquier momento, dentro de su jurisdicción, los reglamentos del trabajo que estuvieren aprobados, con el objeto de que éstos contengan todas las

disposiciones necesarias para la regulación justa de los intereses de empleadores y trabajadores y el pleno cumplimiento de las prescripciones legales pertinentes.

Copia auténtica del reglamento interno, suscrita por el Director o Subdirector del Trabajo, deberá enviarse a la organización de trabajadores de la empresa y fijarse permanentemente en lugares visibles del trabajo, para que pueda ser conocido por los trabajadores. El reglamento podrá ser revisado y modificado por las aludidas autoridades, por causas motivadas en todo caso, siempre que lo soliciten más del cincuenta por ciento de los trabajadores de la misma empresa.

CAPÍTULO II

II. MÉTODOS PARA EXTRAER PATRONES DE DATOS.

La tecnología actual nos permite capturar y almacenar una gran cantidad de datos, se ha estimado que la cantidad de datos almacenados en el mundo en bases de datos se duplica cada 20 meses.

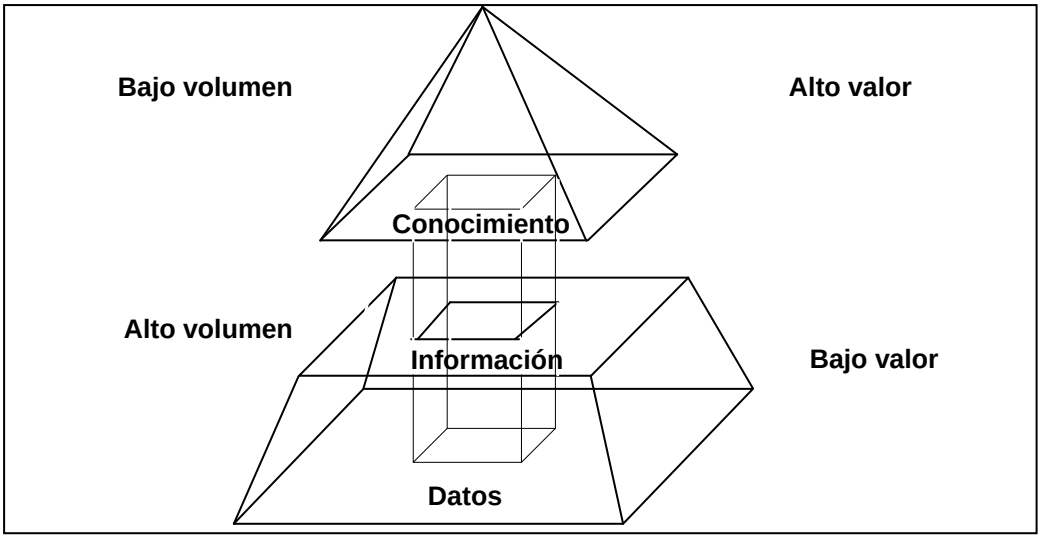
En la actualidad las organizaciones tienen gran cantidad de datos almacenados y organizados. Tratar de encontrar patrones, tendencias y anomalías es uno de los grandes retos de la vida moderna.

Se cree que se está perdiendo una gran cantidad de información y conocimiento valioso que se podría extraer de los datos.

Los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representen un valor agregado, entonces nos referimos al conocimiento.

En la figura siguiente se ilustra la jerarquía que existe en una base de datos entre datos, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento.

Figura 2.1
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Jerarquía que existe en una base de datos entre datos, información y conocimiento



2.1 Descubrimiento de Conocimiento en Bases de Datos (KDD)

El KDD es el “Proceso de extracción no trivial de identificar patrones que sean válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos”.

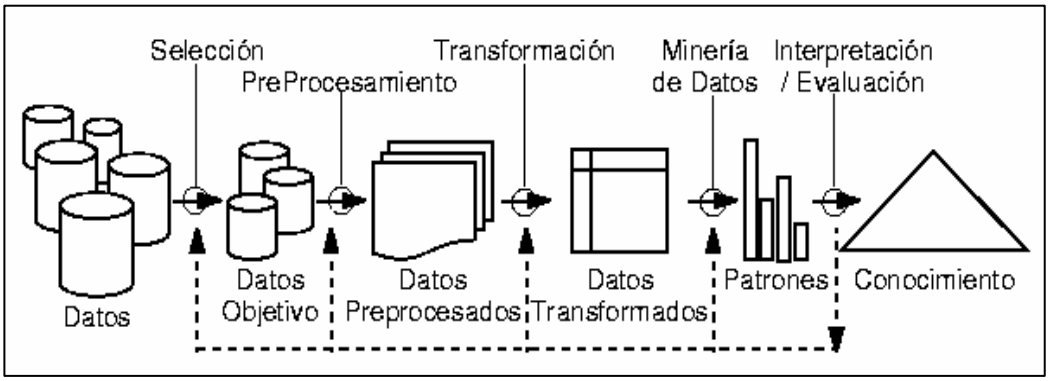
- *Proceso:* KDD involucra varios pasos y es interactivo, al encontrar información útil en los datos, se realizan mejores preguntas.
- *Válido:* se utilizan principalmente los datos y se espera que los patrones puedan aplicarse en el futuro.
- *Novedoso:* desconocido con anterioridad.
- *Útil:* aplicable y cumpliendo las metas del usuario.
- *Entendible:* que nos lleve a la comprensión, muchas veces medido por el tamaño.

2.1.1Metas del KDD

- ✓ Procesar automáticamente grandes cantidades de datos crudos.
- ✓ Identificar los patrones más significativos y relevantes.
- ✓ Presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

2.1.2 El Proceso KDD

Figura 2.2
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Proceso de Extracción de Conocimiento



El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos.

La interpretación de los patrones extraídos es lo que da origen al conocimiento.

Un modelo de conocimiento representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.

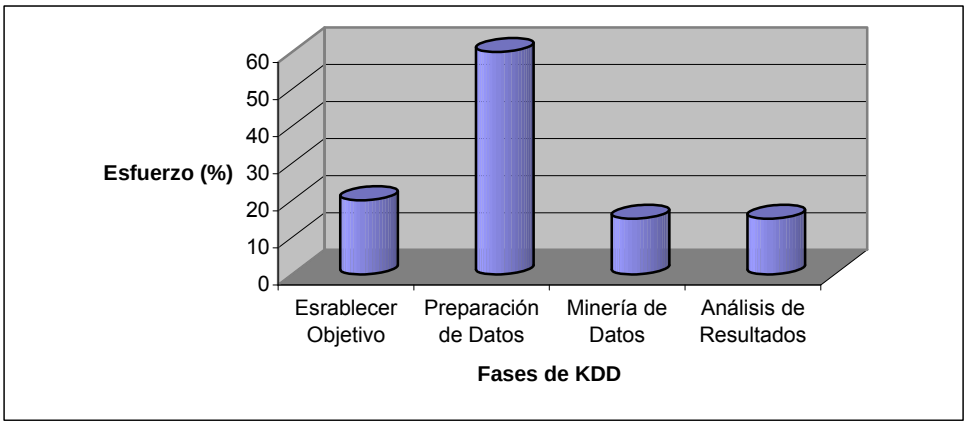
Pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

Al Descubrimiento de Conocimiento de Bases de Datos (KDD) a veces también se le conoce como minería de datos (*Data Mining*).

Sin embargo, muchos autores se refieren al Proceso de Minería de Datos como el de la aplicación de un algoritmo para extraer patrones de datos a partir de datos pre-procesados y a KDD al proceso completo (pre-procesamiento, minería, post-procesamiento).

Se estima que la extracción de patrones (minería) de los datos ocupa solo el 15% - 20% del esfuerzo total del proceso de KDD.

Figura 2.3
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Esfuerzo requerido en cada etapa del Proceso KDD



El proceso de descubrimiento de conocimiento en bases de datos involucra varias fases:

1. Determinar las fuentes de información: que pueden ser útiles y dónde conseguirlas.

2. Diseñar el esquema de un almacén de datos (Data Warehouse): que consiga unificar de manera operativa toda la información recogida.

3. Implantación del almacén de datos: que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados. Esta es la etapa que puede llegar a consumir el mayor tiempo.

4. Selección, limpieza y transformación de los datos que se van a analizar: La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos). La limpieza y pre-procesamiento de datos se logra diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.

5. Seleccionar y aplicar el método de minería de datos apropiado: Esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar. La transformación de los datos al formato requerido por el algoritmo específico de minería de datos. Y llevar a cabo el proceso de minería de datos, se buscan patrones que puedan expresarse como un modelo o simplemente que expresen dependencias de los datos, el modelo encontrado depende de su función (clasificación) y de su forma de representarlo (árboles de decisión, reglas, etc.), se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos, se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está predeterminada en el algoritmo de minería).

6. Evaluación, interpretación, transformación y representación de los patrones extraídos: Interpretar los resultados y posiblemente regresar a los pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias.

Este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes o irrelevantes.

7. Difusión y uso del nuevo conocimiento.

- Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente.
- El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas.

En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de minería de datos.

2.1.3 Representación de patrones

Según la manera de representar los patrones, podemos distinguir entre técnicas no simbólicas y técnicas simbólicas.

2.1.3.1 Técnicas no simbólicas

Las más numerosas y tradicionales son las técnicas no simbólicas, generalmente más apropiadas para variables continuas y con un conocimiento más claro de lo que se quiere buscar. El mayor inconveniente de las técnicas no simbólicas es su poca (o nula) inteligibilidad.

Entre las técnicas no simbólicas, puedo destacar las siguientes:

Redes Neuronales Artificiales, Lógica Difusa, Algoritmos Genéticos y combinaciones entre ellos.

2.1.3.2 Técnicas simbólicas

Las técnicas simbólicas generan un modelo “legible” y además aceptan mayor variedad de variables y mayor riqueza en la estructura de los datos.

Entre las técnicas simbólicas, puedo destacar:

Árboles de Decisión, Programación Inductiva y Otras Técnicas de Inducción Simbólica de Alto Nivel

2.1.4 Tipologías de patrones

Una vez recogidos los datos de interés en un almacén de datos, un explorador puede decidir qué tipos de patrón quiere descubrir.

Es importante destacar que la elección del tipo de conocimiento que se desea extraer va a marcar claramente la técnica de minería de datos a utilizar.

Los propios sistemas de minería de datos se encargan generalmente de elegir la técnica más idónea entre las disponibles para un determinado tipo de patrón a buscar, con lo que el explorador sólo debe determinar el tipo de patrón.

Tipos de conocimiento que se desea extraer:

1. **Asociaciones:** Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta.

Ejemplo: en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.

2. Dependencias: Una dependencia fundamental (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Uno de los mayores problemas de la búsqueda de dependencias es que suelen existir muchas dependencias nada interesantes y en las que la causalidad es justamente la inversa.

Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.

3. Clasificación: Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas. Muchas veces se conoce como aprendizaje supervisado.

Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, grado de miopías y astigmatismo han determinado los pacientes para los que su operación de

cirugía ocular ha sido satisfactoria. Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.

4. Agrupamiento/Segmentación: La segmentación (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

5. Tendencias/Regresión: El objetivo es predecir los valores de una variable continua a partir de la evolución de otra variable continua, generalmente el tiempo.

Ejemplo: se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.

6. Reglas Generales: Evidentemente muchos patrones no se ajustan a los tipos anteriores. Recientemente los sistemas

incorporan capacidad para establecer otros patrones más generales.

Una vez determinado el tipo de patrón a buscar, el sistema (u opcionalmente el usuario) puede elegir la técnica más apropiada.

2.1.5 Técnicas de KDD

Los algoritmos de aprendizaje son una parte integral de KDD. Las técnicas de aprendizaje podrán ser *supervisadas o no supervisadas*. En general, las técnicas de aprendizaje dirigidas disfrutan de un rango de éxito definido por la utilidad del descubrimiento del conocimiento.

Los algoritmos de aprendizaje son complejos y generalmente considerados como la parte más difícil de cualquier técnica KDD.

2.1.5.1 Algoritmos Supervisados o Predictivos.

Los algoritmos supervisados estiman una función f que mejor asocia un conjunto de datos X (variables independientes) con

un conjunto de datos Y (variables dependientes), dado un conjunto anterior de observaciones (datos a priori).

Ejemplos (x1, y1) , (x2, y2) ....(xN, yN).

Esta forma de trabajar se conoce como *aprendizaje supervisado* y se desarrolla en dos fases:

1. Fase de entrenamiento o supervisión
2. Fase Prueba

En cada fase se trabaja con un conjunto de datos diferentes: datos de entrenamiento o diseño y datos de pruebas, ambos conjuntos de datos se sacan del conjunto de datos iniciales.

En la *fase de supervisión*, al algoritmo se le presentan los datos de entrenamiento y éste ajusta sus parámetros internos de su modelo de tal manera que minimice el error de predicción de la variable dependiente Y.

Pasada la fase de entrenamiento se aplica la *fase de prueba* la cual consiste en la estimación del error cometido por el modelo; pero basado en los datos de prueba no en los datos usados en la etapa de supervisión. El error encontrado en la fase prueba

es una aproximación más cercana al error de predicción del modelo. Y lo que se quiere hacer es encontrar modelos que minimicen el error de predicción.

Entre los algoritmos supervisados tenemos:

- *Regresión Lineal:* Clasifica regiones con límites lineales.
- *Los K-ésimos vecinos más cercanos* (K-Nearest Neighbors).
- *Árboles de Decisión:* Clasifica regiones que pueden dividirse mediante rectángulos.
- *Redes Neuronales:* Clasifica regiones arbitrariamente complejas.
- *Clasificadores Bayesianos*: Define regiones de clasificación en base a la regla de Bayes.

2.1.5.2 Algoritmos No Supervisados o del Descubrimiento del Conocimiento.

Dado un conjunto de variables aleatorias x1, x2, .. xN para los cuales no existe ninguna variable Y que clasifique a estas variables. Entonces sólo se puede aplicar los algoritmos de tipo no supervisado los que encarga de estimar o de explorar ciertas propiedades de la distribución conjunta de x1, x2,... xN es decir, P(x1, x2,... xN)

Entre los algoritmos no supervisados tenemos:

- Reglas de Asociación
- Agrupamiento (Clustering)

El descubrimiento de la máquina es uno de los campos más recientes que han contribuido para KDD. Mientras el descubrimiento de la máquina confía solamente en métodos autónomos para el descubrimiento de la información, KDD típicamente combina métodos automatizados con la interacción humana para asegurar resultados exactos, útiles, y entendibles.

Existen muchos métodos diferentes que son clasificados como las técnicas de KDD:

- Métodos cuantitativos, como los probabilísticos y los estadísticos.
- Métodos que utilizan las técnicas de visualización.
- Métodos de clasificación como la clasificación de Bayesian, lógica inductiva, descubrimiento de modelado de datos y análisis de decisión.

- Otros métodos incluyen la desviación y tendencia al análisis, algoritmos genéticos, redes neuronales y los métodos híbridos que combinan dos o más técnicas.

Debido a las maneras en que estas técnicas pueden usarse y combinarse, hay una falta de acuerdos de cómo estas técnicas deben categorizarse.

Por ejemplo, el método Bayesiano puede agruparse lógicamente con los métodos probabilísticos, de clasificación o de visualización.

Por causa de la organización, cada método descrito aquí es incluido en el grupo que mejor encaje. Sin embargo, esta selección no implica una categorización estricta.

1. Método Probabilístico

Esta familia de técnicas KDD utiliza modelos de representación gráfica para comparar las diferentes representaciones del conocimiento. Estos modelos están basados en las probabilidades e independencias de los datos.

Estos son útiles para aplicaciones que involucran incertidumbre y aplicaciones estructuradas tal que una probabilidad puede asignarse a cada uno de los “resultados” o pequeña cantidad del descubrimiento del conocimiento.

Las técnicas probabilísticas pueden usarse en los sistemas de diagnóstico, planeación y sistemas de control.

2. Método Estadístico

Este método usa la regla del descubrimiento y se basa en las relaciones de los datos. El algoritmo de aprendizaje inductivo puede seleccionar automáticamente trayectorias útiles y atributos para construir las reglas de una base de datos con muchas relaciones.

Este tipo de inducción es usado para generalizar los modelos en los datos y construir las reglas de los modelos nombrados.

Parte de las múltiples técnicas estadísticas se pueden utilizar para confirmar asociaciones y dependencias, y para realizar segmentaciones. Una técnica muy importante es el uso de regresión lineal (y no lineal) y redes de regresión para

establecer tendencias. También son originariamente estadísticos los árboles de regresión.

El proceso analítico en línea (OLAP) es un ejemplo de un método orientado a la estadística.

3. Método de Clasificación

La clasificación es probablemente el método más viejo y más usado de todos los métodos de KDD. Este método agrupa los datos de acuerdo a similitudes o clases. Hay muchos tipos de clasificación de técnicas y numerosas herramientas disponible que son automatizadas.

A continuación se describirá brevemente las técnicas más importantes:

- **Reglas de Asociación:** Establece asociaciones en base a los perfiles de los clientes sobre los cuales se está realizando el data mining. Las reglas de Asociación están siempre definidas sobre atributos binarios. No es muy complicado generar reglas en grandes bases de datos.

El problema es que tal algoritmo eventualmente puede dar información que no es relevante. Data Mining envuelve modelos para determinar patterns a partir de los datos observados. Los modelos juegan un rol de conocimiento inferido. Diciendo cuando el conocimiento representa conocimiento útil o no, esto es parte del proceso de extracción de conocimiento en bases de datos (Knowledge Discovery in Databases-KDD).

- **Método del vecino más cercano:** Una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases de k registro/s más similar/es a él en un conjunto de datos históricos. Algunas veces se llama la técnica del vecino k-más cercano.

- **Método Bayesiano:** es un modelo gráfico que usa directamente los arcos exclusivamente para formar un [sic] gráfica acíclica. Usa los medios probabilísticos y gráficos de representación, pero también es considerado un tipo de clasificación.

Se usan muy frecuentemente las **redes Bayesianas** cuando la incertidumbre se asocia con un resultado puede expresarse en términos de una probabilidad.

Este método cuenta con un dominio del conocimiento codificado y ha sido usado para los sistemas de diagnóstico.

- **Programación Inductiva:** Fundamentalmente se usan para obtener patrones de tipo general, que se pueden establecer entre varios datos o son intrínsecamente estructurales. Aunque existen algunas aproximaciones basadas en reglas simples, es la *Programación Lógica Inductiva* (ILP) el área que ha experimentado un mayor avance en la década de los noventa.

ILP se basa en utilizar la lógica de primer orden para expresar los datos, el conocimiento previo y las hipótesis.
Como la mayoría de bases de datos actuales siguen el modelo relacional, ILP puede trabajar directamente con la estructura de la misma, ya que una base de datos relacional se puede ver como una teoría lógica.

Aparte de esta naturalidad que puede evitar o simplificar la fase de preprocesado, ILP permite representar hipótesis o patrones relacionales, aprovechando y descubriendo nuevas relaciones entre individuos.

