DISEÑO E IMPLEMENTACIÓN DE UN MODELO EXPLICATIVO DE LAS CAUSAS DE LAS HORAS DE SOBRETIEMPO QUE TIENE UNA EMPRESA QUE MANEJA PERSONAL DE SERVICIO

Daryna Marycruz Calderón Orozco[1], Ing. Juan Alvarado Ortega[2]

[1]Ingeniero en Estadística Informática 2005; (e-mail: dcaldero@solca.med.ec)
[2]Director de Tesis, Ingeniero en Computación FIEC, Escuela Superior Politécnica del Litoral, Magíster en Administración Publica, Escuela Superior Politécnica del Litoral, Profesor del ICM – ESPOL; (e-mail: jao_ec@yahoo.com)

Resumen. *El presente trabajo analizar las causas por las cuales empleados de una empresa de servicios realizan sobretiempo.*

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. Para llegar a estos patrones se utiliza el proceso de KDD (descubrimiento del conocimiento en bases de datos), a través del cual se creará un modelo para el análisis de una base de datos aplicando la técnica de árboles de decisión.

Este proceso permite 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 este trabajo el beneficio sería de negocio. Los resultados obtenidos, se reflejan en algunas conclusiones que se presentan en este trabajo.

Summary. ***T****he present work analyzes the causes for which employees of one enterprise of service to realize overtime.*

At the present time the organizations have a great quantity of warehouse and organize data. Treat of meet patterns, tendencies and anomalies is one of the great challenges of modern life. In order to extract these patterns we use the process of KDD (knowledge discovery in databases), through which we create a model for the analysis of a database applying the technique of decision trees.

This process permit to identify or discover patterns of overtime and tendencies little obvious intro of the data. The discovery of this information serve to carry something out actions and to obtain a benefit, be this scientist or of business, given the case of this work the benefit would be of business. The obtained results, are reflected in some conclusions that appear in this work.

---

1. INTRODUCCIÓN

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. [2]

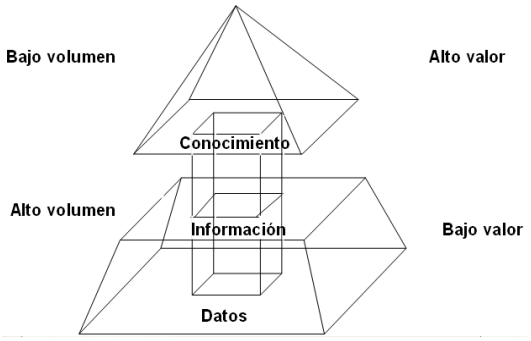

Figura 1: Jerarquía que existe en una base de datos entre datos, información y conocimiento.

1.1 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".

1.2 El Proceso KDD

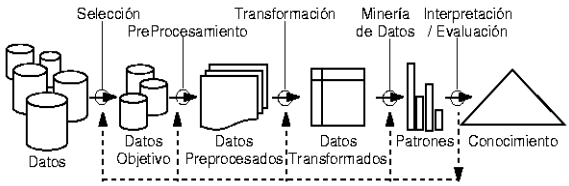

Figura 2: Proceso de extracción de conocimiento.

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.

El proceso de descubrimiento de conocimiento en bases de datos involucra varias fases [4]:

1. Determinar las fuentes de información
2. Diseñar el esquema de un almacén de datos (Data Warehouse
3. Implantación del almacén de datos
4. Selección, limpieza y transformación de los datos que se van a analizar
5. Seleccionar y aplicar el método de minería de datos apropiado
6. Evaluación, interpretación, transformación y representación de los patrones extraídos
7. Difusión y uso del nuevo conocimiento.

Para este trabajo se usó una base de datos de una empresa de servicios cuyo nombre se omite por razones de confidencialidad. Esta base contiene los registros de las diferentes marcaciones que cada uno de los empleados ha realizado en sus jornadas laborales, 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 aplicar el proceso de KDD, para determinar patrones por los cuales los empleados hacen sobretiempo.

