ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística Informática

"Estimación del nivel de deterioro en la madera de las haciendas de una compañía balsera en el Ecuador, mediante el diseño de Data Marts y modelamiento de Árboles de Decisiones"

TESINA DE GRADO

Previa a la obtención del Título de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

Presentado por:
Elizabeth Ponce Santana
Wendy Rodríguez Santiana

GUAYAQUIL – ECUADOR
2010

Tribunal de Graduación

Mat. John Ramírez Figueroa

Delegado

Msc. Pedro Fabricio Echeverría Briones

Director de Tesina

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este Proyecto de Graduación, nos corresponde exclusivamente; y el patrimonio intelectual del mismo a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL".

_____

Elizabeth Ponce Santana

_____

Wendy Rodríguez Santiana

DEDICATORIA

Elizabeth…

A mi madre esta dedicatoria como un homenaje a su grandeza, que con sabios consejos orientó mis pasos por el camino recto de la vida. Michi ten presente que la gloria más grande que tengo es ser tu hija…

Wendy…

Dedicado a una mujer que admiro por su sabiduría, fortaleza y virtudes que posee… Mi madre.

En agradecimiento por su amor incondicional, sus enseñanzas, por compartir mis logros y disfrutarlos tanto como yo.

Te amo mamá.

AGRADECIMIENTO

Elizabeth…

A mi familia que me apoyó anímica, moral, material y económicamente durante todos estos años. En especial a mis primos que siempre me han cuidado y protegido como una hermana.

A mi Abuelita por su fortaleza y valentía para enfrentar la vida hasta sus últimos días. A mi hermana Adriana: por ser y estar, por compartir el espacio y los momentos significativos.

Wendy…

Mis agradecimientos: A Dios, por acompañarme en cada instante y permitirme disfrutar de este momento. A ti mamá, por tu amor y principalmente por tu apoyo en cada reto que me toca vivir. A mi abuelita Fanny, por consentirme y brindarme sus cuidados que me transmitieron su fortaleza. A mi papá por su amor, comprensión y respeto de mis decisiones. A ti Xavier, por ser un hermano genial, por tu apoyo en momentos difíciles y porque sé que me quieres tanto como yo a ti. A Orlando por estar presente y compartir momentos importantes, por sus consejos y amor sincero. En definitiva ¡Gracias!.... A mi familia y a todas las personas especiales que he conocido, por su aportación en mi vida; unas junto a mí, otras en mi memoria, pero todas en mi corazón.

RESUMEN

En este estudio se analizaron los procedimientos utilizados en una empresa del Ecuador exportadora de madera de balsa. Los cuales van desde plantaciones (inicio del proceso), hasta la elaboración en las fábricas del producto final (bloques formados por piezas de balsa).

Luego de mejorar algunos de los procedimientos en la empresa, se estimó el nivel de deterioro de la madera en las haciendas ó plantaciones; con el objetivo de reducirlo para evitar pérdidas futuras.

Se utilizó para este fin la metodología de Data Marts y Árboles de decisión.

ÍNDICE GENERAL

ABREVIATURAS

BFT	Pie Maderero (Board Feet)
CHAID	Chi-Squared Automatic Interaction Detection

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE GRÁFICOS

Pág.

ÍNDICE DE TABLAS

INTRODUCCIÓN

En esta investigación se estimará el nivel de deterioro de las haciendas en una empresa balsera en el Ecuador; la cual se dedica a la exportación de madera de balsa procesándola desde las plantaciones hasta elaborar el producto final.

La empresa se vio afectada en el año 2009 por problemas con la calidad del producto, los cuales de acuerdo a los estudios realizados se originan hipotéticamente desde plantaciones ó haciendas.

A fin de confirmar esta hipótesis se utilizarán técnicas de Minerías de Datos como Árboles de decisión, partiendo de la elaboración de un Modelo Estrella (Data Marts) específico para el Dpto. de Satélites, área del negocio inmerso en este estudio; con el objetivo de reducir el deterioro encontrado en la madera de balsa, y que éste no exista en el producto final.

CAPÍTULO I

1. Planteamiento del Problema

Una empresa ecuatoriana exportadora de balsa requiere analizar los procesos inmersos en el aserrado de la madera en sus haciendas, con el objetivo de reducir el nivel de deterioro.