- ***Regla de inducción:*** la extracción de reglas if-then de datos basados en significado estadístico.

  Estos patrones son imposibles de expresar con representaciones clásicas. Nótese que un árbol de decisión siempre se puede convertir fácilmente en un conjunto de reglas pero no viceversa.

- **Descubrimiento de patrones y de datos:** es otro tipo de clasificación que sistemáticamente reduce una base de datos grande a unos cuantos archivos informativos. Si el dato es redundante y poco interesante se elimina, la tarea de descubrir los patrones en los datos se simplifica. Este método trabaja en la premisa de un dicho viejo, “menos es más".

El descubrimiento de patrones y las técnicas de limpia de datos son útiles para reducir volúmenes enormes de datos en las aplicaciones, tal como aquellos encontrados al analizar las

grabaciones de un censor automatizado. Una vez que las lecturas del censor se reducen a un tamaño manejable usando la técnica de limpia de datos, pueden reconocerse con más facilidad los patrones de datos.

- **El método del árbol de decisión:** usa las reglas de predicción, construidas como figuras gráficas basado en datos premisos, y clasificación de los datos según sus atributos. Este método requiere clases de datos que son discretos y predefinidos. El uso primario de este método es para predecir modelos que pueden ser apropiados para cualquier clasificación o técnicas de regresión.

Los **árboles de decisión** son utilizados fundamentalmente para clasificación y segmentación, consisten en una serie de tests que van separando el problema, siguiendo la técnica del divide y vencerás, hasta llegar a las hojas del árbol que determinan la clase o grupo a la que pertenece el dato.

Existen muchísimas técnicas para inducir árboles de decisión, siendo el más famoso el algoritmo C4.5 de Quinlan. Los

árboles de regresión son similares a los árboles de decisión pero basados en técnicas estadísticas.

Los árboles de decisión incluyen:

- ***CART*** *Árboles de clasificación y regresión:* técnica usada para la clasificación de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando 2 divisiones. Requiere menos preparación de datos que CHAID.

- ***CHAID*** *Detección de interacción automática de Chi cuadrado:* técnica similar a la anterior, pero segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones.

- **Método de desviación y tendencia del análisis:** La base de este método es el *método de detección por filtrad*o. Normalmente las técnicas de análisis y desviación son aplicadas temporalmente en las bases de datos. Una buena

aplicación para este tipo de KDD es el análisis de tráfico en las grandes redes de telecomunicaciones.

- **Redes neuronales:** son modelos predecibles, no lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.

Estos son particularmente útiles para el reconocimiento de patrones y algunas veces se pueden agrupar con los métodos de clasificación.

- **Algoritmos genéticos:** son usados para la clasificación, son similares a las redes neuronales aunque estas son consideradas más poderosos.

Estos algoritmos son técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución.

Inspirados en el principio de la supervivencia de los más aptos. La recombinación de soluciones buenas en promedio produce mejores soluciones. Es una analogía con la evolución natural.

- **Programación Genética:** se basan en la evolución de programas de cómputos que permitan explicar o predecir con mínimo error un determinado fenómeno.

- **Método híbrido:** También es llamado método multi-paradigmático. Combina la potencia de más de un método, aunque la implementación puede ser más difícil. Algunos de los métodos comúnmente usados combinan técnicas de visualización, inducción, redes neuronales y los sistemas basados en reglas para llevar a cabo el descubrimiento de conocimiento deseado. También se han usado bases de datos deductivas y algoritmos genéticos en los métodos híbridos.

En definitiva, existen multitud de técnicas, combinaciones y nuevas variantes que aparecen recientemente, debido al interés del campo. Así, los sistemas de KDD se afanan por incorporar la mayor cantidad de técnicas, así como ciertas heurísticas para

determinar o asesorar al usuario sobre qué métodos son mejores para distintos problemas.

Para estimar el error o determinar el rendimiento del algoritmo se suele usar **validación cruzada**, para esto se realiza una validación cruzada 10-veces (del inglés 10-fold cross validation). Este procedimiento divide los datos disponibles en 10 subconjuntos de instancias y toma sucesivamente un conjunto diferentes como conjunto de prueba usando el resto de instancias (de los otros subconjuntos) como conjunto de prueba. De esta forma, el algoritmo se ejecuta 10 veces aprendiendo en cada ejecución con un 90% de las instancias y se prueba con el 10% de instantes restantes.

2.1.6 Retos del proceso y de su aplicación

Los criterios técnicos de aplicación de un proceso de KDD incluyen consideraciones sobre la disponibilidad de casos suficientes. En general, cuantos más atributos se tengan en consideración, más datos serán necesarios y más difícil será encontrar los patrones.

Sin embargo, el disponer de antemano de cierto conocimiento acerca de los datos podría reducir en gran parte el número de casos necesarios.

El disponer de cierto conocimiento acerca del dominio de los datos puede resultar muy beneficioso, pues puede ayudar a establecer cuales son los atributos o valores de los mismos más importantes, cuales son las posibles dependencias entre ellos así como patrones conocidos con anterioridad.

Otra consideración muy importante tiene que ver con los atributos a tener en cuenta y su relevancia. Es muy importante que los atributos sean relevantes para la tarea de descubrimiento. Por muchos datos que se tengan, si los atributos no son relevantes no se podrá obtener ningún conocimiento.

El ruido es otra variable a tener en cuenta, a menos que se tenga una gran cantidad de datos que mitiguen el ruido y que clarifique los patrones, el conocimiento obtenido no será significativo.

Se estudian a continuación los principales retos con los que se encuentran los diseñadores de sistemas con capacidades de extracción de conocimiento en bases de datos:

1) Bases de datos muy grandes: Es común hoy en día la situación de manejar bases de datos con cientos de tablas donde cada una tiene cientos de campos y millones de tuplas. Por ello los métodos de Data Mining tienen que ser concebidos para tratar con estas cantidades de datos lo que supone la búsqueda de algoritmos más eficientes, posibles muestreos, y procesamiento masivamente paralelo. Asimismo, la construcción de Data Warehouse específicos de Data Mining y en los que ciertas estructuras y valores acerca de los datos que contiene han sido almacenados puede ayudar en alguna medida al éxito de los sistemas de descubrimiento de conocimiento.

2) Alto número de variables o atributos: Como se ha indicado antes no es grande tan sólo el número de registros en las bases de datos a explorar sino también el número de atributos a tener en cuenta. Esto presenta problemas para la eficiente computación de los algoritmos, puesto que agranda el espacio

de búsqueda en explosión combinatoria. Por otra parte, aumenta el riesgo de que el algoritmo encuentre patrones que no son válidos en general.

Para solucionar este problema están los métodos de reducción del número de atributos y la utilización del conocimiento del dominio para desechar las variables que no son relevantes.

3) Datos cambiantes: Cuando se trata con bases de datos en continua evolución, se corre el riesgo de que los patrones descubiertos dejen de ser válidos una vez transcurrido un determinado tiempo. Por otra parte, puede ocurrir incluso que las variables significativas dejen de serlo con el tiempo, que se modifiquen o incluso se borren. Una posible solución son los algoritmos incrementales para actualización de patrones. De todos modos una de las premisas fundamentales cuando se aplican algoritmos de Data Mining pasa por suponer que los datos son estables y los patrones resultados serán aplicables mientras los datos no cambien.

Cuando se haga una actualización de los datos que contiene el Data Warehouse será cuando se tendrán que aplicar las

técnicas de algoritmos incrementales para evitar tener que recalcular todas las estructuras al realizar una consulta de Data Mining.

4) Bases de datos con gran cantidad de valores nulos y/o ruido: Este problema se da sobre todo en bases de datos de negocios. Si la base de datos no se creó y diseñó con las tareas de descubrimiento en mente puede ocurrir que campos muy relevantes contengan gran cantidad de valores nulos.

Para solucionarlo se puede optar por desechar los registros donde hay valores nulos si estos no son muy abundantes, o intentar identificar los valores nulos teniendo en cuenta las posibles dependencias existentes entre las variables. La aplicación de algoritmos de tratamiento de nulos e información desconocida se realiza en la fase de preprocesado de manera que cuando se tiene que aplicar un determinado algoritmo de Data Mining los datos ya no tienen este tipo de impurezas.

5) Relaciones complejas en los datos: Los primeros algoritmos fueron diseñados para valores de atributos simples, sin embargo, con el paso del tiempo se hace necesario el

desarrollo de nuevas técnicas capaces de tratar con estructuras como jerarquías de conceptos y otras más sofisticadas.

6) Comprensión de los patrones: En muchas aplicaciones ocurre que es crucial la manera de presentar los resultados a los usuarios. Esto ha dado lugar al desarrollo de representaciones gráficas, estructurado de las reglas, y técnicas para la visualización de los datos y el conocimiento. Es muy importante también cuando se van a presentar reglas a los usuarios el eliminar todas aquellas que sean implícita o explícitamente redundantes.

7) Interacción con el usuario: Los primeros métodos y herramientas no eran realmente interactivo y no dejaban incorporan conocimiento del dominio. La posibilidad de incorporar conocimiento es importantísima en todos los pasos del proceso.

8) Integración con otros sistemas: Un sistema autónomo cuya misión sea tan sólo el descubrimiento de conocimiento puede que no sea de interés. La integración mas importante (pero no la única), es integrarlo con un gestor de bases de

datos relacional que es el enfoque que se está defendiendo a lo largo de esta unidad. Se puede igualmente integrar y resulta muy útil para el soporte a la decisión el integrarlo con hojas de cálculo y herramientas de visualización.

CAPÍTULO III

III. MODELO PROPUESTO PARA EXTRAER INFORMACIÓN DEL COMPORTAMIENTO DEL SOBRETIEMPO.

En este capítulo se aplicaran los diferentes pasos del proceso de KDD (*descubrimiento del conocimiento en bases de datos*), explicados brevemente en el capítulo anterior, a través de los cuales se creará un modelo para el análisis de la base de datos perteneciente a una empresa de servicios cuyo nombre se omite por razones de confidencialidad.

Este proceso me va ha permitir identificar o descubrir patrones de sobretiempo y tendencias poco obvias dentro de los datos. El descubrimiento de esta información sirve para llevar a cabo acciones y

obtener un beneficio, sea este científico o de negocio, dado el caso de esta tesis el beneficio sería de negocio.

3.1 Determinar las fuentes de información

La fuente de información proviene de una base de datos que esta en Microsoft Visual FoxPro, perteneciente a una empresa de servicios, cuyos empleados han sido registrados al momento de formar parte del personal de la empresa en una tabla llamada *empleado.*

Lo que nos interesa de esta base de datos son los registros de las diferentes marcaciones que cada uno de los empleados ha realizado en sus jornadas laborales, almacenados en una tabla llamada *reloj* desde el 1 de marzo de 2000 hasta el 12 de mayo de 2004, haciendo un total de 1,207,522 registros.

El objetivo de esta información es determinar patrones por los cuales los empleados hacen sobretiempo. El realizar un análisis de los datos de sobretiempo de los empleados proveerá información importante

acerca de sus causas, esto le permitirá al dueño o administrador de la empresa de servicios tomar las decisiones que el considere pertinente.

3.2 Diseño del esquema de un almacén de datos

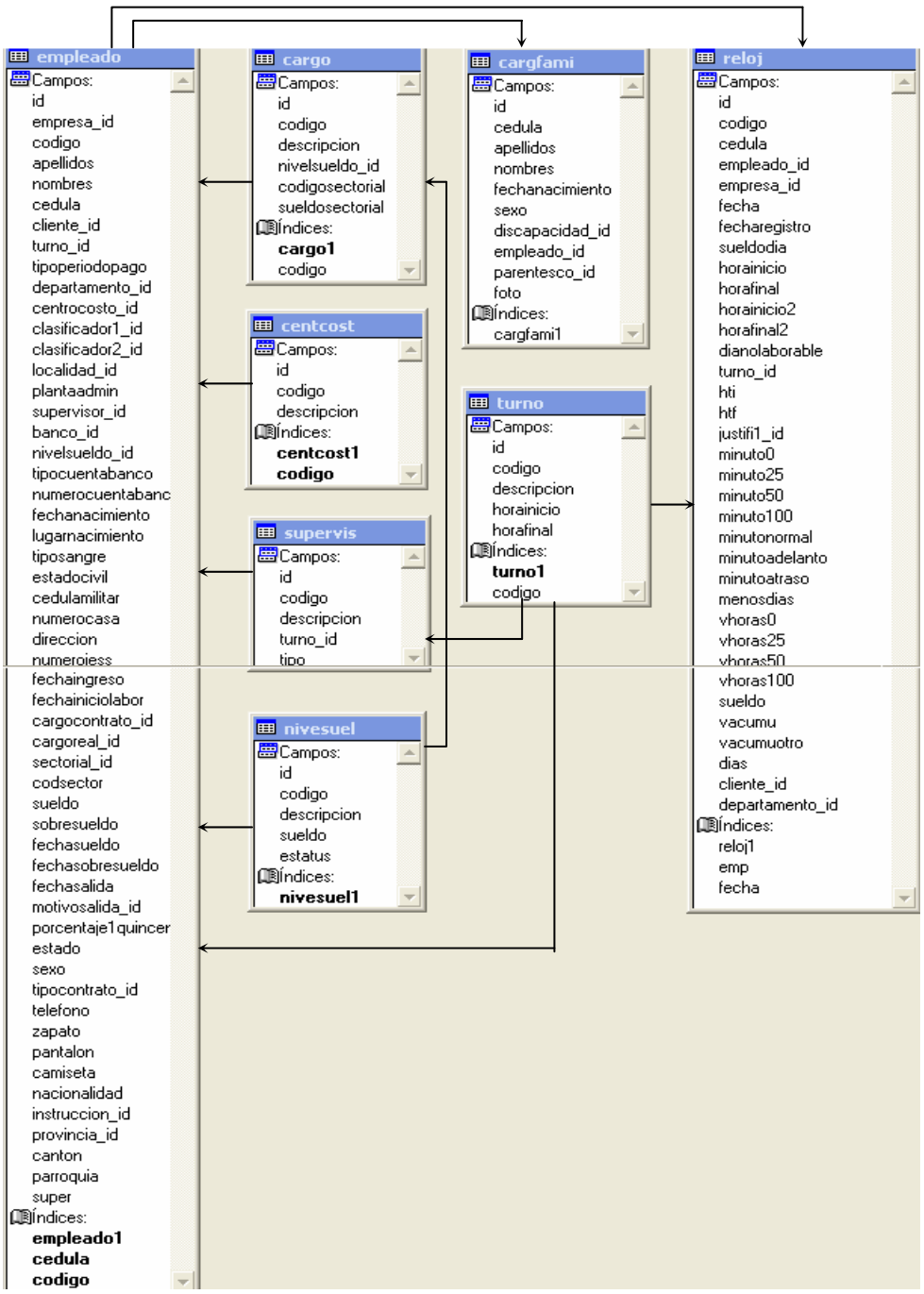
Un almacén de los datos (data warehouse) es típicamente una recogida de centros comerciales de los datos (data mart) que representen la totalidad de la información crítica de un negocio o de una empresa.

La base de datos con la que se trabajo (data mart) consta de las siguientes tablas:

- *empleado* (registra todos los datos relacionados con el empleado),
- *cargo* (registra la descripción de los cargos que tienen los empleados),
- *cargfami* (registra todos los datos de las cargas familiares de los empleados),
- *centcost* (registra la descripción de los centros de costo o departamentos en los que labora el empleado),
- *turno* (registra los diferentes turnos existentes en los que laboran los empleados),
- *supervis* (registra los supervisores de los distintos departamentos),

- *nivesuel* (registra los diferentes niveles de sueldo existentes),
- *reloj* (registra la marcación de entrada y salida laboral de cada empleado, siendo esta la mas importante).

Figura 3.1
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Modelo Entidad-Relación



3.3 Implantación del almacén de datos

El almacén de datos de la empresa de servicios de la cual estoy desarrollando la tesis, en la actualidad está implementado.

Esto nos va ha permitir proporcionar el acceso rápido de los usuarios del negocio a los datos refinados, apresurando el proceso de preguntas y de análisis en los datos de la empresa, para así, obtener una mejor información.

De la base de datos, la información que me ayudará para el objetivo de esta tesis, son todos los datos que relacionen al empleado directamente como por ejemplo: *la edad, el cargo que ocupa, el centro costo en el que labora, las cargas familiares que tiene, las horas que trabaja.*

3.4 Selección, limpieza y transformación de los datos que se van a analizar

Mediante la creación de querys, se ha realizado operaciones básicas sobre los datos, como el filtrado para reducir el ruido y derivación de nuevos atributos. A continuación se mostrará los querys utilizados con las indicaciones correspondientes de cada uno.

Figura 3.2
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Query usado para filtrar los datos

```
clear
set dele on
set safety off
set date to dmy

sele empleado_id , count(*) cargas from cargfami group by 1 into cursor cf

*---------- Query 1 ----------*

*Este query permite filtrar ciertos datos del empleado que se consideran importantes para el
*desarrollo de esta tesis. Debido a que la fecha de nacimiento de ciertos empleados fue
*ingresada incorrectamente, se ha filtrado aquellos cuya fecha de nacimiento es menor al
*01/01/1995 de esta manera los datos serán consistentes, estos datos son almacenados en
*una  tabla temporal llamada Templeado.

select emp.id, ser.codigo codser, cli.codigo codcli, cen.codigo codcen, sueldo, ;
cargo.codigo codcar, str(emp.nivelsueldo_id,1,0) nive, sexo, FechaNacimiento, (date() -
FechaNacimiento) / 365 edad, ;
iif( Empty(emp.estadocivil) , 'NN', emp.estadocivil) estadocivil, plantaadmin, ;
supervis.codigo codsuper, str(iif(isnull(cf.cargas), 0, cf.cargas),1,0) cargas  ;
from empleado emp left outer join cf ON cf.Empleado_id = emp.id , unidadmi ser, cliente cli, ;
cargo, centcost cen, supervis ;
where emp.empresa_id =ser.id and emp.cliente_id =cli.id and emp.cargoreal_id = cargo.id ;
and emp.centrocosto_id = cen.id and emp.supervisor_id = supervis.id AND ;
!Empty(FechaNacimiento) AND FechaNacimiento < ctod('01/01/1995') into cursor templeado
? 'Templeado'

*---------- Query 2 ----------*

*Este query permite filtrar las horas perdidas u horas que no ha laborado el empleado en la
*semana, a su vez se filtra los días no laborables o días festivos,
*estos datos son almacenados en una tabla temporal llamada TMenos.

SELECT reloj.empleado_id, peripago.codigo semana, ;
sum(Reloj.MenosDias*8+ Reloj.MinutoAtraso/60+Reloj.MinutoAdelanto/60 ) MenosHoras ;
FROM reloj, peripago  ;
WHERE Peripago.Tipo='S' AND Reloj.Fecha >= Peripago.FechaInicio AND ;
Reloj.Fecha <= Peripago.FechaFinal AND ;
DiaNoLaborable=0 ;
GROUP BY Reloj.empleado_id, peripago.codigo  ;
INTO cursor TMenos
Index on str(Empleado_id)+semana tag tmp
? 'TMenos'
```

Figura 3.2
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
*---------------------------------------------------- Query 3 ----------------------------------------------------*

*Con este query obtengo los diferentes turnos que ha tenido el empleado en la semana, los
*días que ha laborado por cada turno, se filtra los  días no laborables o días festivos en los
*que no se labora y las horas que ha laborado en la semana mayor a 0,  estos datos  son
*almacenados en una tabla temporal llamada TTurno.