El sobretiempo representa para el jefe o administrador de la empresa un gasto imprevisto, y al analizar sus causas le permitirá tomar las decisiones que el considere convenientes enfocándose en un grupo determinado.

2. APLICACIÓN DEL PROCESO KDD

2.1 Determinar las fuentes de información

La fuente de información proviene de una base de datos que esta en Microsoft Visual FoxPro.

Los registros de las diferentes marcaciones están almacenados en una tabla llamada *reloj*, siendo esta la tabla de mayor interés para el análisis de las causas de sobretiempo.

2.2 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.

Para el desarrollo de este trabajo, se uso un data mart, el cual representa una pequeña parte de la información de la empresa.

2.3 Implantación del almacén de datos

En la actualidad el almacén de datos de la empresa de servicios está implementado.

De la base de datos, la información que me ayudará para el objetivo de este trabajo, 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.*

2.4 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.

Se hizo un query para filtrar: la fecha de nacimiento, los días no laborables o días festivos, el estado civil diferente de NN que no describe el estado civil real del empleado.

También se realizo querys para: obtener los diferentes turnos, determinar las horas extras y los días que labora el empleado en la semana, asignar las clases de sobretiempo.

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

Con los datos filtrados se hizo un query con el atributo que registra los sobretiempos de los empleados y 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.

Los datos obtenidos fueron los siguientes:

Tabla I
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	

Elaborado por: Daryna Calderón Orozco.

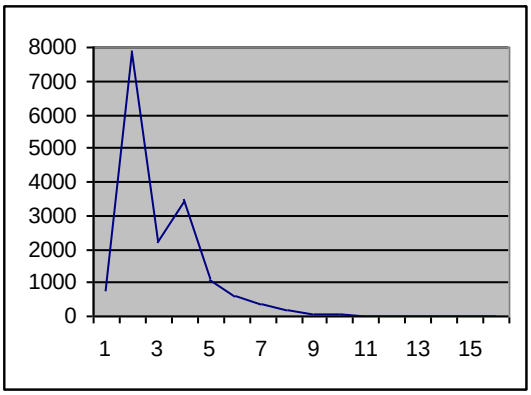

Figura 3: Gráfico de frecuencias de sobretiempo.

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.

Finalmente, la tabla obtenida de las filtraciones (tabla con la que se trabaja) queda con 16,824 registros y 20 atributos.

Se transforma los datos de esta tabla con formato DBF a ARFF, el cual se lo utiliza en un programa llamado WEKA. Para transformar los datos a formato ARFF se utiliza un query.

2.5 Seleccionar y aplicar el método de minería de datos apropiado

La técnica que mejor se adapta al objetivo de este trabajo son los *árboles de decisión.* Esta técnica es de fácil entendimiento y es apropiada para predicciones numéricas.

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 aplica esta técnica mediante el uso de WEKA, el cual explicaré brevemente.

2.5.1 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.

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.

2.5.2 Aplicación del programa WEKA

Cargamos el archivo con formato *ARFF*, en el que disponemos de 16,824 instancias o registros de 20 atributos.

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.

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).

2.5.2.1 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.

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.*

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

El tipo de predicciones correctas e incorrectas cuando se aplico el modelo sobre el conjunto de prueba son mostradas en una Matriz de Confusión, donde 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.

El árbol resultante se clasificó o ramificó primero por el atributo *Supervisor*, en esta parte no se muestra el árbol obtenido por ser muy extenso; pero si se lo detallará para que se tenga una idea clara del mismo.

2.6.1 Resumen de Datos Obtenidos

Del árbol obtenido se contó el número de hojas en las que se clasificó el atributo *Supervisor* hasta obtener el resultado final, mientras más alta es la cantidad, la complejidad para la explicación del atributo es mayor, caso contrario es menor. A su vez, se sumo el número muestreado de los registros semanales de los empleados dándonos 5,946 registros.