1.1 Descripción del proceso [1]

La empresa objeto de este estudio se dedica al tratamiento de madera de balsa, desde plantaciones (haciendas) hasta obtener el producto final; diferentes presentaciones para exportación según requerimiento del comprador.

Posee plantaciones (haciendas) que se encargan del proceso de sembrar los árboles de madera de balsa. Cuando están listos para la cosecha pasan a ser cortados, proceso que es denominado ***"tumba"*** del árbol. A cada árbol tumbado se le retiran las hojas quedando únicamente el tronco, el cual es llamado ***"troza"***. Cada troza es pintada en sus extremos de un color de acuerdo al día de la semana.

Figura 1.1
Hacienda de madera de balsa

Figura 1.2
Figura 1.2
Hacienda de madera de balsa con parte cosechada

Posteriormente cada árbol tumbado, pasa a la siguiente fase (proceso de ***aserrado***). Dicho proceso consiste en cortar el árbol en pequeñas piezas que pueden tener tres medidas: 4, 5 ó 6 pies de largo; así mismo, pueden ser de diferente espesor o ancho.

Figura 1.3
Piezas de balsa, pueden tener diferentes espesor y ancho

[1] Tomado de: Manual de procedimientos de la empresa.

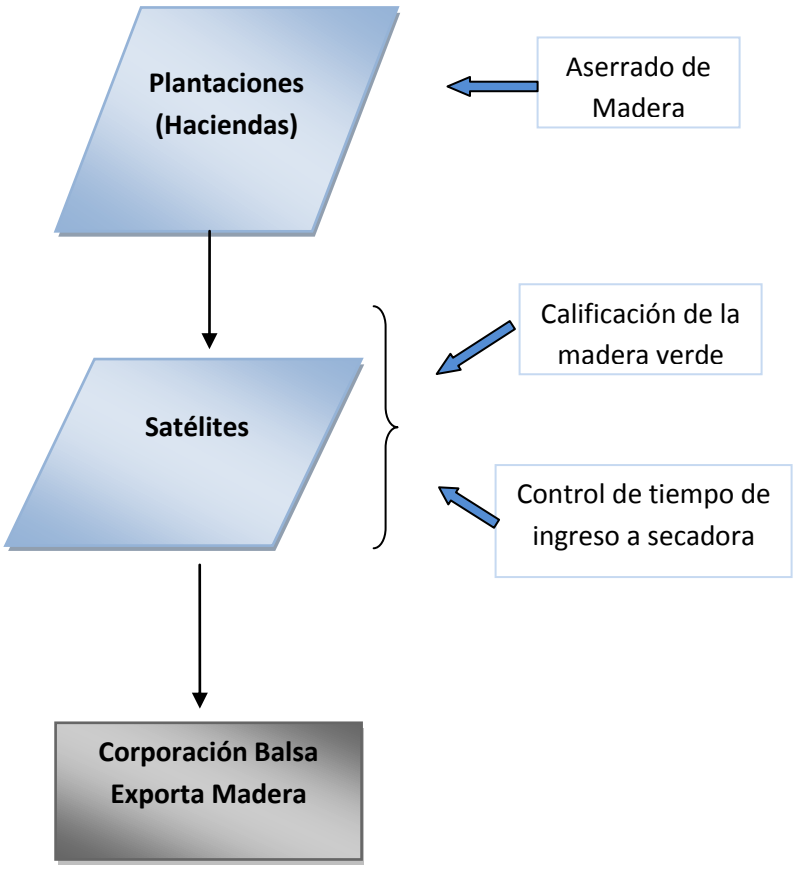
Un conjunto de 70 piezas aproximadamente es reunido para formar un ***"bulto"***. A cada bulto se lo sella con la fecha en que fue realizado este proceso.

Figura 1.4
Armado de bultos

Estos bultos son distribuidos a las fábricas de madera denominadas ***"satélites",*** que prestan servicios de procesamiento de madera de balsa para la empresa en estudio.

1.1.1 Flujo del Proceso[2]

Gráfico 1.1
Flujo del proceso



[2] Tomado de: Manual de procedimiento de la empresa.

La distribución de los bultos es por camiones que transportan 80 bultos aproximadamente. Los bultos son recibidos por las satélites, se registra la fecha de recepción y se realiza un muestreo de cada camionada que llega, para determinar y registrar los defectos encontrados en los bultos muestreados. La muestra es el 5% del total de bultos en la camionada.