Create Cursor TSemana( empleado_id integer, semana char(8), Entropia n(10,4))

SELECT reloj.empleado_id, peripago.codigo semana, Reloj.Turno_id, ;
sum(Reloj.MinutoNormal)/(60*8) DiasTurno ;
FROM reloj, peripago ;
WHERE Peripago.Tipo='S' AND Reloj.Fecha >= Peripago.FechaInicio AND ;
Reloj.Fecha <= Peripago.FechaFinal  ;
AND DiaNoLaborable=0 ;
GROUP BY Reloj.empleado_id, peripago.codigo, Reloj.Turno_id  ;
HAVING sum(Reloj.MinutoNormal) > 0 ;
INTO CURSOR TTurno
Index on str(Empleado_id)+semana tag tmp
? 'TTurno'

*si el empleado tiene diferentes turnos en la semana le asignamos 1, caso contrario 0. El
*atributo que almacena esta información en la tabla Tturno se llama Entropía.

Go Top
Do While !eof()
  nEmpleado = Empleado_id
  nSemana = Semana
  nDiasSemana = 0
  nRecno = Recno()
  Do While nEmpleado = Empleado_id AND nSemana = Semana AND !eof()
    nDiasSemana = nDiasSemana + DiasTurno
    skip
  Enddo
  Go nRecno
  nEntropia = 0
  Do While nEmpleado = Empleado_id AND nSemana = Semana AND !eof()
    nEntropia = nEntropia + DiasTurno/nDiasSemana * Log( DiasTurno/nDiasSemana )
    skip
  Enddo
  INSERT INTO TSemana VALUES ( nEmpleado, nSemana, -nEntropia )
Enddo

*---------------------------------------------------- Query 4 ----------------------------------------------------*

*con este query determinamos las horas extras y los días que labora el empleado en la
*semana, estos datos los almacenamos en una tabla temporal llamada TSobre.
```

Figura 3.2
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
SELECT reloj.empleado_id, peripago.codigo semana,  ;
sum(Reloj.Minuto100+Reloj.Minuto50)/60 SobreTiempo, Sum(MinutoNormal)/480 DiasSemana, ;
(Max(reloj.fecha) - emp.FechaNacimiento) / 365 edad ;
FROM reloj, peripago, Empleado Emp ;
WHERE Peripago.Tipo='S' AND Reloj.Fecha >= Peripago.FechaInicio AND ;
Reloj.Fecha <= Peripago.FechaFinal and emp.id=reloj.empleado_id ;
GROUP BY Reloj.empleado_id, peripago.codigo  ;
INTO cursor TSobre
Index on str(Empleado_id)+semana tag tmp
* aqui se quedo "Operator/Operand type mismatch"
? 'TSobre'

*---------- Query 5 ----------*

*Por medio de este query filtramos los datos que se considera más importantes del empleado,
*a su vez se filtra el estado civil diferente de NN, que no describe el estado civil real del
*empleado y se filtra además, los empleados que realizan sobretiempo, estos datos se
*almacenan en una tabla temporal llamada TReloj_modificado.

SELECT TSobre.empleado_id Emp_id, Tsobre.semana, emp.codcli, emp.codcen, emp.sueldo, ;
emp.codcar, emp.nive, emp.sexo, Tsobre.Edad, TMenos.MenosHoras, ;
emp.estadocivil , emp.plantaadmin, emp.codsuper, emp.cargas, TSobre.DiasSemana, ;
TSemana.Entropia, 00.0000 EntroAnt, TSobre.SobreTiempo sobre, 00.0000 SobreAnt, ' ' as SobreTmp ;
FROM TEmpleado Emp, Tsobre, Tsemana, Tmenos ;
WHERE emp.id = Tsobre.empleado_id  AND ;
emp.estadocivil <> 'NN' AND ;
TSemana.Empleado_id = TSobre.Empleado_id AND TSemana.Semana=Tsobre.Semana AND ;
TMenos.Empleado_id = TSobre.Empleado_id AND TMenos.Semana=Tsobre.Semana  AND;
TSobre.SobreTiempo > 0;
GROUP BY TSobre.empleado_id, TSobre.Semana  ;
INTO dbf TReloj_modificado
Index on str(Emp_id)+semana tag tmp
? 'TReloj_modificado'

*---------- Query 6 ----------*

*Esta sección  permite asignar las clases a los diferentes intervalos de sobretiempo.
*Para determinar las clases de sobretiempo, se hizo un query con la tabla TReloj_modificado,
*antes de añadir esta sección.

replace all sobretmp with '1' for sobre <=  3
replace all sobretmp with '2' for sobre > 3 AND sobre <= 12
replace all sobretmp with '3' for sobre > 12 AND sobre <= 23
replace all sobretmp with '4' for sobre > 23
Return
```

3.4.1 Obtención de clases de sobretiempo por medio del gráfico de sus frecuencias

Antes de continuar explicaré brevemente la forma en que se asignaron las clases de sobretiempo. Para esto, se hizo un query con el *atributo sobre* de la *tabla TReloj_modificado*, la cual registra los sobretiempos de los empleados, con este atributo se determinó en intervalos la frecuencia de sobretiempos, aplicando la siguiente ecuación:

$$y_i = k * \left(\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \right)$$

siendo k el número de intervalos, $x_{\min}$ el sobretiempo mínimo y $x_{\max}$ el sobretiempo máximo.

Sabiendo que $x_{\min}$ = 0.25, $x_{\max}$ = 80 y con k = 16 el query queda de la siguiente foma:

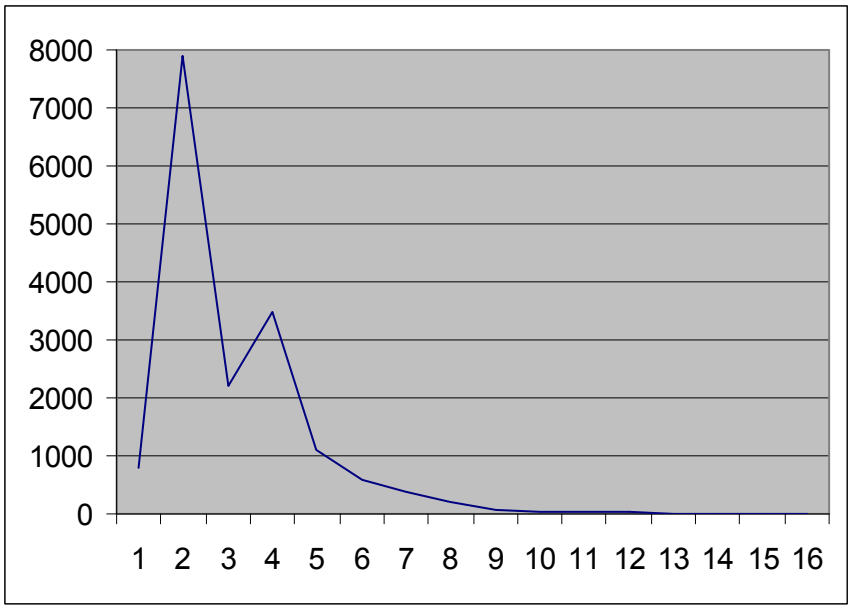
Figura 3.3
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Query usado para obtener las frecuencias de sobretiempo

```
select floor (16*(sobre-0.25)/(80-0.25)) as intervalo, count(*) frecuencia,;
sobre sobretiempo;
from Treloj_mod;
group by intervalo
```

Tabla 3.1
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Datos obtenidos del query de frecuencias de sobretiempo

Número	Intervalo	Frecuencia	Sobretiempo
1	[0 - 1)	800	3
2	[1 - 2)	7909	8
3	[2 - 3)	2194	12
4	[3 - 4)	3483	16
5	[4 - 5)	1095	23
6	[5 - 6)	594	28
7	[6 - 7)	367	32
8	[7 - 8)	204	36
9	[8 - 9)	72	44
10	[9 - 10)	41	48
11	[10 - 11)	22	52
12	[11 - 12)	25	60
13	[12 - 13)	3	64
14	[13 - 14)	11	68
15	[14 - 16)	3	71,75
16	[16]	1	80
		16824	

Figura 3.4
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Gráfico de frecuencias de sobretiempo



Por medio del grafico se puede determinar visualmente las clases de sobretiempo existentes, según las variaciones observadas en sus frecuencias. Las variaciones observadas están en 1, 3, 5, quedando las siguientes clases:

1. **Clase 1.** Indica aquellas personas que hacen un sobretiempo de 3 o menos horas a la semana.
2. **Clase 2.** Indica aquellas personas que hacen un sobretiempo de (3,12] horas a la semana.
3. **Clase 3.** Indica aquellas personas que hacen un sobretiempo de (12,23] horas a la semana.
4. **Clase 4.** Indica aquellas personas que hacen un sobretiempo de mas de 23 horas a la semana.

Estas clases son las que se registraron en el atributo *sobretmp* de la tabla temporal *TReloj_modificado,* quedando finalmente esta tabla con 16,824 registros y 20 atributos.

Se transforma los datos de la tabla *TReloj_modificado* con formato DBF a ARFF, el cual se lo utiliza en un programa llamado WEKA. Este programa lo explicare detalladamente en la siguiente sección de este capítulo, para transformar los datos se utiliza el siguiente query:

Figura 3.5
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Query que transforma datos en formato DBF a ARFF

```
PARAMETER sDir

If Empty( sDir )
  sDir = ''
Endif

set dele on
sTabla = alias()
If Empty(sTabla)
  MessageBox( "Tabla No Abierta" )
  Return
Endif

nFile = fCreate( sDir + sTabla + '.arff')

fPuts(nFile , "@relation '" + sTabla + "'" )

Dime aCol[1]
nCols = aField( aCol)

For i=1 To nCols
   sTipo = aCol[i,2]
   sColumna = lower(aCol[i,1])
   If sColumna = 'id' OR sColumna = 'codigo'
     Loop
   Endif
   If sTipo = 'C' OR sTipo = 'M'
     Select DISTINCT strtran(alltrim(&sColumna), ' ', '_') FROM (sTabla) INTO ARRAY
aVals
    Else
     Select DISTINCT &sColumna FROM (sTabla) INTO ARRAY aVals
   Endif

    nVals = _Tally
    if sTipo = 'N' OR sTipo = 'I'
     fPuts(nFile , "@attribute " + sColumna + " real" )
    Else
     sVals  = ' {'
     For j=1 To nVals
        If j > 1
          sVals = sVals + ','
        Endif
        sVals = sVals + alltrim(aVals[j])
     Endfor
     If j = nVals + 1
       sVals = sVals + '}'
     Endif
```

Figura 3.5
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
      fPuts(nFile , "@attribute " + sColumna + sVals )

    Endif
Endfor

fPuts(nFile , "@data" )

Sele (sTabla)
Go Top
Do While !eof()
  sLinea = ''
  For i=1 To nCols
     If !empty(sLinea)
       sLinea = sLinea + ','
     Endif
     sTipo = aCol[i,2]
     sColumna = aCol[i,1]
     nValor = Eval( sColumna)
     If sTipo = 'C' OR sTipo = 'M'
       sLinea = sLinea + alltrim( strtran(alltrim(nValor), ' ', '_') )
      Else
       sLinea = sLinea + alltrim( str(nValor))
     Endif
  Endfor
  fPuts(nFile , sLinea )
  skip
Enddo

fClose( nFile )
```

A continuación le mostraré la primera sección del archivo con formato ARFF que se obtuvo de la ejecución del query de la Figura 3.5.

Figura 3.6
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Datos en formato ARFF