De los datos obtenidos se determinó que 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.

2.6.2 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 default 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.

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

```
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
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
```

Figura 4: Árbol de Decisión obtenido de la modificación del parámetro minNumObj del Algoritmo J4.8

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.

2.6.2.1 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 II, III y IV muestran el cuadro de las ganancias por cada nodo final del Árbol de Decisión mostrado en la Figura 4, 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 [3]:

- ✓ *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.*

De los Cuadros de Ganancia obtenidos, se mostrará una parte del CUADRO DE GANANCIAS POR NODO DE LA CLASE 2, debido a que es extenso para mostrarlo completamente. El resto de cuadros se muestran en su totalidad.

Tabla II
CUADRO DE GANANCIAS POR NODO
DE LA CLASE 2

Nodo por Nodo						
Nodo	**Nodo n**	**Nodo %**	**Resp n**	**Resp %**	**Ganancia (%)**	**Index (%)**
7	275	1.76	156	2.68	**56.73**	152.61
16	288	1.84	151	2.60	52.43	141.05
21	303	1.94	152	2.61	50.17	134.95
…						
24	30	0.19	3	0.05	10.00	**26.90**
Acumulado						
Nodo	**Nodo n**	**Nodo %**	**Resp n**	**Resp %**	**Ganancia (%)**	**Index (%)**
7	275	1.76	156	2.68	56.73	152.61
16	563	3.60	307	5.28	54.53	146.69
21	866	5.54	459	7.89	53.00	142.59
…						
24	15,638	100.00	5,813	100.00	37.17	100.00

Tabla III
CUADRO DE GANANCIAS POR NODO
DE LA CLASE 3

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

Tabla IV
CUADRO DE GANANCIAS POR NODO
DE LA CLASE 4

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

Elaborado por: Daryna Calderón Orozco.

2.6.2.2 Análisis del Cuadro de Ganancias por nodo de cada categoría

De la Tabla III 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 la muestra de registros semanales de los empleados que hacen sobretiempo de la clase 3.

De la Tabla IV 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 la muestra de registros semanales de los empleados que hacen sobretiempo de la clase 4.

Análisis similar se hizo para el caso de la Tabla II.

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

De las Tabla II, III y IV destacamos los nodos más importantes y el *codsuper* (Supervisor) que se asignó a ese nodo, según el Árbol de Decisión mostrado en la Figura 4.

Tabla V
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	
21	(1750) MENDOZA PAZ SOCRATES CONTARDO	Empleados que no tienen cargas familiares

Tabla VI
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 VII
NODOS A DESTACAR SEGÚN CUADRO DE GANANCIAS DE LA CLASE 4

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

Elaborado por: Daryna Calderón Orozco.

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 V y VI respectivamente.

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

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:

- ✓ 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.

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 VI.

- ✓ 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 V.

REFERENCIAS BIBLIOGRÁFICAS

1. **Daryna M. Calderón**, "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, Instituto de Ciencias Matemáticas, Escuela Superior Politécnica del Litoral, 2005).

2. **Verónica S. Bogado y Mariana C. Arruzazabala,** septiembre 2003. Descubrimiento de Conocimiento en Bases de Datos (KDD), http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonografiaMD.PDF

3. **Jean Pierre Lerry Mangin y Jesús Varela Mallou**, Análisis Multivariables par alas Ciencias Sociales (Madrid, Pearson, 2003), Capítulo 13.

4. **Juan Alvarado Ortega**, "Algoritmos de Minería de Datos", Revista Tecnológica, Vol. 1, No. 2 (Guayaquil, 2003).

5. **Ian H. Witten and Eibe Frank**, Practical Machine Learning Tools and Techniques with Java Implementations (San Francisco, Morgan Kaufman Publichers, 2000).