Figura 1.5
Muestra como se realiza la distribución en camiones

Cada satélite recibe la madera verde y la debe ingresar a las máquinas de secado. Dependiendo de la capacidad y la cantidad de secadoras que tengan las satélites deben organizar el ingreso de la balsa sin exceder de un lapso mayor a 21 días. Este tiempo ha sido hasta la actualidad ***determinado empíricamente***, y transcurre desde la fecha de tumba del árbol hasta el momento que la madera ingresa a las secadoras. Si se excede este lapso la madera puede deteriorarse; lo que ocasionaría grandes pérdidas a la empresa.

Para prever que las satélites tengan mayor cantidad de madera verde a la que pueden secar, la empresa está realizando actualmente un control semanal de distribución para secado de la madera verde.

Tabla I
Previsión de secado

Previsión de Secado							Fecha de Elaboración:
SECADORAS							
Días	1	2	3	4	5	6	7
	Capacidad	Capacidad	Capacidad	Capacidad	Capacidad	Capacidad	Capacidad
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							

Fuente: Satélites
Elaborado por: Jefe planta

La Tabla I muestra el formato utilizado para coordinar los días que tarda cada satélite en secar la madera de acuerdo a la cantidad y capacidad con que cuenta cada secadora. Esta información es enviada semanalmente y de acuerdo a ella se distribuye las camionadas de madera.

Una vez que es secada la madera verde pasa a los talleres de las satélites o de la empresa, para ser procesada y convertida en el producto final - paneles de balsa que son exportados y se utilizan en los pisos de los barcos y otros.

1.2 Antecedentes

La empresa está atravesando serios problemas con la calidad del producto; en Septiembre del 2009 iniciaron reclamos por parte de sus clientes, por esta razón ha tomado acciones correctivas y preventivas para evitar reclamos posteriores.

Realizando un análisis se determinó el problema ha tratar en este proyecto. Se sabe que uno de los factores que producen el problema, son las diferentes formas que es aserrada la madera en cada una de las haciendas; otro factor es la forma como se califica y procesa la madera en las satélites, por ejemplo hay defectos que son calificados como madera correcta, ocasionando que dicha madera inconforme sea aceptada en los siguientes procesos llegando hasta el producto listo de exportación. Finalmente, un factor que influye también es el tiempo en exceso que tiene que esperar la madera verde en los patios de las satélites, produciendo así uno de los mayores defectos que causa pérdidas a la empresa es el deterioro.

A finales del año 2008, la empresa en mención tomó la decisión de cerrar una de las fábricas que les secaba madera, siendo ésta la que mayor capacidad de secado tenía. Sin preveer el invierno y las consecuencias futuras, esta decisión ocasionó que haya un sobre-stock en las otras satélites y que gran cantidad de madera verde no ingresara a tiempo a las secadoras. Sin embargo, aún cuando el tiempo había sido excedido, se realizó el proceso de secado dicha madera, fue procesada y calificada, presentando muchos rechazos por deterioro; además hubo madera verde que tenía en su interior podredumbre que no fue detectaba y

continuó en el proceso hasta el producto final y fue exportada; los clientes al revisar la calidad del producto detectaron el problema, devolvieron los contenedores y fue un impacto económico grande para la empresa balsera. Actualmente, con la reapertura de la fábrica se espera disminuir el grado del problema.