```
@relation 'TRELOJ_MODIFICADO'
@attribute emp_id real
@attribute semana
{200001,200002,200003,200004,200005,200006,200007,200008,200009,200010,200011,200012,200013,200014,200015,200016,
200017,200018,200019,200020,200021,200022,200023,200024,200025,200026,200027,200028,200029,200030,200031,200032,
200033,200034,200035,200036,200037,200038,200039,200040,200041,200042,200043,200044,200045,200046,200047,200048,
200049,200050,200051,200052,200101,200102,200103,200104,200105,200106,200107,200108,200109,200110,200111,200112,
200113,200114,200115,200116,200117,200118,200119,200120,200121,200122,200123,200124,200125,200126,200127,200128,
200129,200130,200131,200132,200133,200134,200135,200136,200137,200138,200139,200140,200141,200142,200143,200144,
200145,200146,200147,200148,200149,200150,200151,200152,200201,200202,200203,200204,200205,200206,200207,200208,
200209,200210,200211,200212,200213,200214,200215,200216,200217,200218,200219,200220,200221,200222,200223,200224,
200225,200226,200227,200228,200229,200230,200231,200232,200233,200234,200235,200236,200237,200238,200239,200240,
200241,200242,200243,200244,200245,200246,200247,200248,200249,200250,200251,200252,200301,200302,200303,200304,
200305,200306,200307,200308,200309,200310,200311,200312,200313,200314,200315,200316,200317,200318,200319,200320,
200321,200322,200323,200324,200325,200326,200327,200328,200329,200330,200331,200332,200333,200334,200335,200336,
200337,200338,200339,200340,200341,200342,200343,200344,200345,200346,200347,200348,200349,200350,200351,200352,
200401,200402,200403,200404,200405,200406,200407,200408,200409,200410,200411,200412,200413,200414,200415,200416,
200417,200418,200419,200420,200421,200422,200423,200424,200425,200426,200427,200428,200429,200430,200431,200432,
200433,200434,200435,200436,200437,200438,200439,200440,200441,200442,200443,200444,200445,200446,200447,200448,
200449}
@attribute codcli {001,002,003}
@attribute codcen
{012,013,014,015,016,017,019,020,021,022,024,025,027,030,031,032,033,034,035,040,050,060,070,075,080,113,116,117,122,12
3,124,125,130,132,150}
@attribute sueldo real
@attribute codcar
{01,02,03,04,05,06,07,08,10,100,101,102,104,105,107,109,11,115,12,120,121,122,129,13,14,15,16,17,18,19,24,28,31,33,34,38,41
,42,47,48,51,53,54,55,57,58,59,64,69,70,71,72,73,74,75,76,79,80,81,83,84,85,86,88,90,91,97,98,99}
@attribute nive {0,1,2,3,4,5}
@attribute sexo {F,M}
@attribute edad real
@attribute menoshoras real
@attribute estadocivi {C,S,U}
@attribute plantaadmi {A,P}
@attribute codsuper {0001,0002,0003,0004,0007,0008,0009,0010,0012,01,02,03,0486,0541,0864,10,1665,1750,20,30,50,60}
@attribute cargas {0,1,2,3,4,5}
@attribute diassemana real
@attribute entropia real
@attribute entroant real
@attribute sobre real
@attribute sobreant real
@attribute sobretmp {1,2,3,4}
@data
6,200318,003,125,160,41,5,M,27,0,S,P,10,0,7,0,0,8,0,2
6,200319,003,125,160,41,5,M,27,0,S,P,10,0,7,1,0,16,0,3
6,200323,003,125,160,41,5,M,27,8,S,P,10,0,5,0,0,8,0,2
6,200341,003,125,160,41,5,M,28,0,S,P,10,0,7,0,0,8,0,2
6,200345,003,125,160,41,5,M,28,0,S,P,10,0,6,0,0,8,0,2
6,200352,003,125,160,41,5,M,28,0,S,P,10,0,7,0,0,8,0,2
6,200401,003,125,160,41,5,M,28,0,S,P,10,0,7,0,0,8,0,2
6,200408,003,125,160,41,5,M,28,3,S,P,10,0,7,0,0,5,0,2
6,200422,003,125,160,41,5,M,28,0,S,P,10,0,7,0,0,8,0,2
6,200427,003,125,160,41,5,M,28,0,S,P,10,0,7,0,0,3,0,1
6,200430,003,125,160,41,5,M,28,0,S,P,10,0,7,0,0,8,0,2
6,200431,003,125,160,41,5,M,29,0,S,P,10,0,7,0,0,3,0,1
6,200441,003,125,160,41,5,M,29,0,S,P,10,0,7,0,0,2,0,1
6,200442,003,125,160,41,5,M,29,0,S,P,10,0,7,0,0,1,0,1
22,200304,003,113,160,55,5,M,49,0,S,P,10,0,7,0,0,32,0,4
22,200305,003,113,160,55,5,M,49,0,S,P,10,0,7,1,0,48,0,4
22,200306,003,113,160,55,5,M,49,0,S,P,10,0,7,0,0,24,0,4
22,200307,003,113,160,55,5,M,49,3,S,P,10,0,7,0,0,3,0,1
22,200308,003,113,160,55,5,M,49,0,S,P,10,0,7,1,0,24,0,4
```

3.5 Seleccionar y aplicar el método de minería de datos apropiado

De las técnicas explicadas brevemente en el capítulo anterior, la técnica que mejor se adapta al objetivo de esta tesis son los *árboles de decisión.* Esta técnica es de fácil entendimiento y es apropiada para predicciones numéricas.

Les recuerdo que los árboles de decisión son utilizados fundamentalmente para clasificación y segmentación, siguiendo la técnica del divide y vencerás, hasta llegar a las hojas del árbol que determinan la clase o grupo a la que pertenece el dato.

Se aplicará esta técnica mediante el uso de WEKA, el cual me ayudará a extraer información del comportamiento de las horas extras. A continuación explicaré brevemente en que consiste este programa.

3.5.1 Introducción al programa WEKA

WEKA (Waikato Environment for Knowledge Analysis) fue desarrollado en la universidad de Waikato en Nueva Zelanda. Se trata de un programa o entorno para el análisis de conocimientos.

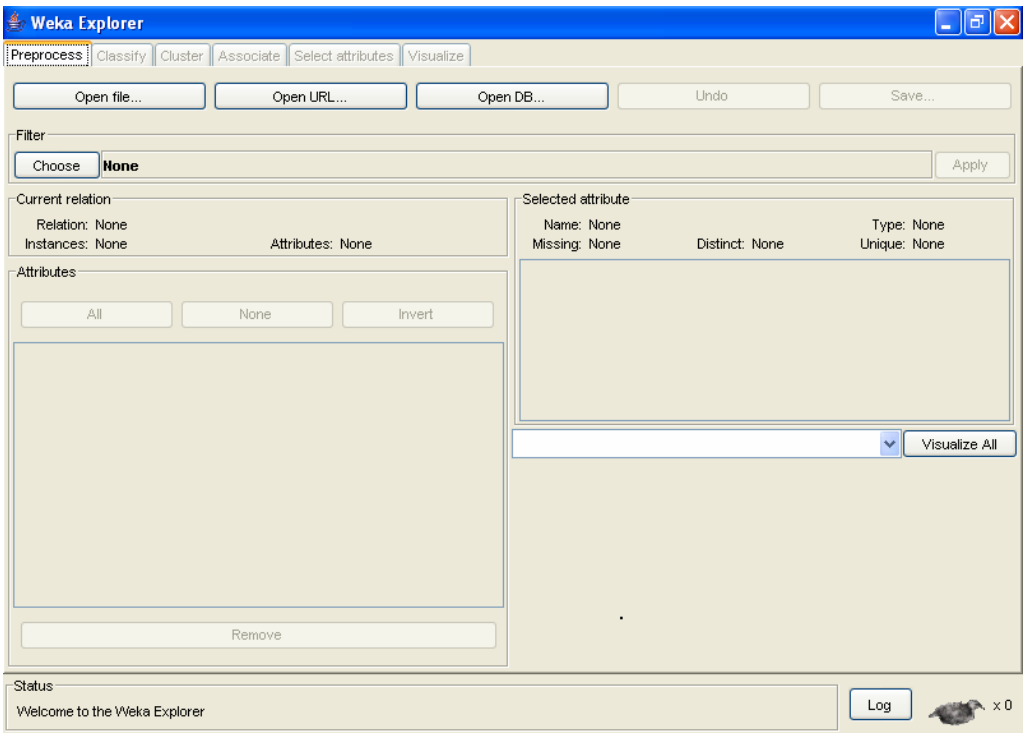
Este programa esta escrito en Java por lo que se convierte en un sistema multiplataforma. Implementa numerosos algoritmos de aprendizaje y múltiples herramientas para transformar las bases de datos y realizar un exhaustivo análisis.

Figura 3.7
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Primera pantalla de WEKA

Los algoritmos pueden ser aplicados directamente a un conjunto de datos o llamado de su propio código Java. Weka contiene aplicaciones para el pre-proceso de datos, la clasificación, regresión, clustering, reglas de asociación, y visualización. También es útil para desarrollar nuevos esquemas de aprendizaje. Weka es un software libre, el cual

tiene acceso en la página Web www.mkp.com/datamining o en www.cs.waikato.ac.nz/ml/weka.

Figura 3.8
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Pantalla principal del Explorador de WEKA



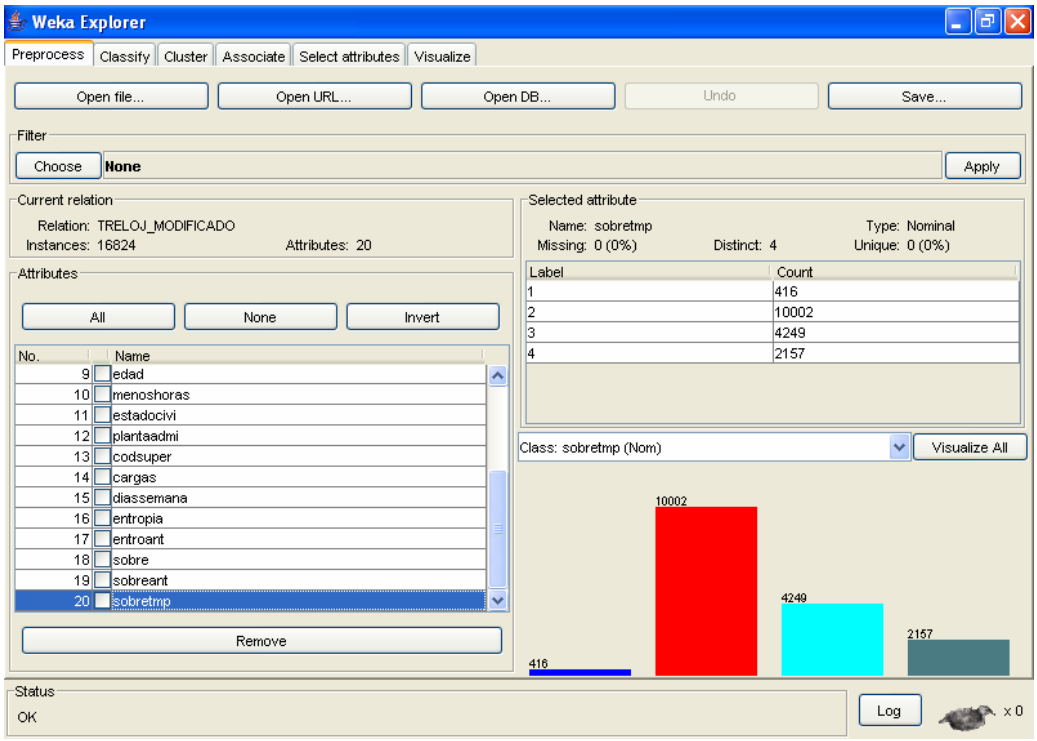
Esta es la página principal del Explorador de WEKA, donde podemos hacer un pre-procesamiento de la información. Desde aquí podemos seleccionar el archivo de datos con formato ARFF, que incluirá los ejemplos de entrenamiento y los atributos de que está compuesto el ejemplo. También podemos ver, una vez abierto un experimento, como

están distribuidos los distintos atributos de el conjunto de entrenamiento, así como editarlo, eliminando atributos.

Las restantes secciones serán descritas según sean necesarias para el desarrollo de la tesis.

3.5.2 Aplicación del programa WEKA

Figura 3.9
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Archivo TRELOJ_MODIFICADO.ARFF abierto



Podemos ver que hemos cargado el archivo *TRELOJ_MODIFICADO.ARFF,* en el que disponemos de 16,824 instancias o registros de 20 atributos.

Al tener seleccionado el atributo sobretmp, que es el que nos informa de las clases de sobretiempo, podemos ver que las clases están bastante diferenciadas y distribuidas de la siguiente manera:

1. La *primera clase* tiene 416 registros,
2. la *segunda clase* tiene 10,002 registros,
3. la *tercera clase* tiene 4,249 registros, y
4. la *cuarta clase* tiene 2,157 registros.

Antes de continuar, selecciono por lógica los atributos que considero menos importantes y los remuevo, quedando 14 de los 20 atributos, siendo estos los más relevantes:

1. codcen (código del centro costo o departamento)
2. sueldo
3. codcar (código cargo)
4. nive (nivel de sueldo)
5. sexo
6. edad
7. menoshoras (horas que no labora o perdidas en la semana)

8. estadocivi (estado civil del empleado)
9. plantaadmi (planta en la que labora)
10. codsuper (código del supervisor)
11. cargas (número de cargas familiares del empleado)
12. diassemana (días que labora en la semana)
13. entropia (indica si el empleado tiene turnos variables o no)
14. sobretmp (indica las clases de sobretiempo que han tenido los empleados).

Utilizando Tablas de Contingencia verificaré si existe independencia del atributo sobretmp (atributo a estudiar) con los demás atributos, de esta forma podré remover atributos que no tengan dependencia con este.

3.5.2.1 Tabla de Contingencia

Es una tabla que cruza dos variables cualitativas -una en filas y otra en columnas- de modo que cada celda recoge el número de asociaciones o frecuencias que existen entre las diferentes categorías de las variables.

Denominamos *variables cualitativas* a aquellas cuyo resultado es un valor o categoría de entre un conjunto finito de respuestas posibles.

Figura 3.10
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Ejemplo de Tabla de Contingencia

Factores	Nivel 1º factor B	Nivel 2º factor B	$n_{i.}$
Nivel 1º factor A	n_{11}	n_{12}	$\sum_j n_{ij}$ para i=1
Nivel 2º factor A	n_{21}	n_{22}	$\sum_j n_{ij}$ para i=2
$n_{.j}$	$\sum_i n_{ij}$ para j=1	$\sum_i n_{ij}$ para j=2	$n = \sum_i \sum_j n_{ij}$

Los n_{ij} representan el número de individuos observados en cada combinación de los niveles de los factores A, B y se consideran como la realización de una variable aleatoria con valores enteros y positivos.

Contraste X^2 (chi cuadrado) de independencia

Contrastación de la hipótesis de independencia en una tabla de contingencia bidimensional.

Las hipótesis a contrastar son: $H_0 : P_{ij} = P_{i.}P_{.j}$
$H_1 : P_{ij} \neq P_{i.}P_{.j}$

donde H_0 es la hipótesis nula y H_1 la hipótesis alternativa.

El estadístico propuesto para realizar este contraste es el siguiente:

$$X^2 = \sum_i \sum_j \frac{(N_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

teniendo en cuenta que bajo H_0 tenemos

$$\hat{m}_{ij} = \frac{n_{i.} n_{.j}}{n}.$$

Dicho estadístico se distribuye según una X^2 con (I-1)(J-1) grados de libertad. Además, si el valor observado supera al esperado, rechazaremos H_0.

Otra forma de contrastar la hipótesis es:

H_0: El factor A es independiente del factor B

H_1: ┐ *H_0*

Donde se compara el p-valor con el nivel de significación α prefijado.

Si $p > \alpha$ aceptamos H_0.

Si $p \leq \alpha$ rechazamos H_0.

Esta es la forma más sencilla de determinar si dos atributos o variables son independientes o no.

3.5.2.2 Análisis del atributo SOBRETMP con los demás atributos

Para este análisis utilice SPSS, donde el α prefijado es 0.05 de X^2. El inconveniente que se presenta con el test de independencia X^2 es que si la tabla de contingencia calculada por SPSS contiene más del 10% de las celdas con valores nulos o menores de 5, el valor p de la chi-

cuadrado no es confiable, razón por la cual se obvió el contraste de hipótesis de ciertos atributos como: codcar, entropia, menoshoras.

Atributo SOBRETMP con NIVE

Tabla de Contingencia

Tabla 3.2
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Tabla de Contingencia de SOBRETMP con NIVE

		NIVE						
		0	1	2	3	4	5	Total
SOBRETMP	1	154	4	6	16	2	234	416
	2	2301	490	418	302	322	6169	10002
	3	1424	210	247	149	79	2140	4249
	4	705	126	128	49	33	1116	2157
Total		4584	830	799	516	436	9659	16824

Contraste de Hipótesis

H_0: El factor SOBRETMP es independiente del factor NIVE

H_1: ┐H_0

Tabla 3.3
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Contraste de Hipótesis de SOBRETMP con NIVE

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	344.339[a]	15	.000
Likelihood Ratio	356.870	15	.000
Linear-by-Linear Association	162.567	1	.000
N of Valid Cases	16824		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.78.

El nivel de significancia obtenido es 0, el cual me permite rechazar H_0 a favor de H_1, *en otras palabras se puede decir que, el sobretiempo depende del nivel de sueldo de los empleados.*

Atributo SOBRETMP con SEXO

Tabla de Contingencia

Tabla 3.4
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Tabla de Contingencia de SOBRETMP con SEXO

		SEXO		
		F	M	Total
SOBRETMP	1	41	375	416
	2	570	9432	10002
	3	284	3965	4249
	4	41	2116	2157
Total		936	15888	16824

Contraste de Hipótesis

H_0: El factor SOBRETMP es independiente del factor SEXO

H_1: ┐H_0

Tabla 3.5
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Contraste de Hipótesis de SOBRETMP con SEXO

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	80.165[a]	3	.000
Likelihood Ratio	94.888	3	.000
N of Valid Cases	16824		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 23.14.

El nivel de significancia obtenido es 0, el cual me permite rechazar H_0 a favor de H_1, *en otras palabras se puede decir que, el sobretiempo depende del sexo de los empleados.*

Atributo SOBRETMP con ESTACIVI

Tabla de Contingencia

Tabla 3.6
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Tabla de Contingencia de SOBRETMP con ESTACIVI

		ESTACIVI			
		C	S	U	Total
SOBRETMP	1	62	347	7	416
	2	1448	8424	130	10002
	3	855	3338	56	4249
	4	500	1641	16	2157
Total		2865	13750	209	16824

Contraste de Hipótesis

H_0: El factor SOBRETMP es independiente del factor ESTACIVI

H_1: ┐H_0

Tabla 3.7
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Contraste de Hipótesis de SOBRETMP con ESTACIVI

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	138.025[a]	6	.000
Likelihood Ratio	134.766	6	.000
N of Valid Cases	16824		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.17.

El nivel de significancia obtenido es 0, el cual me permite rechazar H_0 a favor de H_1, *en otras palabras se puede decir que, el sobretiempo depende del estado civil de los empleados.*

Atributo SOBRETMP con CODSUPER

Tabla de Contingencia

Tabla 3.8
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Tabla de Contingencia de SOBRETMP con CODSUPER

		CODSUPER																	
		1	2	3	4	7	8	9	10	12	20	30	50	60	486	541	864	1665	1750	Total
SOBRETMP	1	82	25	33	7	4	9	22	106		75	23				1	1	3	25	416
	2	1809	2064	2854	61	207	119	652	828	72	237	44	1	4	47	137	134	433	299	10002
	3	667	827	897	36	120	138	229	282	9	102	13	6	9	59	100	51	138	566	4249
	4	266	521	436	11	67	245	57	142		48	7	14	32	47	50	21	36	157	2157
Total		2824	3437	4220	115	398	511	960	1358	81	462	87	21	45	153	288	207	610	1047	16824

Contraste de Hipótesis

H_0: El factor SOBRETMP es independiente del factor CODSUPER

H_1: ┐H_0

Tabla 3.9
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Contraste de Hipótesis de SOBRETMP con CODSUPER

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2566.829[a]	51	.000
Likelihood Ratio	1963.057	51	.000
Linear-by-Linear Association	109.951	1	.000
N of Valid Cases	16824		

a. 7 cells (9.7%) have expected count less than 5. The minimum expected count is .52.

El nivel de significancia obtenido es 0, el cual me permite rechazar H_0 a favor de H_1, *en otras palabras se puede decir que, el sobretiempo depende del supervisor de cada empleado.*

Atributo SOBRETMP con CARGAS

Tabla de Contingencia

Tabla 3.10
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Tabla de Contingencia de SOBRETMP con CARGAS

		CARGAS						
		0	1	2	3	4	5	Total
SOBRETMP	1	184	74	71	58	29		416
	2	3932	1587	2258	1313	880	32	10002
	3	1442	745	963	539	336	224	4249
	4	655	381	491	255	129	246	2157
Total		6213	2787	3783	2165	1374	502	16824

Contraste de Hipótesis

H_0: El factor SOBRETMP es independiente del factor CARGAS

H_1: ┐ H_0

Tabla 3.11
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Contraste de Hipótesis de SOBRETMP con CARGAS

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	928.613[a]	15	.000
Likelihood Ratio	865.899	15	.000
Linear-by-Linear Association	195.987	1	.000
N of Valid Cases	16824		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.41.

El nivel de significancia obtenido es 0, el cual me permite rechazar H_0 a favor de H_1, *en otras palabras se puede decir que, el sobretiempo depende de las cargas familiares que tiene cada empleado.*

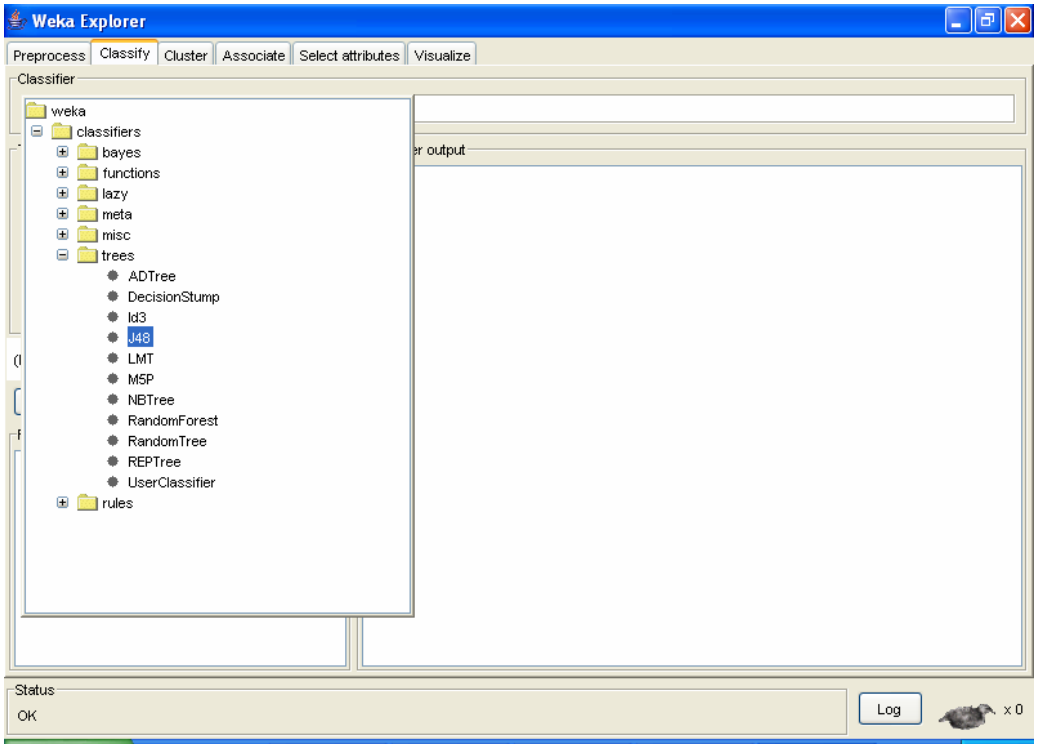
Aplicando el test de independencia chi-cuadrado he podido verificar que las variables que se utilizan para el estudio de sobretiempo son dependientes, de esta manera puedo decir que no es necesario remover más variables.

A continuación explicaré brevemente los pasos a seguir en la aplicación del programa WEKA.

3.5.2.3 Aplicación de algoritmo J4.8

De los muchos algoritmos de aprendizaje que WEKA implementa voy a trabajar con algoritmos cuya clasificación de datos está basada en árboles de decisión. En particular para análisis de datos nominales esta J4.8, el cual se trata de una implementación propia de WEKA para el algoritmo C4.5.

Figura 3.11
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Selección del Algoritmo J4.8

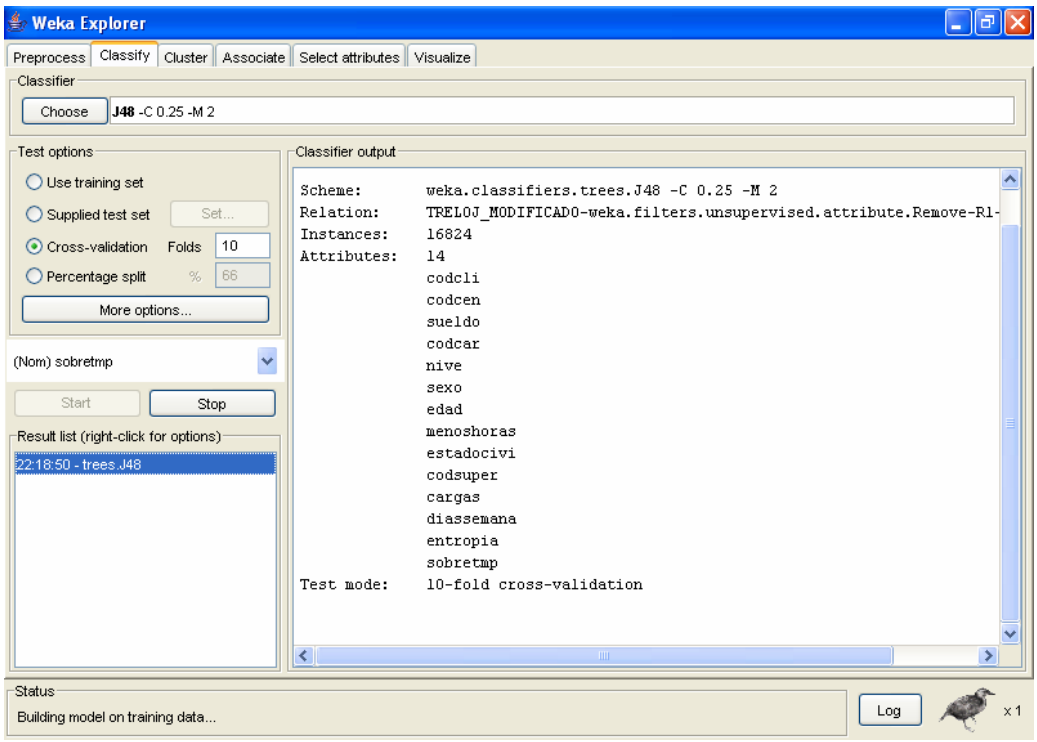


Para comenzar con el experimento de aprendizaje, iré a la sección de Clasificación (Classify), desde la cual se puede configurar los distintos parámetros del experimento.

Lo primero será seleccionar el clasificador, para esto voy a la carpeta de árboles (trees) y escojo el algoritmo J4.8, tal como se muestra en la Figura 3.11.

En la ventana mostrada anteriormente podemos ver la inmensa cantidad de algoritmos de aprendizaje que tiene implementados este software.

Figura 3.12
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Aplicación del Algoritmo J4.8



Después de la breve explicación acerca del entorno de WEKA y una vez configurado todo, solo queda seleccionar el campo que quiero aprender, en este caso “sobretmp” que corresponde al atributo de clases y pulso el botón “Start” para proceder al aprendizaje o aplicación del algoritmo J4.8, del cual se obtuvo la siguiente información:

Figura 3.13

Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio

Árbol de Decisión obtenido mediante la utilización del Algoritmo J4.8

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

J48 pruned tree
------------------

codsuper = 0001: 2 (278.0/87.0)
codsuper = 0002: 3 (17.0/6.0)
codsuper = 0003
|   menoshoras <= 1: 2 (110.0/54.0)
|   menoshoras > 1: 3 (10.0/4.0)
codsuper = 0004
|   edad <= 35
|   |   edad <= 34: 2 (43.0/15.0)
|   |   edad > 34: 3 (35.0/17.0)
|   edad > 35: 2 (37.0/17.0)
codsuper = 0007: 2 (398.0/191.0)
codsuper = 0008
|   edad <= 53
|   |   edad <= 33: 2 (104.0/36.0)
|   |   edad > 33: 3 (171.0/81.0)
|   edad > 53: 4 (236.0/22.0)
codsuper = 0009: 2 (960.0/308.0)
codsuper = 0010: 2 (320.0/73.0)
codsuper = 0012: 2 (81.0/9.0)
codsuper = 01: 2 (2546.0/928.0)
codsuper = 02: 2 (3420.0/1360.0)
codsuper = 03: 2 (4100.0/1306.0)
codsuper = 0486
|   entropia <= 0
|   |   edad <= 30
|   |   |   edad <= 27: 2 (27.0/12.0)
|   |   |   edad > 27
|   |   |   |   menoshoras <= 2
|   |   |   |   |   menoshoras <= 0: 4 (81.0/43.0)
|   |   |   |   |   menoshoras > 0: 2 (3.0/1.0)
|   |   |   |   menoshoras > 2: 3 (4.0/1.0)
|   |   edad > 30: 3 (24.0/6.0)
|   entropia > 0: 2 (14.0/4.0)
codsuper = 0541: 2 (288.0/151.0)
codsuper = 0864: 2 (207.0/73.0)
codsuper = 10
|   estadocivi = C: 2 (211.0/39.0)
|   estadocivi = S
|   |   menoshoras <= 0
|   |   |   codcar = 02: 2 (99.0/51.0)
|   |   |   codcar = 04
|   |   |   |   codcen = 113
|   |   |   |   |   entropia <= 0
|   |   |   |   |   |   edad <= 30: 3 (10.0/4.0)
```

Figura 3.13
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
|  |  |  |  |  |  edad > 30: 2 (20.0/10.0)
|  |  |  |  |  entropia > 0: 2 (15.0/7.0)
|  |  |  |  codcen = 116
|  |  |  |  |  diassemana <= 6
|  |  |  |  |  |  edad <= 22: 2 (2.0/1.0)
|  |  |  |  |  |  edad > 22: 1 (6.0/3.0)
|  |  |  |  |  diassemana > 6
|  |  |  |  |  |  cargas = 0
|  |  |  |  |  |  |  edad <= 26
|  |  |  |  |  |  |  |  edad <= 25
|  |  |  |  |  |  |  |  |  edad <= 24
|  |  |  |  |  |  |  |  |  |  edad <= 23: 2 (17.0/7.0)
|  |  |  |  |  |  |  |  |  |  edad > 23: 3 (24.0/11.0)
|  |  |  |  |  |  |  |  |  edad > 24: 2 (16.0/4.0)
|  |  |  |  |  |  |  |  edad > 25: 3 (11.0)
|  |  |  |  |  |  |  edad > 26: 2 (16.0/4.0)
|  |  |  |  |  |  cargas = 1: 2 (1.0)
|  |  |  |  |  |  cargas = 4: 2 (96.0/45.0)
|  |  |  |  codcen = 117
|  |  |  |  |  edad <= 34
|  |  |  |  |  |  cargas = 0
|  |  |  |  |  |  |  edad <= 28: 2 (36.0/14.0)
|  |  |  |  |  |  |  edad > 28: 3 (10.0/1.0)
|  |  |  |  |  |  cargas = 1: 2 (60.0/31.0)
|  |  |  |  |  edad > 34
|  |  |  |  |  |  entropia <= 0: 3 (16.0/1.0)
|  |  |  |  |  |  entropia > 0: 2 (3.0)
|  |  |  |  codcen = 122
|  |  |  |  |  entropia <= 0
|  |  |  |  |  |  edad <= 24: 2 (3.0/1.0)
|  |  |  |  |  |  edad > 24: 1 (6.0/2.0)
|  |  |  |  |  entropia > 0: 3 (2.0)
|  |  |  codcar = 05
|  |  |  |  codcen = 116: 2 (7.0/1.0)
|  |  |  |  codcen = 117: 1 (3.0/1.0)
|  |  |  codcar = 109: 2 (38.0/16.0)
|  |  |  codcar = 120: 2 (16.0/6.0)
|  |  |  codcar = 121: 1 (2.0/1.0)
|  |  |  codcar = 122
|  |  |  |  codcen = 117: 2 (40.0/15.0)
|  |  |  |  codcen = 123
|  |  |  |  |  edad <= 20: 2 (12.0/2.0)
|  |  |  |  |  edad > 20: 1 (4.0/1.0)
|  |  |  codcar = 16: 2 (30.0/15.0)
|  |  |  codcar = 41
|  |  |  |  edad <= 28: 2 (9.0/2.0)
|  |  |  |  edad > 28: 1 (3.0)
```

Figura 3.13
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
|  |  |  codcar = 53: 1 (2.0)
|  |  |  codcar = 55
|  |  |  |  edad <= 49: 4 (14.0/2.0)
|  |  |  |  edad > 49: 2 (18.0/10.0)
|  |  menoshoras > 0
|  |  |  codcen = 113
|  |  |  |  entropia <= 0: 1 (27.0/4.0)
|  |  |  |  entropia > 0
|  |  |  |  |  menoshoras <= 1: 3 (4.0/2.0)
|  |  |  |  |  menoshoras > 1: 1 (5.0/2.0)
|  |  |  codcen = 116
|  |  |  |  menoshoras <= 3
|  |  |  |  |  edad <= 23
|  |  |  |  |  |  menoshoras <= 2: 1 (4.0/1.0)
|  |  |  |  |  |  menoshoras > 2: 2 (5.0/1.0)
|  |  |  |  |  edad > 23: 1 (27.0/11.0)
|  |  |  |  menoshoras > 3: 2 (13.0/2.0)
|  |  |  codcen = 117: 2 (44.0/15.0)
|  |  |  codcen = 122: 2 (1.0)
|  |  |  codcen = 123
|  |  |  |  edad <= 20: 2 (6.0/2.0)
|  |  |  |  edad > 20: 1 (3.0/1.0)
|  |  |  codcen = 125: 2 (2.0)
|  |  |  codcen = 130
|  |  |  |  diassemana <= 6: 1 (2.0/1.0)
|  |  |  |  diassemana > 6
|  |  |  |  |  menoshoras <= 1: 4 (4.0/1.0)
|  |  |  |  |  menoshoras > 1: 2 (10.0/1.0)
codsuper = 1665
|  nive = 0: 1 (3.0)
|  nive = 5
|  |  edad <= 35
|  |  |  edad <= 34: 2 (484.0/119.0)
|  |  |  edad > 34: 3 (51.0/27.0)
|  |  edad > 35: 2 (72.0/21.0)
codsuper = 1750
|  cargas = 0
|  |  codcar = 100: 2 (7.0)
|  |  codcar = 104: 3 (30.0/3.0)
|  |  codcar = 57
|  |  |  edad <= 30: 3 (92.0/47.0)
|  |  |  edad > 30: 2 (76.0/34.0)
|  |  codcar = 58: 2 (0.0)
|  |  codcar = 59: 2 (98.0/33.0)
|  cargas = 1: 3 (228.0/113.0)
|  cargas = 2
|  |  edad <= 26
```

Figura 3.13
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
|  |  |  edad <= 25: 2 (23.0/14.0)
|  |  |  edad > 25: 4 (52.0/4.0)
|  |  edad > 26: 3 (166.0/42.0)
|  cargas = 4: 2 (30.0/3.0)
|  cargas = 5: 3 (245.0/51.0)
codsuper = 20
|  menoshoras <= 1: 2 (375.0/180.0)
|  menoshoras > 1
|  |  menoshoras <= 3
|  |  |  codcen = 113
|  |  |  |  edad <= 41: 1 (40.0/8.0)
|  |  |  |  edad > 41: 2 (2.0)
|  |  |  codcen = 116: 2 (3.0/1.0)
|  |  |  codcen = 117: 2 (2.0)
|  |  |  codcen = 130: 2 (3.0/1.0)
|  |  menoshoras > 3: 2 (37.0/8.0)
codsuper = 30
|  menoshoras <= 0: 2 (56.0/23.0)
|  menoshoras > 0
|  |  menoshoras <= 3
|  |  |  edad <= 21
|  |  |  |  menoshoras <= 2: 1 (2.0)
|  |  |  |  menoshoras > 2: 2 (2.0)
|  |  |  edad > 21: 1 (19.0/3.0)
|  |  menoshoras > 3: 2 (8.0/1.0)
codsuper = 50: 4 (21.0/7.0)
codsuper = 60: 4 (45.0/13.0)

Number of Leaves  :         397
Size of the tree :  457
Time taken to build model: 111.27 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10749               63.8909 %
Incorrectly Classified Instances      6075               36.1091 %
Kappa statistic                          0.194
Mean absolute error                      0.2534
Root mean squared error                  0.3584
Relative absolute error                 89.5592 %
Root relative squared error             95.2877 %
Total Number of Instances            16824

=== Confusion Matrix ===

   a     b     c     d   <-- classified as
 105   291    20     0 |   a = 1
  53  9618   297    34 |   b = 2
  15  3484   675    75 |   c = 3
   5  1626   175   351 |   d = 4
```

Para estimar el error o determinar el rendimiento del algoritmo se utilizó validación cruzada. Si el conjunto de entrenamiento tiene pocos subconjuntos no es aconsejable *el uso del estimador de validación cruzada por conjunto de prueba ya que reduce aún más el tamaño efectivo del conjunto de aprendizaje. Razón por la cual suele utilizarse como mínimo validación cruzada con V = 10 conjuntos, valor que se utilizó cuando se aplicó el algoritmo J4.8.*

3.6 Evaluación, interpretación, transformación y representación de los patrones extraídos

La Matriz de Confusión muestra el tipo de predicciones correctas e incorrectas cuando se aplico el modelo sobre el conjunto de prueba (Tabla 3.12).

Las predicciones correctas están representadas por los valores que aparecen sobre la diagonal, sumando así 10,749 registros, los cuales están clasificados de la siguiente manera:

- 105 registros corresponden a la clase 1,
- 9,618 registros corresponden a la clase 2,
- 675 registros corresponden a la clase 3 y
- 351 registros corresponden a la clase 4.

El resto de los valores indican el tipo de error cometido (qué valor ha predicho el modelo y cual es el valor verdadero). Los valores en la Matriz de Confusión son proporcionales al peso de los registros que representan.

Tabla 3.12
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Matriz de Confusión

a	**b**	**c**	**d**	
105	291	20	0	a = 1
53	**9618**	297	34	b = 2
15	3484	**675**	75	c = 3
5	1626	175	**351**	d = 4

Después de obtener el algoritmo J4.8, podemos ver que el árbol de decisión se clasificó o ramificó primeramente por el atributo *codsuper* (Supervisor).

Para mostrar la clasificación obtenida de manera clara y sencilla utilizaré una tabla, señalando con una X los atributos que intervinieron en la clasificación del atributo *Supervisor*, para ello seleccione los atributos que hicieron la primera ramificación, de esta manera puedo dar una idea de los atributos mas relevantes que intervienen en las causas de los sobretiempos de los empleados.

Tabla 3.13
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Clasificación del atributo Supervisor

Supervisor	Supervisor	Horas Perdidas	Edad	Turno	Estado Civil	Nivel de	Cargas Familiares
(0001) ALZA LUCAS	X						
(0002) ANDRADE OSWALDO	X						
(0007) MORENO FERNANDO	X						
(0009) PEREZ ROLANDO	X						
(0010) CUESTA GAVICA CRISTOBAL	X						
(0012) MELO XAVIER	X						
(01) ROSALES PEREZ FREDY	X						
(02) SEGARRA BOLIVAR	X						
(03) MORA QUIJIJE RODOLFO SEGUNDO	X						
(0541) LUCIO GUERRERO VICTOR MANUEL	X						
(0864) PIN RODRIGUEZ EDISON DANILO	X						
(50) CALDERON MORAN FELIX	X						
(60) HERNANDEZ MACIAS EDUARDO	X						
(0003) DEL PEZO ABDON		X					
(20) SALVADOR SARAGURO JOSE MANUEL		X					
(30) CORREA SOLORZANO LUIS ENRIQUE		X					
(0004) LEON ENRIQUE			X				
(0008) ORTIZ JOSE			X				
(0486) PILLOAJO ORTEGA GENARO				X			
(10) RODRIGUEZ DASTON APOLINARIO					X		
(1665) DE LUNA VINCES ENRIQUE GAL						X	
(1750) MENDOZA PAZ SOCRATES CONTARDO							X

3.6.1 Resumen de Datos Obtenidos

Este resumen es realizado con la finalidad de dar una breve explicación del comportamiento del atributo *Supervisor*, siendo esto de mucha ayuda para los dueños, jefes o administradores de una empresa en la toma de decisiones.

La tabla siguiente le indicará la *cantidad de hojas* en las que se clasificó el atributo *Supervisor* hasta obtener el resultado final, mientras mas alta es la cantidad la complejidad para la explicación del atributo es mayor, caso contrario es menor.

Además, presentare a continuación los *atributos que* intervinieron en la clasificación de cada Supervisor no solo el atributo que hizo la primera ramificación, sino todos los atributos que intervinieron en la ramificación, también se muestra *el número de empleados* que realizan horas extras *y* el *porcentaje* correspondiente en las diferentes *clases.*

Es importante recalcar, que los datos mostrados en la Tabla 3.14, indican la clase mayoritaria de cada Supervisor, razón por la cual al sumar el número de empleados nos da como resultado 5,946 registros.

Tabla 3.14
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Resumen de datos obtenidos

Supervisor	Cantidad de Hojas	Número de Empleados	Clase (%)				Atributos que intervienen en la clasificación
			1	2	3	4	
(0001) ALZA LUCAS	1	87		100			
(0002) ANDRADE OSWALDO	1	6			100		
(0007) MORENO FERNANDO	1	191		100			
(0009) PEREZ ROLANDO	1	308		100			
(0010) CUESTA GAVICA CRISTOBAL	1	73		100			
(0012) MELO XAVIER	1	9		100			
(01) ROSALES PEREZ FREDY	1	928		100			
(02) SEGARRA BOLIVAR	1	1360		100			
(03) MORA QUIJIJE RODOLFO SEGUNDO	1	1306		100			
(0541) LUCIO GUERRERO VICTOR MANUEL	1	151		100			
(0864) PIN RODRIGUEZ EDISON DANILO	1	73		100			
(50) CALDERON MORAN FELIX	1	7				100	
(60) HERNANDEZ MACIAS EDUARDO	1	13				100	
(0003) DEL PEZO ABDON	2	58		93	7		Edad
(20) SALVADOR SARAGURO JOSE MANUEL	5	202	4	96			Horas Perdidas Centro Costo Edad
(30) CORREA SOLORZANO LUIS ENRIQUE	5	31	16	84			Horas Perdidas Edad
(0004) LEON ENRIQUE	2	49		65	25		Edad
(0008) ORTIZ JOSE	3	139		26	58	16	Edad
(0486) PILLOAJO ORTEGA GENARO	6	67		25	11	64	Turno Variable Edad Horas Perdidas

Tabla 3.14
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

Supervisor	Cantidad de Hojas	Número de Empleados	Clase (%				Atributos que intervienen en la clasificación
			1	2	3	4	
(1665) DE LUNA VINCES ENRIQUE GAL	4	170	2	82	16		Nivel de Sueldo Edad
(1750) MENDOZA PAZ SOCRATES CONTARDO	4	351		26	73	1	Carga Familiar Cargo Edad
(10) RODRIGUEZ DASTON APOLINARIO	11	367	9	82	8	1	Estado Civil Horas Perdidas Cargo Centro Costo Turno Variable Edad Días que labora Carga Familiar

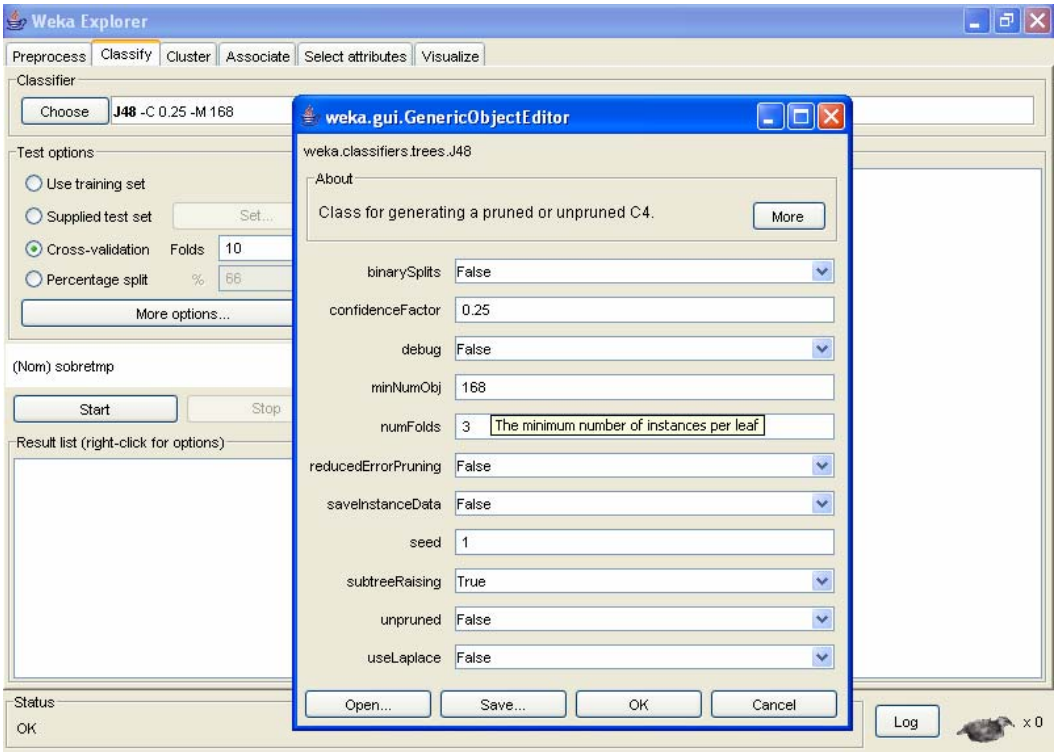
Como podemos observar, la clasificación del Supervisor 10 (RODRIGUEZ DASTON APOLINARIO) es la más compleja, debido a la cantidad de hojas mostradas y atributos que intervienen para su explicación.

3.6.2 Modificación del Algoritmo J4.8

Debido a que el Árbol de Decisión obtenido de la aplicación del Algoritmo J4.8 es de gran tamaño, es necesario modificar los parámetros definidos por defecto por el programa WEKA, especialmente en lo relativo al mínimo número de instancias con que debe contar una hoja, el cual se ha establecido en 168 instancias en vez de 2 instancias, que corresponde al 1% de registros filtrados. El resultado de esta modificación es que desaparecen ciertas hojas, de esta manera será menos complicado analizar la información obtenida.

Para modificar este parámetro se da click sobre el nombre del Algoritmo **J4.8** –C 0.25 –M 2 que aparece en la parte superior, inmediatamente aparecerá un cuadro con los parámetros que pueden ser modificados, de los cuales sólo se modifica minNumObj de 2 a 168 instancias y a continuación se presiona “OK”. Luego de configurados los parámetros se pulsa “Start”.

Figura 3.14
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Modificación de parámetros del Algoritmo J4.8



El árbol obtenido de la modificación del parámetro minNumObj es más fácil de entender e interpretar. El resultado obtenido es el siguiente:

Figura 3.15
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Árbol de Decisión obtenido de la modificación del parámetro minNumObj del Algoritmo J4.8

J48 pruned tree

----------	Nodo
codsuper = 0001: 2 (278.0/87.0)	1
codsuper = 0002: 3 (17.0/6.0)	2
codsuper = 0003: 2 (120.0/60.0)	3
codsuper = 0004: 2 (115.0/54.0)	4
codsuper = 0007: 2 (398.0/191.0)	5

Figura 3.15
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
codsuper = 0008                                  6
|   edad <= 53: 2 (275.0/156.0)                  7
|   edad > 53: 4 (236.0/22.0)                    8
codsuper = 0009: 2 (960.0/308.0)                 9
codsuper = 0010: 2 (320.0/73.0)                  10
codsuper = 0012: 2 (81.0/9.0)                    11
codsuper = 01: 2 (2546.0/928.0)                  12
codsuper = 02: 2 (3420.0/1360.0)                 13
codsuper = 03: 2 (4100.0/1306.0)                 14
codsuper = 0486: 3 (153.0/94.0)                  15
codsuper = 0541: 2 (288.0/151.0)                 16
codsuper = 0864: 2 (207.0/73.0)                  17
codsuper = 10: 2 (1038.0/457.0)                  18
codsuper = 1665: 2 (610.0/177.0)                 19
codsuper = 1750                                  20
|   cargas = 0: 2 (303.0/152.0)                  21
|   cargas = 1: 3 (228.0/113.0)                  22
|   cargas = 2: 3 (241.0/106.0)                  23
|   cargas = 4: 2 (30.0/3.0)                     24
|   cargas = 5: 3 (245.0/51.0)                   25
codsuper = 20: 2 (462.0/225.0)                   26
codsuper = 30: 2 (87.0/43.0)                     27
codsuper = 50: 4 (21.0/7.0)                      28
codsuper = 60: 4 (45.0/13.0)                     29

Number of Leaves  :     28
Size of the tree :      31
Time taken to build model: 3.33 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10574               62.8507 %
Incorrectly Classified Instances      6250               37.1493 %
Kappa statistic                          0.143
Mean absolute error                      0.2608
Root mean squared error                  0.3615
Relative absolute error                 92.1804 %
Root relative squared error             96.1171 %
Total Number of Instances            16824

=== Confusion Matrix ===

   a     b     c     d   <-- classified as
   0   408     8     0 |   a = 1
   0  9778   219     5 |   b = 2
   0  3676   536    37 |   c = 3
   0  1695   202   260 |   d = 4
```