1.3 Definición del Problema

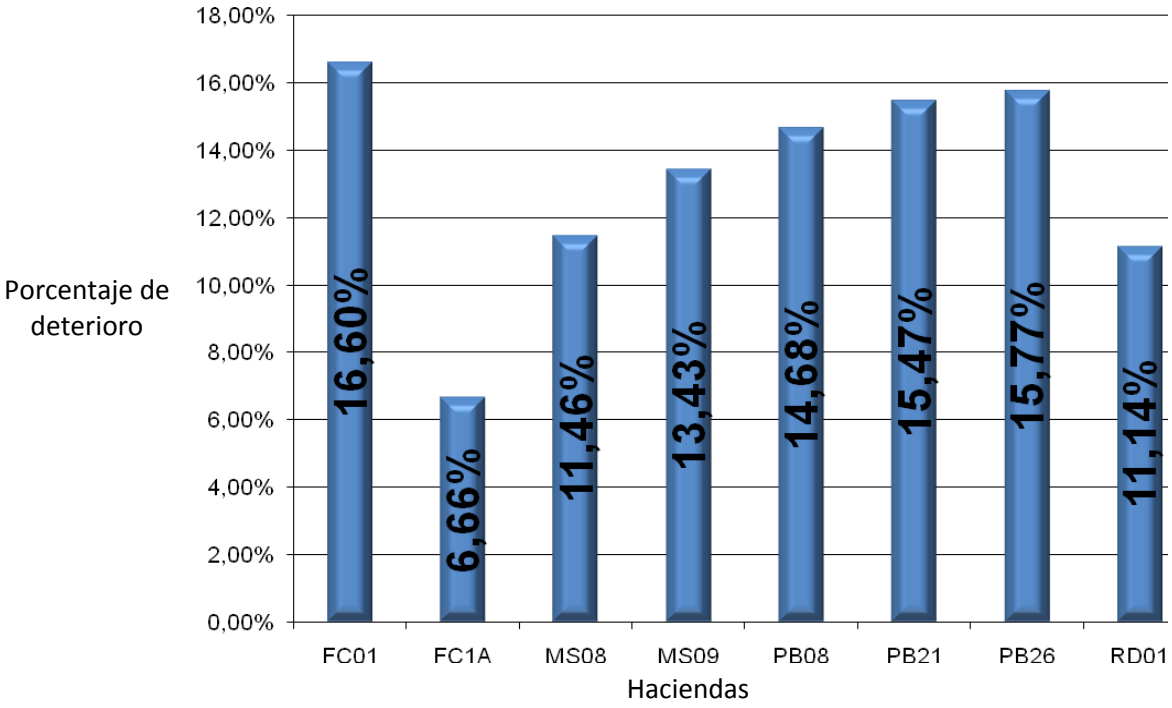
Necesidad de mejorar los procesos de control establecidos, desde las haciendas hasta que la madera sea ingresada a secadoras. Estimar el nivel de deterioro de la madera a fin de reducirlo para que la madera cumpla con los estándares establecidos por la empresa exportadora.

Tabla II
Nombres de las Haciendas de la empresa

Códigos	Nombre de la Hacienda
FC01	Finca 01
FC1A	Finca 1A
MS08	Maseca 08
MS09	Maseca 09
PB08	Plantabal 08
PB21	Plantabal 21
PB26	Plantabal 26
RD01	Redonda 01