Las predicciones correctas representadas por los valores que aparecen sobre la diagonal de la Matriz de Confusión suman 10,574 registros, los cuales están clasificados de la siguiente manera:

- 0 registros corresponden a la clase 1,
- 9,778 registros corresponden a la clase 2,
- 536 registros corresponden a la clase 3 y
- 260 registros corresponden a la clase 4.

3.6.2.1 Cuadro de Ganancias por nodos de cada categoría

La información que se obtiene del Cuadro de Ganancias por nodos, le permite tomar importantes decisiones al dueño o administrador de la empresa, enfocándose en el grupo de empleados que más sobretiempo realiza.

Las Tablas 3.15, 3.16 y 3.17 muestran el cuadro de las ganancias por cada nodo final del Árbol de Decisión mostrado en la Figura 3.15, agrupados por clase y ordenados en sentido descendente, de mayor a menor proporción dependiendo del sobretiempo que hacen los empleados.

Antes de mostrar las tablas explicaré el significado de cada columna:

- ✓ *Nodo n.* Indica el tamaño de cada nodo de una determinada clase.

- ✓ *Nodo %.* Expresa la proporción que cada nodo representa sobre el total de registros de una determinada clase (suma de *Nodo n*).
- ✓ *Resp n:* Recoge el número de casos pertenecientes a la categoría o clase que se esta analizando presentes en cada nodo.
- ✓ *Resp %:* Expresa la proporción que una determinada categoría representa en cada nodo sobre el total de la categoría que se está analizando (suma de *Resp n*).
- ✓ *Ganancia (%):* Indica la proporción de *Resp n* sobre *Nodo n,* calculado de la misma forma para la sección de Nodo por Nodo como para el Acumulado.
- ✓ *Index (%):* Indica la relación existente entre la proporción de ganancia de cada nodo de una clase determinada sobre el porcentaje obtenido de la suma de *Resp n* sobre la suma de *Nodo n.*

Tabla 3.15
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Cuadro de Ganancias por nodo de la Clase 2

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
7	275	1.76	156	2.68	**56.73**	152.61	275	1.76	156	2.68	56.73	152.61
16	288	1.84	151	2.60	52.43	141.05	563	3.60	307	5.28	54.53	146.69
21	303	1.94	152	2.61	50.17	134.95	866	5.54	459	7.89	53.00	142.59
3	120	0.77	60	1.03	50.00	134.51	986	6.31	519	8.92	52.64	141.60
27	87	0.56	43	0.74	49.43	132.96	1,073	6.86	562	9.66	52.38	140.90
26	462	2.95	225	3.87	48.70	131.02	1,535	9.82	787	13.53	51.27	137.93
5	398	2.55	191	3.29	47.99	129.10	1,933	12.36	978	16.82	50.59	136.11
4	115	0.74	54	0.93	46.96	126.32	2,048	13.10	1,032	17.75	50.39	135.56
18	1,038	6.64	457	7.86	44.03	118.44	3,086	19.74	1,489	25.61	48.25	129.80
13	3,420	21.87	1,360	23.40	39.77	106.98	6,506	**41.61**	2,849	**49.01**	43.79	117.80
12	2,546	16.28	928	15.96	36.45	**98.06**	9,052	57.89	3,777	64.97	41.73	112.25

Tabla 3.15

Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio

Continuación

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
17	207	1.32	73	1.26	35.27	**94.87**	9,259	59.21	3,850	66.23	41.58	111.86
9	960	6.14	308	5.30	32.08	**86.31**	10,219	65.35	4,158	71.53	40.69	109.46
14	4,100	26.22	1,306	22.47	31.85	**85.69**	14,319	91.57	5,464	93.99	38.16	102.65
1	278	1.78	87	1.50	31.29	**84.19**	14,597	93.34	5,551	95.49	38.03	102.30
19	610	3.90	177	3.04	29.02	**78.06**	15,207	97.25	5,728	98.53	37.67	101.33
10	320	2.05	73	1.26	22.81	**61.37**	15,527	99.29	5,801	99.79	37.36	100.51
11	81	0.52	9	0.15	11.11	**29.89**	15,608	99.81	5,810	99.94	37.22	100.14
24	30	0.19	3	0.05	10.00	**26.90**	15,638	100.00	5,813	100.00	37.17	100.00

Tabla 3.16

Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio

Cuadro de Ganancias por nodo de la Clase 3

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
15	153	17.31	94	25.41	**61.44**	146.77	153	17.31	94	25.41	61.44	146.77
22	228	25.79	113	30.54	49.56	118.40	381	43.10	207	55.95	54.33	129.79
23	241	27.26	106	28.65	43.98	105.07	622	**70.36**	313	**84.60**	50.32	120.21
2	17	1.92	6	1.62	35.29	**84.31**	639	72.29	319	86.22	49.92	119.26
25	245	27.71	51	13.78	20.82	**49.73**	884	100.00	370	100.00	41.86	100.00

Tabla 3.17

Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio

Cuadro de Ganancias por nodo de la Clase 4

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
28	21	6.95	7	16.67	33.33	239.64	21	6.95	7	16.67	33.33	239.64
29	45	14.90	13	30.95	28.89	207.68	66	**21.85**	20	**47.62**	30.30	217.85
8	236	78.15	22	52.38	9.32	**67.02**	302	100.00	42	100.00	13.91	100

3.6.2.2 Análisis del Cuadro de Ganancias por nodo de cada categoría

De la Tabla 3.15 se puede observar que los nodos 12, 17, 9, 14, 1, 19, 10, 11 y 24 ostentan una proporción de ganancia inferior a la que presenta la totalidad de la categoría 2 que se está analizando (suma de *Resp n*), por lo que su índice individual asociado es inferior a 100, tal como aparece reflejado en la columna Index de la sección Nodo por Nodo. Estos grupos serían menos interesantes para el análisis de sobretiempo que se desea realizar. En cambio, si se observa el resto de nodos, se alcanza un 41.61 por ciento de la columna *Nodo %* en la sección de *Acumulado* y la columna *Resp %* nos dice que estos nodos agrupan el 49.01 por ciento de los empleados que hacen sobretiempo de la clase 2.

De la Tabla 3.16 se puede observar que los nodos 2 y 25 ostentan una proporción de ganancia inferior a la que presenta la totalidad de la categoría 3 que se está analizando (suma de *Resp n*), por lo que su índice individual asociado es inferior a 100, tal como aparece reflejado en la columna Index de la sección Nodo por Nodo. Estos grupos serían menos interesantes para el análisis de sobretiempo que se desea realizar. En cambio, si se observa el resto de nodos, se alcanza un 70.36 por ciento de la columna *Nodo %* en la sección de *Acumulado* y la columna *Resp %* nos dice que estos nodos agrupan el 84.60 por ciento de los empleados que hacen sobretiempo de la clase 3.

De la Tabla 3.17 se puede observar que el nodo 8 ostenta una proporción de ganancia inferior a la que presenta la totalidad de la categoría 4 que se está analizando (suma de *Resp n*), por lo que su índice individual asociado es inferior a 100, tal como aparece reflejado en la columna Index de la sección Nodo por Nodo. Este grupo sería menos interesante para el análisis de sobretiempo que se desea realizar. En cambio, si se observa el resto de nodos, se alcanza un 21.85 por ciento de la columna *Nodo %* en la sección de *Acumulado* y la columna *Resp %* nos dice que estos nodos agrupan el 47.62 por ciento de los empleados que hacen sobretiempo de la clase 4.

3.6.2.3 Nodos a destacar según el Análisis del Cuadro de Ganancias por nodo de cada categoría

De las Tabla 3.15, 3.16 y 3.17 destacamos los nodos más importantes y el *codsuper* (Supervisor) que se asignó a ese nodo, según el Árbol de Decisión mostrado en la Figura 3.15.

Tabla 3.18
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Nodos a destacar según Cuadro de Ganancias de la Clase 2

Nodo	Supervisor	Característica
7	**(0008) ORTIZ JOSE**	Empleados que tienen hasta 53 años de edad
16	(0541) LUCIO GUERRERO VICTOR MANUEL	

Tabla 3.18
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

Nodo	Supervisor	Característica
21	(1750) MENDOZA PAZ SOCRATES CONTARDO	Empleados que no tienen cargas familiares
3	(0003) DEL PEZO ABDON	
27	(30) CORREA SOLORZANO LUIS ENRIQUE	
26	(20) SALVADOR SARAGURO JOSE MANUEL	
5	(0007) MORENO FERNANDO	
4	(0004) LEON ENRIQUE	
18	(10) RODRIGUEZ DASTON APOLINARIO	
13	(02) SEGARRA BOLIVAR	

Tabla 3.19
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Nodos a destacar según Cuadro de Ganancias de la Clase 3

Nodo	Supervisor	Característica
15	**(0486) PILLOAJO ORTEGA GENARO**	
22	(1750) MENDOZA PAZ SOCRATES CONTARDO	Empleados que tienen una carga familiar
23	(1750) MENDOZA PAZ SOCRATES CONTARDO	Empleados que tienen dos cargas familiares

Tabla 3.20
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Nodos a destacar según Cuadro de Ganancias de la Clase 4

Nodo	Supervisor	Característica
28	(50) CALDERON MORAN FELIX	
29	(60) HERNANDEZ MACIAS EDUARDO	

De todos estos nodos el más importante es el nodo 15 que corresponde al Supervisor **(0486) PILLOAJO ORTEGA GENARO** con una proporción de ganancia del 61.44 por ciento de la muestra, seguido del nodo 7 que corresponde al Supervisor **(0008) ORTIZ JOSE** con una proporción de ganancia del 56.73 por ciento de la muestra, según la Tabla 3.16 y 3.15 respectivamente.

Los empleados que están que están en estos grupos son los que realizan más sobretiempo, a diferencia de los otros grupos.

3.7 Aplicación del Algoritmo J4.8 con sobre-muestreo

Para que todas las clases tengan la misma probabilidad de ser seleccionadas se aplica el sobre-muestreo, de esta manera se evita el sesgo de los datos hacia una determinada clase, como en el caso anterior donde los empleados que realizan más sobretiempo correspondían a la Clase 2.

Para hacer el sobre-muestreo se creó el siguiente query, donde se asignan más registros de las diferentes clases de forma aleatoria, hasta que todos

tengan la misma cantidad de registros, tomando como base la Clase 2 con 10,002 registros.

Figura 3.16
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Query utilizado para hacer el sobre-muestreo a la tabla Treloj modificado

```
clear
close DATABASES

use TReloj_modificado ALIAS TReloj_modificado
Copy stru to Tmp
Select 0
use Tmp

SELECT treloj.Emp_id, treloj.semana, treloj.codcli, treloj.codcen, treloj.sueldo, ;
treloj.codcar, treloj.nive, treloj.sexo, treloj.Edad, treloj.MenosHoras, ;
treloj.estadocivi, treloj.plantaadmi, treloj.codsuper, treloj.cargas, treloj.DiasSemana, ;
treloj.Entropia, 00.0000 EntroAnt, treloj.sobre, 00.0000 SobreAnt, treloj.SobreTmp ;
FROM TReloj_modificado treloj  ;
WHERE treloj.SobreTmp  = '2';
GROUP BY treloj.Emp_id, treloj.semana;
INTO cursor clase2
nClase2 = Reccount()

* Sobremuestreo para la clase 1
SELECT treloj.Emp_id, treloj.semana, treloj.codcli, treloj.codcen, treloj.sueldo, ;
treloj.codcar, treloj.nive, treloj.sexo, treloj.Edad, treloj.MenosHoras, ;
treloj.estadocivi, treloj.plantaadmi, treloj.codsuper, treloj.cargas, treloj.DiasSemana, ;
treloj.Entropia, 00.0000 EntroAnt, treloj.sobre, 00.0000 SobreAnt, treloj.SobreTmp ;
FROM TReloj_modificado treloj  ;
WHERE treloj.SobreTmp  = '1';
GROUP BY treloj.Emp_id, treloj.semana;
INTO cursor clase1
nClase1 = Reccount()

For i=0 To nClase2 - nClase1 - 1
   nRecno = Int( nClase1*Rand() ) + 1
   Selec Clase1
   Go nRecno
   * Grabo el registro de Clase1  al arreglo aReloj
   Scatter To aReloj
   Select Tmp
   * Agrego un registro en blanco a la tabla Tmp
   Append Blank
   * Cargo los valores de aReloj en Tmp
   Gather from aReloj
Endfor
```

Figura 3.16
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
* Sobremuestreo para la clase 3
SELECT treloj.Emp_id, treloj.semana, treloj.codcli, treloj.codcen, treloj.sueldo, ;
treloj.codcar, treloj.nive, treloj.sexo, treloj.Edad, treloj.MenosHoras, ;
treloj.estadocivi, treloj.plantaadmi, treloj.codsuper, treloj.cargas, treloj.DiasSemana, ;
treloj.Entropia, 00.0000 EntroAnt, treloj.sobre, 00.0000 SobreAnt, treloj.SobreTmp ;
FROM TReloj_modificado treloj  ;
WHERE treloj.SobreTmp  = '3';
GROUP BY treloj.Emp_id, treloj.semana;
INTO cursor clase3
nClase3 = Reccount()

For i=0 To nClase2 - nClase3 - 1
   nRecno = Int( nClase3*Rand() ) + 1
   Selec Clase3
   Go nRecno
   * Grabo el registro de Clase3  al arreglo aReloj
   Scatter To aReloj
   Select Tmp
   * Agrego un registro en blanco a la tabla Tmp
   Append Blank
   * Cargo los valores de aReloj en Tmp
   Gather from aReloj
Endfor


* Sobremuestreo para la clase 4
SELECT treloj.Emp_id, treloj.semana, treloj.codcli, treloj.codcen, treloj.sueldo, ;
treloj.codcar, treloj.nive, treloj.sexo, treloj.Edad, treloj.MenosHoras, ;
treloj.estadocivi, treloj.plantaadmi, treloj.codsuper, treloj.cargas, treloj.DiasSemana, ;
treloj.Entropia, 00.0000 EntroAnt, treloj.sobre, 00.0000 SobreAnt, treloj.SobreTmp ;
FROM TReloj_modificado treloj  ;
WHERE treloj.SobreTmp  = '4';
GROUP BY treloj.Emp_id, treloj.semana;
INTO cursor clase4
nClase4 = Reccount()

For i=0 To nClase2 - nClase4 - 1
   nRecno = Int( nClase4*Rand() ) + 1
   Selec Clase4
   Go nRecno
   * Grabo el registro de Clase4  al arreglo aReloj
   Scatter To aReloj
   Select Tmp
   * Agrego un registro en blanco a la tabla Tmp
   Append Blank
   * Cargo los valores de aReloj en Tmp
   Gather from aReloj
Endfor
```

Figura 3.16
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
*Añado valores aleatorios a la tabla Treloj_modificado
Select Treloj_modificado
Append from Tmp
```

Luego de obtenidos los 40,008 registros con sobre-muestreo se aplica el query mostrado en la Figura 3.5, que transforma los datos de formato DBF a formato ARFF.