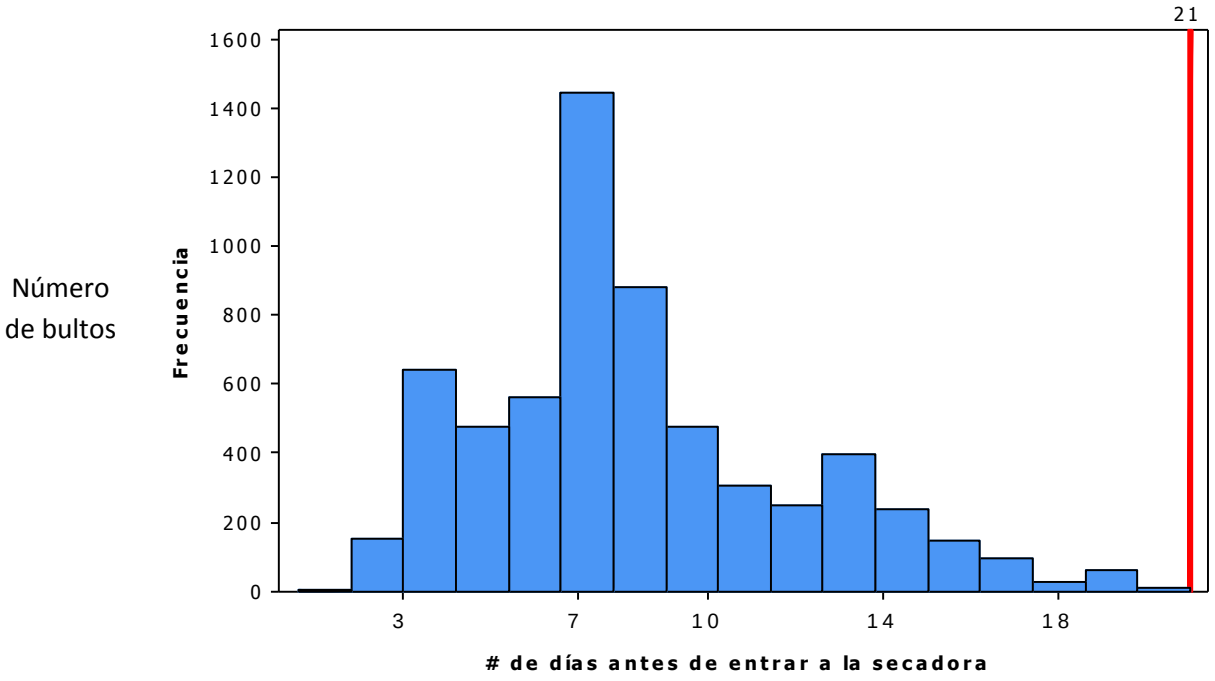
Fuente: Empresa

Gráfico 1.2
Nivel de Deterioro



En el Gráfico 1.2 se observa el nivel o porcentaje de deterioro de la madera de acuerdo a la hacienda donde procede. Como se muestra en este mismo gráfico, las haciendas son conocidas y están representadas por códigos. Sin embargo para esta investigación, se muestra en la Tabla II, los nombres exactos de las haciendas para conocer el significado de los códigos mostradas en el Gráfico 1.2.

Gráfico 1.3
Histograma de días en espera de secado



Revisado el nivel de deterioro, se puede observar en el Gráfico 1.3 el número de días que las satélites tardan en ingresar la madera de las haciendas a las secadoras. Cabe mencionar que la madera está organizada en bultos, por lo cual el Gráfico 1.3 muestra el número de días que determinada cantidad de bultos espera hasta ingresar al proceso de secado. Si el lapso que espera la madera en las satélites, sobrepasa los días establecidos por los estándares de la empresa, puede ocurrir que esto afecte la calidad de la balsa y se produzcan problemas en el producto final.

1.4 Propuesta de Soluciones

Una vez definido el problema se considera que las soluciones más viables para la empresa son:

- Aplicación de una de las técnicas de Minería de Datos como son los Árboles de decisiones, a fin de obtener información confiable para las tomas de decisiones con respecto a los niveles aceptables de defectos encontrados en la madera.

- Realización de un Data Marts para el Departamento que controla las satélites, con el objetivo que puedan tener acceso a esta información las diversas partes involucradas: haciendas, satélites, Dpto. de Control y otros.

1.5 Objetivo General

Estimar el nivel de deterioro promedio controlando que la madera que llega de las plantaciones ha sido procesada bajo los estándares establecidos por la empresa exportadora.

1.6 Objetivos Específicos

- Realización de Data Mart para las satélites.
- Utilizando la información contenida en el Data Mart, aplicar modelamiento de árbol de decisión para la toma de decisiones y mejora de resultados.

CAPÍTULO II

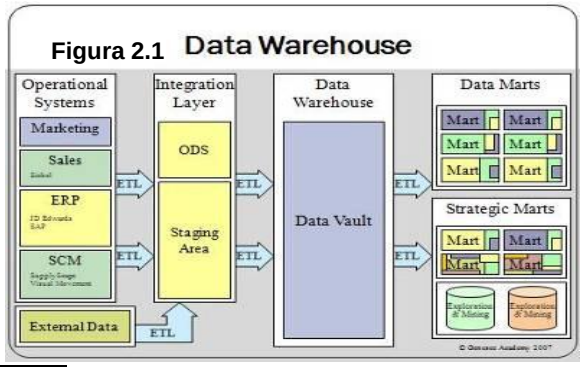
2. Definiciones técnicas

A partir de la aplicación de una técnica de Minería de Datos como los árboles de decisión se pretende estimar el nivel de deterioro en madera de balsa, basado en las relaciones que se establecen entre el centro de procedencia de la madera, llegada a las satélites, y tiempo de espera entre tumba y secado. Estos resultados pueden mejorar el proceso de envío, recepción y elevar la calidad de la madera en la empresa en estudio.

2.1 Data Warehouse [3]

Un Data Warehouse es una solución que permite centralizar en un solo punto, toda la información definida por la compañía como relevante para la gestión de su negocio y la toma de decisiones. La distribución de la información se realiza a través de herramientas que permiten a los usuarios finales construir sus propios informes de forma autónoma.