Estos datos son recuperados en el programa WEKA para aplicar el Algoritmo J4.8 como se hizo con los datos filtrados originalmente, se utiliza los mismos atributos con que se trabajo anteriormente y se cambia el *minNumObj* a 400 instancias por hoja, que corresponde al 1% de 40,008 registros con sobre-muestreo.

El Árbol de Decisión obtenido es el siguiente:

Figura 3.17
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Árbol de Decisión obtenido de datos sobre-muestreados y de la modificación del parámetro minNumObj del Algoritmo J4.8

J48 pruned tree ----------	Nodo
codsuper = 0001: 1 (1451.0/299.0)	1
codsuper = 0002: 3 (34.0/13.0)	2
codsuper = 0003: 1 (479.0/190.0)	3
codsuper = 0004: 1 (348.0/207.0)	4
codsuper = 0007: 4 (911.0/590.0)	5
codsuper = 0008	6
| sueldo <= 169: 3 (687.0/433.0)	7
| sueldo > 169: 4 (1124.0/75.0)	8
codsuper = 0009	9
| edad <= 30: 1 (765.0/400.0)	10
| edad > 30: 2 (1222.0/765.0)	11
codsuper = 0010: 2 (494.0/247.0)	12
codsuper = 0012: 2 (94.0/22.0)	13
codsuper = 01	13
| sueldo <= 160	15
| | entropia <= 0: 2 (2187.0/1232.0)	16
| | entropia > 0: 4 (675.0/389.0)	17
| sueldo > 160	18
| | codcen = 012: 1 (437.0/217.0)	19
| | codcen = 013: 2 (167.0/96.0)	20
| | codcen = 014: 2 (173.0/102.0)	21
| | codcen = 017: 3 (260.0/180.0)	22
| | codcen = 031: 2 (53.0/31.0)	23
| | codcen = 032: 2 (177.0/107.0)	24
| | codcen = 040: 1 (1019.0/412.0)	25
codsuper = 02	26
| codcen = 013	27
| | edad <= 28: 4 (443.0/257.0)	28
| | edad > 28: 3 (459.0/283.0)	29
| codcen = 014: 2 (173.0/102.0)	30
| codcen = 015: 4 (315.0/153.0)	31
| codcen = 016: 4 (653.0/379.0)	32
| codcen = 017: 2 (493.0/285.0)	33
| codcen = 020: 3 (178.0/110.0)	34
| codcen = 024: 2 (262.0/148.0)	35
| codcen = 025: 4 (111.0/54.0)	36
| codcen = 027: 2 (197.0/124.0)	37
| codcen = 030	38
| | cargas = 1: 4 (1104.0/551.0)	39
| | cargas = 3: 3 (844.0/556.0)	40
| codcen = 031: 1 (540.0/269.0)	41
| codcen = 032: 3 (537.0/333.0)	42
| codcen = 034: 2 (226.0/115.0)	43
| codcen = 040: 4 (438.0/306.0)	44

Figura 3.17
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
codsuper = 03                                   45
|   codcen = 012: 1 (256.0/64.0)                46
|   codcen = 013: 2 (66.0/35.0)                 47
|   codcen = 014: 4 (310.0/197.0)               48
|   codcen = 015: 4 (548.0/328.0)               49
|   codcen = 016: 2 (644.0/366.0)               50
|   codcen = 017: 2 (431.0/219.0)               51
|   codcen = 019: 4 (118.0/81.0)                52
|   codcen = 020: 3 (235.0/150.0)               53
|   codcen = 021: 3 (234.0/152.0)               54
|   codcen = 022: 2 (248.0/146.0)               55
|   codcen = 024: 2 (140.0/69.0)                56
|   codcen = 025: 2 (970.0/591.0)               57
|   codcen = 027: 2 (265.0/147.0)               58
|   codcen = 030: 4 (476.0/312.0)               59
|   codcen = 032: 2 (469.0/211.0)               60
|   codcen = 033: 1 (72.0/43.0)                 61
|   codcen = 034: 2 (280.0/157.0)               62
|   codcen = 035: 1 (176.0/122.0)               63
|   codcen = 040: 2 (935.0/507.0)               64
|   codcen = 060: 4 (376.0/215.0)               65
codsuper = 0486: 4 (422.0/196.0)                66
codsuper = 0541: 3 (653.0/400.0)                67
codsuper = 0864: 2 (385.0/251.0)                68
codsuper = 10                                   69
|   menoshoras <= 0                             70
|   |   codcen = 113: 4 (382.0/228.0)           71
|   |   codcen = 116: 1 (672.0/434.0)           72
|   |   codcen = 117: 1 (932.0/512.0)           73
|   |   codcen = 122: 1 (137.0/12.0)            74
|   |   codcen = 123: 1 (96.0/21.0)             75
|   |   codcen = 124: 1 (58.0)                  76
|   |   codcen = 125: 1 (115.0/8.0)             77
|   |   codcen = 130: 4 (155.0/83.0)            78
|   |   codcen = 132: 1 (45.0/20.0)             79
|   |   codcen = 150: 2 (117.0/42.0)            80
|   menoshoras > 0                              81
|   |   codcen = 113: 1 (648.0/27.0)            82
|   |   codcen = 116: 1 (550.0/45.0)            83
|   |   codcen = 117: 1 (142.0/77.0)            84
|   |   codcen = 122: 2 (1.0)                   85
|   |   codcen = 123: 1 (63.0/11.0)             86
|   |   codcen = 125: 2 (2.0)                   87
|   |   codcen = 130: 1 (47.0/22.0)             88
|   |   codcen = 132: 2 (4.0)                   89
|   |   codcen = 150: 2 (93.0/39.0)             90
codsuper = 1665: 2 (994.0/561.0)                91
```

Figura 3.17
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

```
codsuper = 1750                                          92
|   cargas = 0: 1 (972.0/478.0)                         93
|   cargas = 1: 3 (549.0/282.0)                         94
|   cargas = 2: 4 (789.0/331.0)                         95
|   cargas = 4: 2 (37.0/10.0)                           96
|   cargas = 5: 3 (586.0/140.0)                         97
codsuper = 20                                            98
|   cargas = 0: 1 (1042.0/307.0)                        99
|   cargas = 1: 1 (309.0/67.0)                          100
|   cargas = 2: 1 (962.0/188.0)                         101
|   cargas = 3: 1 (42.0/9.0)                            102
|   cargas = 4: 4 (121.0/72.0)                          103
codsuper = 30: 1 (660.0/108.0)                           104
codsuper = 50: 4 (83.0/12.0)                             105
codsuper = 60: 4 (164.0/24.0)                            106

Number of Leaves  :         213

Size of the tree :  228


Time taken to build model: 13.78 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       20599               51.4872 %
Incorrectly Classified Instances     19409               48.5128 %
Kappa statistic                          0.3532
Mean absolute error                      0.2919
Root mean squared error                  0.3826
Relative absolute error                 77.8526 %
Root relative squared error             88.3469 %
Total Number of Instances            40008

=== Confusion Matrix ===

     a     b     c     d   <-- classified as
  8433   464   490   615 |   a = 1
  1705  5409   174  1714 |   b = 2
  1856  3912  2191  2043 |   c = 3
  1244  2927  1265  4566 |   d = 4
```

Las predicciones correctas representadas por los valores que aparecen sobre la diagonal de la Matriz de Confusión suman 20,599 registros, los cuales están clasificados de la siguiente manera:

- 8,433 registros corresponden a la clase 1,
- 5,409 registros corresponden a la clase 2,
- 2,191 registros corresponden a la clase 3 y
- 4,566 registros corresponden a la clase 4.

3.7.1 Cuadro de Ganancias por nodos de cada categoría

Las Tablas 3.21, 3.22, 3.23 y 3.24 muestran el cuadro de las ganancias por cada nodo final del Árbol de Decisión mostrado en la Figura 3.17, agrupados por clase y ordenados en sentido descendente, de mayor a menor proporción dependiendo del sobretiempo que hacen los empleados.

Tabla 3.21

Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio

Cuadro de Ganancias por nodo de la Clase 1

(datos sobremuestreados)

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
76	58	0.44	58	1.25	**100.00**	281.69	58	0.44	58	1.25	100.00	281.69
63	176	1.35	122	2.64	**69.32**	195.26	234	1.79	180	3.89	76.92	216.68
72	672	5.16	434	9.38	**64.58**	181.92	906	6.95	614	13.27	67.77	190.90
61	72	0.55	43	0.93	59.72	168.23	978	7.50	657	14.20	67.18	189.23
4	348	2.67	207	4.47	59.48	167.56	1,326	10.17	864	18.67	65.16	183.54
73	932	7.15	512	11.07	54.94	154.75	2,258	17.32	1,376	29.73	60.94	171.66
84	142	1.09	77	1.66	54.23	152.75	2,400	18.41	1,453	31.40	60.54	170.54
10	765	5.87	400	8.64	52.29	147.29	3,165	24.28	1,853	40.04	58.55	164.92
41	540	4.14	269	5.81	49.81	140.32	3,705	28.42	2,122	45.86	57.27	161.34
19	437	3.35	217	4.69	49.66	139.88	4,142	31.77	2,339	50.55	56.47	159.07
93	972	7.46	478	10.33	49.18	138.53	5,114	39.23	2,817	60.88	55.08	155.17
88	47	0.36	22	0.48	46.81	131.85	5,161	39.59	2,839	61.35	55.01	154.95
79	45	0.35	20	0.43	44.44	125.20	5,206	39.93	2,859	61.79	54.92	154.70
25	1,019	7.82	412	8.90	40.43	113.89	6,225	47.75	3,271	70.69	52.55	148.02
3	479	3.67	190	4.11	39.67	111.74	6,704	**51.43**	3,461	**74.80**	51.63	145.43
99	1,042	7.99	307	6.63	29.46	**82.99**	7,746	59.42	3,768	81.43	48.64	137.03
46	256	1.96	64	1.38	25.00	**70.42**	8,002	61.38	3,832	82.81	47.89	134.90
75	96	0.74	21	0.45	21.88	**61.62**	8,098	62.12	3,853	83.27	47.58	134.03
100	309	2.37	67	1.45	21.68	**61.08**	8,407	64.49	3,920	84.72	46.63	131.35
102	42	0.32	9	0.19	21.43	**60.36**	8,449	64.81	3,929	84.91	46.50	130.99
1	1,451	11.13	299	6.46	20.61	**58.05**	9,900	75.94	4,228	91.37	42.71	120.30
101	962	7.38	188	4.06	19.54	**55.05**	10,862	83.32	4,416	95.44	40.66	114.52
86	63	0.48	11	0.24	17.46	**49.18**	10,925	83.81	4,427	95.67	40.52	114.15
104	660	5.06	108	2.33	16.36	**46.09**	11,585	88.87	4,535	98.01	39.15	110.27
74	137	1.05	12	0.26	8.76	**24.67**	11,722	89.92	4,547	98.27	38.79	109.27
83	550	4.22	45	0.97	8.18	**23.05**	12,272	94.14	4,592	99.24	37.42	105.40
77	115	0.88	8	0.17	6.96	**19.60**	12,387	95.02	4,600	99.41	37.14	104.61
82	648	4.97	27	0.58	4.17	**11.74**	13,035	100.00	4,627	100.00	35.50	100.00

Tabla 3.22
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Cuadro de Ganancias por nodo de la Clase 2
(datos sobremuestreados)

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
85	1	0.01	1	0.01	**100.00**	178.19	1	0.01	1	0.01	100.00	178.19
87	2	0.02	2	0.03	**100.00**	178.19	3	0.03	3	0.04	100.00	178.19
89	4	0.03	4	0.06	**100.00**	178.19	7	0.06	7	0.10	100.00	178.19
68	385	3.21	251	3.73	**65.19**	116.17	392	3.27	258	3.83	65.82	117.28
37	197	1.64	124	1.84	62.94	112.16	589	4.91	382	5.67	64.86	115.57
11	1,222	10.18	765	11.36	62.60	111.55	1,811	15.09	1,147	17.03	63.34	112.86
57	970	8.08	591	8.78	60.93	108.57	2,781	23.18	1,738	25.80	62.50	111.36
24	177	1.48	107	1.59	60.45	107.72	2,958	24.65	1,845	27.39	62.37	111.14
21	173	1.44	102	1.51	58.96	105.06	3,131	26.10	1,947	28.91	62.18	110.81
30	173	1.44	102	1.51	58.96	105.06	3,304	27.54	2,049	30.42	62.02	110.51
55	248	2.07	146	2.17	58.87	104.90	3,552	29.60	2,195	32.59	61.80	110.11
23	53	0.44	31	0.46	58.49	104.22	3,605	30.05	2,226	33.05	61.75	110.03
33	493	4.11	285	4.23	57.81	103.01	4,098	34.15	2,511	37.28	61.27	109.18
20	167	1.39	96	1.43	57.49	102.43	4,265	35.55	2,607	38.71	61.13	108.92
50	644	5.37	366	5.44	56.83	101.27	4,909	40.91	2,973	44.14	60.56	107.92
35	262	2.18	148	2.20	56.49	100.66	5,171	43.10	3,121	46.34	60.36	107.55
91	994	8.28	561	8.33	56.44	100.57	6,165	51.38	3,682	54.67	59.72	106.42
16	2,187	18.23	1,232	18.30	56.33	100.38	8,352	**69.61**	4,914	**72.97**	58.84	104.84
62	280	2.33	157	2.33	56.07	**99.91**	8,632	71.94	5,071	75.30	58.75	104.68
58	265	2.21	147	2.18	55.47	**98.84**	8,897	74.15	5,218	77.48	58.65	104.51
64	935	7.79	507	7.53	54.22	**96.62**	9,832	81.94	5,725	85.01	58.23	103.76
47	66	0.55	35	0.52	53.03	**94.49**	9,898	82.49	5,760	85.53	58.19	103.69
43	226	1.88	115	1.71	50.88	**90.67**	10,124	84.38	5,875	87.24	58.03	103.40
51	431	3.59	219	3.25	50.81	**90.54**	10,555	87.97	6,094	90.49	57.74	102.88
12	494	4.12	247	3.67	50.00	**89.09**	11,049	92.08	6,341	94.16	57.39	102.26
56	140	1.17	69	1.02	49.29	**87.82**	11,189	93.25	6,410	95.18	57.29	102.08
60	469	3.91	211	3.13	44.99	**80.17**	11,658	97.16	6,621	98.32	56.79	101.20
90	93	0.78	39	0.58	41.94	**74.72**	11,751	97.93	6,660	98.90	56.68	100.99
80	117	0.98	42	0.62	35.90	**63.97**	11,868	98.91	6,702	99.52	56.47	100.63
96	37	0.31	10	0.15	27.03	**48.16**	11,905	99.22	6,712	99.67	56.38	100.46
13	94	0.78	22	0.33	23.40	**41.70**	11,999	100.00	6,734	100.00	56.12	100.00

Tabla 3.23
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Cuadro de Ganancias por nodo de la Clase 3
(datos sobremuestreados)

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
22	260	4.95	180	5.94	**69.23**	120.00	260	4.95	180	5.94	69.23	120.00
40	844	16.06	556	18.34	**65.88**	114.19	1,104	21.01	736	24.28	66.67	115.56
54	234	4.45	152	5.01	64.96	112.60	1,338	25.46	888	29.29	66.37	115.04
53	235	4.47	150	4.95	63.83	110.64	1,573	29.93	1,038	34.24	65.99	114.38
7	687	13.07	433	14.28	63.03	109.25	2,260	43.00	1,471	48.52	65.09	112.82
42	537	10.22	333	10.98	62.01	107.49	2,797	53.22	1,804	59.50	64.50	111.80
34	178	3.39	110	3.63	61.80	107.12	2,975	56.61	1,914	63.13	64.34	111.52
29	459	8.73	283	9.33	61.66	106.87	3,434	65.34	2,197	72.46	63.98	110.90
67	653	12.42	400	13.19	61.26	106.18	4,087	**77.76**	2,597	**85.66**	63.54	110.15
94	549	10.45	282	9.30	51.37	**89.04**	4,636	88.21	2,879	94.96	62.10	107.65
2	34	0.65	13	0.43	38.24	**66.28**	4,670	88.85	2,892	95.39	61.93	107.34
97	586	11.15	140	4.62	23.89	**41.41**	5,256	100.00	3,032	100.00	57.69	100.00

Tabla 3.24
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Cuadro de Ganancias por nodo de la Clase 4
(datos sobremuestreados)

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
44	438	4.51	306	6.33	**69.86**	140.48	438	4.51	306	6.33	69.86	140.48
52	118	1.21	81	1.68	**68.64**	138.03	556	5.72	387	8.01	69.60	139.96
59	476	4.90	312	6.46	**65.55**	131.80	1,032	10.62	699	14.46	67.73	136.20
5	911	9.37	590	12.21	64.76	130.23	1,943	20.00	1,289	26.67	66.34	133.40
48	310	3.19	197	4.08	63.55	127.79	2,253	23.19	1,486	30.75	65.96	132.63
49	548	5.64	328	6.79	59.85	120.36	2,801	28.83	1,814	37.53	64.76	130.23
71	382	3.93	228	4.72	59.69	120.02	3,183	32.76	2,042	42.25	64.15	129.00
103	121	1.25	72	1.49	59.50	119.65	3,304	34.00	2,114	43.74	63.98	128.66
32	653	6.72	379	7.84	58.04	116.71	3,957	40.72	2,493	51.58	63.00	126.69
28	443	4.56	257	5.32	58.01	116.66	4,400	45.28	2,750	56.90	62.50	125.68
17	675	6.95	389	8.05	57.63	115.89	5,075	52.23	3,139	64.95	61.85	124.38

Tabla 3.24
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

Nodo por Nodo							Acumulado					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Index (%)
65	376	3.87	215	4.45	57.18	114.98	5,451	56.09	3,354	69.40	61.53	123.73
78	155	1.59	83	1.72	53.55	107.68	5,606	57.69	3,437	71.11	61.31	123.28
39	1,104	11.36	551	11.40	49.91	100.36	6,710	**69.05**	3,988	**82.51**	59.43	119.51
36	111	1.14	54	1.12	48.65	**97.83**	6,821	70.19	4,042	83.63	59.26	119.16
31	315	3.24	153	3.17	48.57	**97.67**	7,136	73.43	4,195	86.80	58.79	118.21
66	422	4.34	196	4.06	46.45	**93.40**	7,558	77.78	4,391	90.85	58.10	116.83
95	789	8.12	331	6.85	41.95	**84.36**	8,347	85.90	4,722	97.70	56.57	113.76
106	164	1.69	24	0.50	14.63	**29.43**	8,511	87.58	4,746	98.20	55.76	112.13
105	83	0.85	12	0.25	14.46	**29.07**	8,594	88.44	4,758	98.45	55.36	111.33
8	1,124	11.57	75	1.55	6.67	**13.42**	9,718	100.00	4,833	100.00	49.73	100.00

3.7.2 Análisis del Cuadro de Ganancias por nodo de cada categoría

De la Tabla 3.21 se puede observar que los nodos 99, 46, 75, 100, 102, 1, 101, 86, 104, 74, 83, 77 y 82 ostentan una proporción de ganancia inferior a la que presenta la totalidad de la categoría 1 que se está analizando (suma de *Resp n*), por lo que su índice individual asociado es inferior a 100, tal como aparece reflejado en la columna Index de la sección Nodo por Nodo. Estos grupos serían menos interesantes para el análisis de sobretiempo que se desea realizar. En cambio, si se observa el resto de nodos, se alcanza un 51.43 por ciento de la columna *Nodo %* en la sección de *Acumulado* y la columna *Resp %* nos dice que estos nodos agrupan el 74.80 por ciento de los empleados que hacen sobretiempo de la clase 1.

De la Tabla 3.22 se puede observar que los nodos 62, 58, 64, 47, 43, 51, 12, 56, 60, 90, 80, 96 y 13 ostentan una proporción de ganancia inferior a la que presenta la totalidad de la categoría 2 que se está analizando (suma de *Resp n*), por lo que su índice individual asociado es inferior a 100, tal como aparece reflejado en la columna Index de la sección Nodo por Nodo. Estos grupos serían menos interesantes para el análisis de sobretiempo que se desea realizar. En cambio, si se observa el resto de nodos, se alcanza un 69.61 por ciento de la columna *Nodo %* en la sección de *Acumulado* y la columna *Resp %* nos dice que estos nodos agrupan el 72.97 por ciento de los empleados que hacen sobretiempo de la clase 2.

De la Tabla 3.23 se puede observar que los nodos 94, 2 y 97 ostentan una proporción de ganancia inferior a la que presenta la totalidad de la categoría 3 que se está analizando (suma de *Resp n*), por lo que su índice individual asociado es inferior a 100, tal como aparece reflejado en la columna Index de la sección Nodo por Nodo. Estos grupos serían menos interesantes para el análisis de sobretiempo que se desea realizar. En cambio, si se observa el resto de nodos, se alcanza un 77.76 por ciento de la columna *Nodo %* en la sección de *Acumulado* y la columna *Resp %* nos dice que estos nodos agrupan el 85.66 por ciento de los empleados que hacen sobretiempo de la clase 3.

De la Tabla 3.24 se puede observar que los nodos 36, 31, 66, 95, 106, 105 y 8 ostentan una proporción de ganancia inferior a la que presenta la totalidad de la categoría 4 que se está analizando (suma de *Resp n*), por lo que su índice individual asociado es inferior a 100, tal como aparece reflejado en la columna Index de la sección Nodo por Nodo. Estos grupos serían menos interesantes para el análisis de sobretiempo que se desea realizar. En cambio, si se observa el resto de nodos, se alcanza un 69.05 por ciento de la columna *Nodo %* en la sección de *Acumulado* y la columna *Resp %* nos dice que estos nodos agrupan el 82.51 por ciento de los empleados que hacen sobretiempo de la clase 4.

3.7.3 Nodos a destacar según el Análisis del Cuadro de Ganancias por nodo de cada categoría

De las Tabla 3.25, 3.26, 3.27 y 3.28 destacamos los nodos más importantes y el *codsuper* (Supervisor) que se asignó a ese nodo, según el Árbol de Decisión mostrado en la Figura 3.17.

Tabla 3.25
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Nodos a destacar según Cuadro de Ganancias de la Clase 1
(datos sobremuestreados)

Nodo	Supervisor	Característica
76	**(10) RODRIGUEZ DASTON APOLINARIO**	Empleados que no tienen horas pérdidas y que trabajan en el Centro Costo (124) SEGURIDAD INDUSTRIAL
63	**(03) MORA QUIJIJE RODOLFO SEGUNDO**	Empleados que laboran en el Centro Costo (035) DESARROLLO
72	(10) RODRIGUEZ DASTON APOLINARIO	Empleados que no tienen horas pérdidas y que trabajan en el Centro Costo (116) IMPRENTA #1
61	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (033) ARTE
4	(0004) LEON ENRIQUE	
73	(10) RODRIGUEZ DASTON APOLINARIO	Empleados que no tienen horas pérdidas y que trabajan en el Centro Costo (117) IMPRENTA #2
84	(10) RODRIGUEZ DASTON APOLINARIO	Empleados que tienen horas pérdidas y que trabajan en el Centro Costo (117) IMPRENTA #2
10	(0009) PEREZ ROLANDO	Empleados que tienen hasta 30 años de edad
41	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (031) MANEJO DE MATERIALES
19	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo mayor a 160 dólares y trabajan en el Centro Costo (012) PREPARACION ALMIDON
93	(1750) MENDOZA PAZ SOCRATES CONTARDO	Empleados que no tienen cargas familiares
88	(10) RODRIGUEZ DASTON APOLINARIO	Empleados que tienen horas pérdidas y que trabajan en el Centro Costo (130) MANTENIMIENTO
79	(10) RODRIGUEZ DASTON APOLINARIO	Empleados que no tienen horas pérdidas y que trabajan en el Centro Costo (132) FABRICA GENERAL
25	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo mayor a 160 dólares y trabajan en el Centro Costo (040) EMPLEADO FCA. GENERAL
3	(0003) DEL PEZO ABDON	

Tabla 3.26
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Nodos a destacar según Cuadro de Ganancias de la Clase 2
(datos sobremuestreados)

Nodo	Supervisor	Característica
85	**(10) RODRIGUEZ DASTON APOLINARIO**	Empleados que tienen horas pérdidas y que trabajan en el Centro Costo (122) HOJEADORA
87	**(10) RODRIGUEZ DASTON APOLINARIO**	Empleados que tienen horas pérdidas y que trabajan en el Centro Costo (125) CONTROL DE PAPEL
89	**(10) RODRIGUEZ DASTON APOLINARIO**	Empleados que tienen horas pérdidas y que trabajan en el Centro Costo (132) FABRICA GENERAL
68	(0864) PIN RODRIGUEZ EDISON DANILO	
37	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (027) IMPRENTA UNITED 3
11	(0009) PEREZ ROLANDO	Empleados que tienen más de 30 años de edad
57	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (025) IMPRENTA WARD
24	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo mayor a 160 dólares y trabajan en el Centro Costo (032) FABRICA GENERAL
21	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo mayor a 160 dólares y trabajan en el Centro Costo (014) CORRUGADORA LANGSTON
30	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (014) CORRUGADORA LANGSTON
55	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (022) CORTDORA DE PADS
23	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo mayor a 160 dólares y trabajan en el Centro Costo (031) MANEJO DE MATERIALES
33	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (017) IMPRENTA UNITED 1
20	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo mayor a 160 dólares y trabajan en el Centro Costo (013) CORRUGADORA S&S
50	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (016) IMPRENTA UNITED 2
35	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (024) IMPRENTA S&S
91	(1665) DE LUNA VINCES ENRIQUE GAL	
16	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo menor o igual a 160 dólares y que no tienen turnos variables.

Tabla 3.27
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Nodos a destacar según Cuadro de Ganancias de la Clase 3
(datos sobremuestreados)

Nodo	Supervisor	Característica
22	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo mayor a 160 dólares y trabajan en el Centro Costo (017) IMPRENTA UNITED 1
40	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (030) MANTENIMIENTO y tienen 3 cargas familiares
54	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (021) CLISE
53	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (020) EMBALADORA
7	(0008) ORTIZ JOSE	Empleados que tienen un sueldo hasta 169 dólares
42	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (032) FABRICA GENERAL
34	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (020) EMBALADORA
29	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (013) CORRUGADORA S&S y tienen más de 28 años de edad
67	(0541) LUCIO GUERRERO VICTOR MANUEL	

Tabla 3.28
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Nodos a destacar según Cuadro de Ganancias de la Clase 4
(datos sobremuestreados)

Nodo	Supervisor	Característica
44	**(02) SEGARRA BOLIVAR**	Empleados que trabajan en el Centro Costo (040) EMPLEADO FCA.GENERAL
52	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (019) ADITAMENTOS
59	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (030) MANTENIMIENTO
5	(0007) MORENO FERNANDO	
48	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (014) CORRUGADORA LANGSTON
49	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (015) IMPRENTA H.S.

Tabla 3.28
Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio
Continuación

Nodo	Supervisor	Característica
71	(10) RODRIGUEZ DASTON APOLINARIO	Empleados que no tienen horas pérdidas y que trabajan en el Centro Costo (113) CORRUGADORA
103	(20) SALVADOR SARAGURO JOSE MANUEL	Empleados que tienen 4 cargas familiares
32	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (016) IMPRENTA UNITED 2
28	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (013) CORRUGADORA S&S y tienen hasta 28 años de edad
17	(01) ROSALES PEREZ FREDY	Empleados que tienen un sueldo menor o igual a 160 dólares y tienen turnos variables.
65	(03) MORA QUIJIJE RODOLFO SEGUNDO	Empleados que laboran en el Centro Costo (060) VENTAS
78	(10) RODRIGUEZ DASTON APOLINARIO	Empleados que no tienen horas pérdidas y que trabajan en el Centro Costo (130) MANTENIMIENTO
39	(02) SEGARRA BOLIVAR	Empleados que trabajan en el Centro Costo (030) MANTENIMIENTO y tienen una carga familiar

De todos estos nodos los más importantes son el nodo 76, 85, 87 y 89 que corresponde al Supervisor **(10) RODRIGUEZ DASTON APOLINARIO** con una proporción de ganancia del 100.00 por ciento de la muestra, tendiendo como diferencia el Centro Costo en el que laboran, seguidos del nodo 44 que corresponde al Supervisor **(02) SEGARRA BOLIVAR** con una proporción de ganancia del 69.86 por ciento de la muestra y del nodo 63 que corresponde al Supervisor **(03) MORA QUIJIJE RODOLFO SEGUNDO** con una proporción de ganancia del 69.32 por ciento de la muestra, según las Tablas 3.21, 3.22 y 3. 24. Los empleados que están que están en estos grupos son los que realizan más sobretiempo, a diferencia de los otros grupos.

CAPÍTULO IV

IV. CONCLUSIONES Y RECOMENDACIONES.

4.1 CONCLUSIONES

* A través de la aplicación de minería de datos se ha obtenido un modelo de conocimiento a partir de un volumen de datos que ha servido para ayudar al dueño o administrador de la empresa a tomar decisiones acerca de los empleados que más sobretiempo realizan.

* Los métodos de minería de datos llevan asociados una serie de mecanismos como son: la estimación de errores, matrices de confusión, análisis sensitivo de entradas, entre otros, los cuales nos permiten realizar una mejor validación del modelo que se aplica, de esta manera el análisis de resultados es más completo y fiable.

* El número de técnicas que engloba la minería de datos es amplio, las cuales se aplican dependiendo del análisis que deseamos realizar.

* WEKA es un programa sencillo de manejar que implementa numerosos algoritmos de aprendizaje y múltiples herramientas para transformar las bases de datos y realizar un exhaustivo análisis.

* Aplicando el algoritmo J4.8 basado en clasificación por árbol de decisión, el cual se trata de una implementación propia de WEKA para el algoritmo C4.5, se obtuvo los siguientes resultados:

1. Del árbol original con minNumObj (mínimo número de instancias con que debe contar una hoja) 2 por default, se obtuvo según Tabla 3.14 que:

   - ✓ La clasificación de los empleados que tienen como Supervisor 10 (RODRIGUEZ DASTON APOLINARIO) es la más completa, debido a la cantidad de hojas mostradas para su explicación, de los cuales el 82 por ciento de 367 registros semanales de empleados correspondientes a este grupo hacen sobretiempo de

clase 2, es decir realizan horas extras entre 12 y 23 horas a la semana.

- ✓ Los empleados que tienen como Supervisor 02 (SEGARRA BOLIVAR) corresponden al número mayoritario de empleados que hacen sobretiempo, de 1,360 registros semanales de empleados que pertenecen a este grupo el 100 por ciento de ellos hacen sobretiempo de clase 2, es decir, realizan horas extras entre 12 y 23 horas a la semana.

- ✓ Al grupo anterior le siguen los empleados que tienen como Supervisor 03 (MORA QUIJIJE RODOLFO SEGUNDO), donde el 100 por ciento de 1,306 registros semanales de empleados que pertenecen a este grupo hacen sobretiempo de clase 2, es decir, realizan horas extras entre 12 y 23 horas a la semana.

2. Del árbol con parámetro minNumObj (mínimo número de instancias con que debe contar una hoja)168, se obtuvo que:

- ✓ De todos los nodos obtenidos el más importante es el nodo 15 que corresponde al Supervisor (0486) PILLOAJO ORTEGA GENARO con una proporción de ganancia del 61.44 por ciento

de la muestra de registros semanales de empleados que hacen sobretiempo de la clase 3, es decir, realizan horas extras entre 12 y 23 horas por semana, según Tabla 3.16.

- ✓ Al nodo 15 le sigue el nodo 7 que corresponde al Supervisor (0008) ORTIZ JOSE con una proporción de ganancia del 56.73 por ciento de la muestra de registros semanales de empleados que hacen sobretiempo de la clase 2, es decir, realizan horas extras entre 3 y 12 horas por semana, según Tabla 3.15.

3. Del árbol con sobremuestreo, al cual se le modificó el parámetro minNumObj (mínimo número de instancias con que debe contar una hoja) a 400, se obtuvo:

- ✓ De todos los nodos obtenidos los más importantes son el nodo 76, 85, 87 y 89 que corresponde al Supervisor (10) RODRIGUEZ DASTON APOLINARIO con una proporción de ganancia del 100.00 por ciento de la muestra, tendiendo como diferencia el Centro Costo en el que laboran:

    - Los registros semanales de empleados que corresponden al nodo 76 no tienen horas pérdidas y que trabajan en el

Centro Costo (124) SEGURIDAD INDUSTRIAL, los cuales hacen semanalmente 3 o menos horas extras, según Tabla 3.21.

- Los registros semanales de empleados que corresponden al nodo 85 tienen horas pérdidas y que trabajan en el Centro Costo (122) HOJEADORA, hacen horas extras de la clase 2, es decir, entre 3 y 12 horas por semana, según Tabla 3.22.
- Los registros semanales de los empleados que corresponden al nodo 87 tienen horas pérdidas y que trabajan en el Centro Costo (125) CONTROL DE PAPEL, hacen horas extras entre 3 y 12 semanalmente, según Tabla 3.22.
- Los registros semanales de los empleados que corresponden al nodo 89 tienen horas pérdidas y que trabajan en el Centro Costo (132) FABRICA GENERAL, hacen horas extras de la clase 2, es decir, 3 a 12 horas semanalmente, según Tabla 3.22.

✓ A los nodos anteriores le sigue el nodo 44 que corresponde al Supervisor (02) SEGARRA BOLIVAR con una proporción de ganancia del 69.86 por ciento de la muestra. Este nodo corresponde a los registros semanales de empleados que

trabajan en el Centro Costo (040) EMPLEADO FCA. GENERAL, los cuales hacen mas de 23 horas extras semanalmente, según Tabla 3.24.

- ✓ Al nodo 44 le sigue el nodo 63 que corresponde al Supervisor (03) MORA QUIJIJE RODOLFO SEGUNDO con una proporción de ganancia del 69.32 por ciento de la muestra. En este nodo se encuentran los registros semanales de empleados que laboran en el Centro Costo (035) DESARROLLO y hacen 3 o menos horas extras semanalmente, según Tabla 3.21.

Con el árbol original y con el árbol de sobremuestreo podemos ver que la mayoría de los registros semanales de empleados hacen horas extras de la clase 2, es decir, 3 a 12 horas semanales, la mayoría de estos tienen como Supervisor (10) RODRIGUEZ DASTON APOLINARIO, seguido del Supervisor (02) SEGARRA BOLIVAR y del Supervisor (03) MORA QUIJIJE RODOLFO SEGUNDO.

4.2 RECOMENDACIONES

* Para la instalación del programa WEKA se requiere que la computadora tenga como mínimo 256 MB de memoria RAM.

* De los algoritmos que WEKA tiene implementado aplique el algoritmo dependiendo de la información que desea obtener, existen diferentes técnicas de minería de datos que se puede seleccionar.

* Si el árbol es muy amplio, se recomienda modificar el parámetro del mínimo número de instancias por hoja.

BIBLIOGRAFÍA

[1] **Congreso del Ecuador**. 2004. "*Código del Trabajo*", Guayaquil-Ecuador.

[2] **Juan Alvarado Ortega**. 2003. "*Algoritmos de Minería de Datos*", Vol. 1 Num. 2. Guayaquil-Ecuador.

[3] **Ian H. Witten and Eibe Frank**. 2000. "*Practical Machimne Learning Tools and Techniques with Java Implementations*", Morgan Kaufman Publichers. San Francisco-California.

[4] **Jean Pierre Lerry Mangin y Jesús Varela Mallou**. 2003. "*Análisis Multivariables par alas Ciencias Sociales*", Pearson. Madrid-España.

[5] **Verónica S. Bogado y Mariana C. Arruzazabala.** 2003. "*Descubrimiento de Conocimiento en Bases de Datos (KDD)*", Sistemas Operativos. Web http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonografiaMD.PDF