Figura 2.1 Data Warehouse



[3] Tomado de: http://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos
http://empresas.telefonica.es/asp/catalogo_servicios/aplicaciones_negocio/aplicaciones_soporte/serv_gestion/data.htm

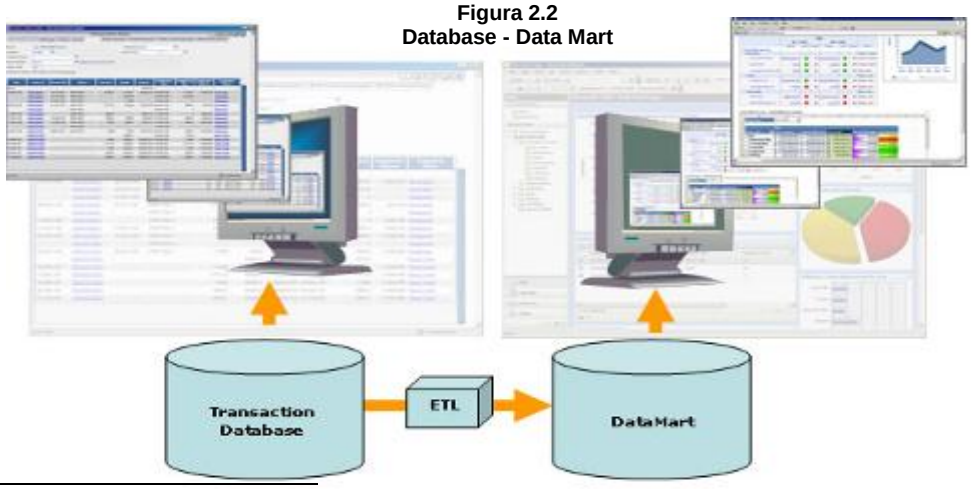
2.2 DataMart [4]

Un DataMart es una solución, que compartiendo tecnología con el Data Warehouse (pero con contenidos específicos, volumen de datos más limitado y un alcance histórico menor), permite dar soporte a una empresa pequeña, o un departamento o área de negocio de una empresa grande.

Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro de la compañía para que pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas para que diversos grupos de usuarios realicen la explotación de los mismos de la forma más conveniente según sus necesidades.

En síntesis, se puede decir que los **Data Marts** son pequeños **Data Warehouse** centrados en un tema o área dentro de una organización.

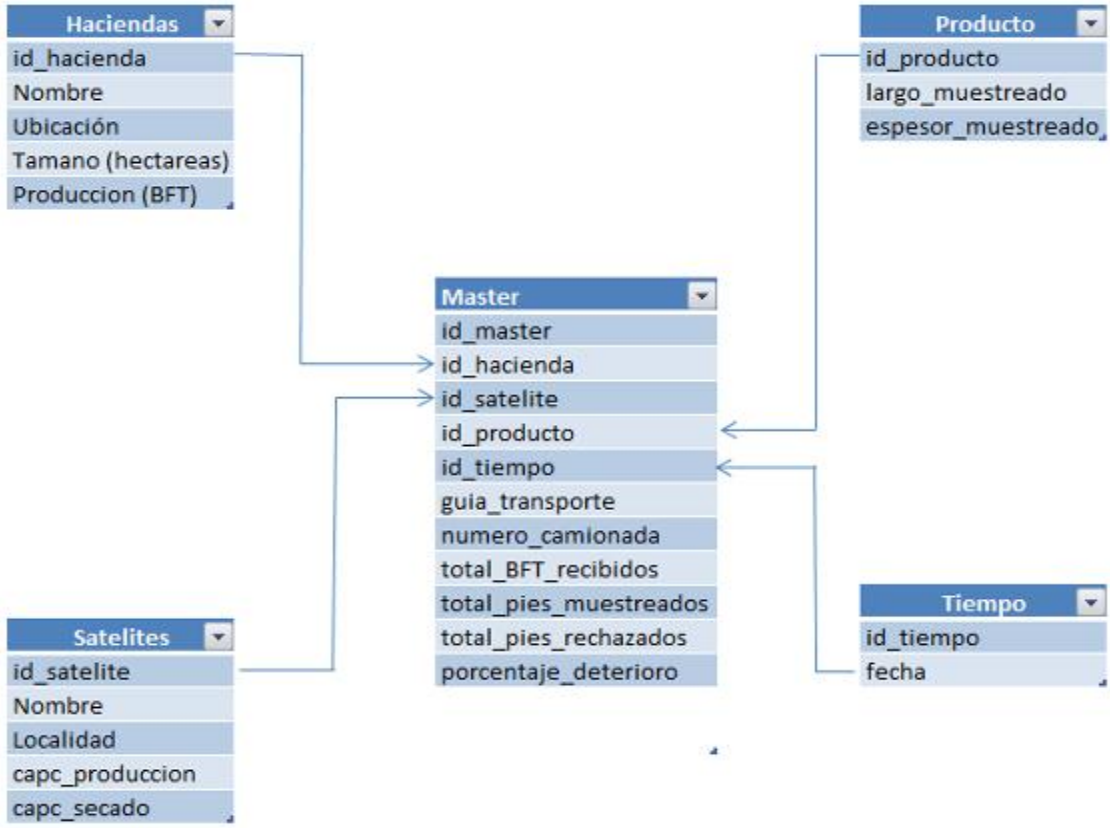
Figura 2.2
Database - Data Mart



[4] Tomado de: http://empresas.telefonica.es/asp/catalogo_servicios/aplicaciones_negocio/aplicaciones_soporte/serv_gestion/data.htm

2.3 Modelo Estrella

Figura 2.3
Data Mart (Modelo Estrella)



Este es el Data Mart de donde se obtendrá la información para aplicar la técnica de minería de datos, Árboles de decisión, para la estimación del deterioro en las haciendas de la empresa balsera.

2.4 Descripción de las Tablas

Lo siguiente es el diccionario de datos del modelo estrella (ver Sección 2.3).

Tabla haciendas, esta tabla contiene los datos de cada hacienda destacando como campos importantes:

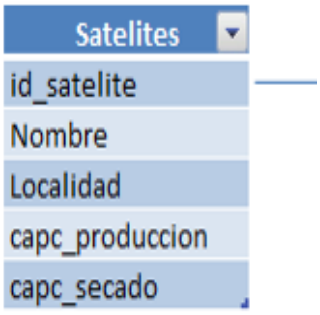
Nombre, es la identificación que se le ha dado a la hacienda hace algunos años.

Ubicación, es donde se encuentra localizada geográficamente.

Tamaño, nos indica el número de hectáreas que tiene la hacienda.

Producción, indica la producción estimada de la hacienda en BFT que es la medida en volumen de la madera de balsa.

Tabla Satélites, esta tabla contiene los datos de cada hacienda destacando como campos importantes:



Nombre, es la denominación de la satélite asignada por los dueños de las mismas.

Localidad, es la ubicación geográfica.

Capc_producción, capacidad mensual de producción.

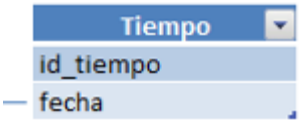
Capc_secado, capacidad de secado de madera del total de sus secadoras.

Tabla Producto, esta tabla contiene los datos del producto que llega a las satélites y del cual se toma las muestras para establecer el nivel de deterioro.

Largo_muestreado, es el largo de la pieza de balsa que pertenece a la muestra que se obtiene y que se revisa para determinar si tiene o no deterioro y estimar el total de defectos en toda la camionada.

Espesor_muestreado, es el espesor de la pieza de balsa que se toma como muestra de la camionada y se revisa el nivel de deterioro.

Tabla Tiempo, esta tabla contiene los datos referentes al tiempo que sean necesarios.



Fecha, campo que determina el día, mes y año en que son registrados los datos.

CAPÍTULO III

3. Árbol de decisiones [5]

Se define un árbol de decisión como una estructura en forma de árbol en la que las ramas representan conjuntos de decisiones. Estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos disjuntos y exhaustivos. Las ramificaciones se realizan de forma recursiva hasta que se cumplen ciertos criterios de parada.

El objetivo de estos métodos es obtener individuos más homogéneos con respecto a la variable que se desea discriminar dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas a partir de las cuales se va a realizar la discriminación de la población en subgrupos.

El programa AID (Automatic Interaction Detection) de Sonquist, Baker y Morgan (1.971), representa uno de los primeros métodos de ajuste de los datos basados en modelos de árboles de clasificación. AID esta basado en un algoritmo recursivo con sucesivas particiones de los datos originales en otros subgrupos menores y más homogéneos mediante secuencias binarias de particiones. Posteriormente surgió un sistema recursivo binario similar denominado CART (Classification And Regression Tree, Árboles de Clasificación y Regresión) desarrollado por Breiman en 1.984. Un algoritmo recursivo de clasificación no binario, denominado CHAID (Chi Square Automatic Interaction Detection, Detección de Interacción Automática de Chi Cuadrado) fue desarrollado por

[5] Tomado de: http://www.eustat.es/document/datos/ct_04_c.pdf

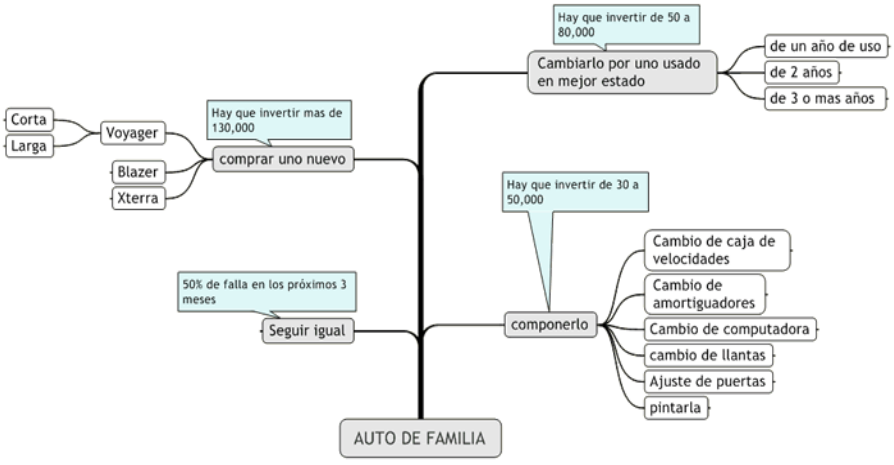
Kass en 1.980. Dentro de los métodos basados en árboles se pueden distinguir dos tipos dependiendo de tipo de variable a discriminar:

- · Árboles de clasificación. Este tipo de árboles se emplea para variables categóricas, tanto nominales como ordinales.
- · Árboles de regresión. Este tipo de discriminación se aplica a variables continuas.

Teniendo en cuenta el tipo de variable con que estamos trabajando se calcula distintas medidas para el estudio de la homogeneidad. En todos los casos las variables explicativas son tratadas como variables categóricas. En particular en el caso de tener una variable explicativa continua, salvo que haya sido categorizada previamente, será tratada como una variable categórica con el número de clases igual al número de valores distintos de la variable en el fichero de datos. Dependiendo de la estructura del árbol, del número de ramas que se permiten generar a partir de un nodo, se distinguen dos tipos:

- Árboles basados en la metodología CART: Técnica de árbol de decisión que permite generar únicamente dos ramas a partir de un nodo.
- Árboles basados en la metodología CHAID: genera distinto número de ramas a partir de un nodo.

Figura 3.1
Ejemplo de Árboles de decisiones



3.1 Software utilizado para la elaboración del árbol de decisiones [6]

SPSS AnswerTree es una herramienta para la creación de sistemas de clasificación en forma de árbol de decisión, lo cual nos ofrece una fácil interpretación de los resultados. Permite crear perfiles, describir segmentos, observar tendencias ocultas y segmentar bases de datos, sus aplicaciones son muy variadas: mejora de los resultados de campañas de marketing directo, investigación médica, detección de operaciones fraudulentas, análisis de venta cruzada, identificación de clientes potenciales, etc.

Nos ofrece la posibilidad de usar cuatro potentes algoritmos de segmentación. Los resultados son sencillos de interpretar y comprender en la estructura de árbol que posee.

CHAID - es un rápido algoritmo de árbol estadístico de múltiples variables que explora eficazmente los datos.

CHAID exhaustivo - es un completo algoritmo de árbol estadístico de múltiples variables que permite una exploración exhaustiva de los datos.

C&RT - Árbol de clasificación y regresión - es un completo algoritmo de árbol binario para dividirlos datos y generar subconjuntos homogéneos precisos.

QUEST - es un algoritmo estadístico para seleccionar variables sin sesgos y crear rápida y eficientemente un preciso modelo de árbol binario.

---

[6] Tomado de: http://www.telematica.com.pe/Product/spss_AnsTree.htm
http://www.spss.com/la/soluciones/data-mining2.htm

3.2 Conocimientos aplicados en Árboles de decisiones [7]

3.2.1. Prueba de Bondad de Ajuste

Con mucha frecuencia no se conoce la distribución de probabilidad de la variable aleatoria en estudio, digamos X, y se desea probar la hipótesis de que X sigue una distribución de probabilidad particular.

Existen dos procedimientos para realizar pruebas de bondad de ajuste que son los más conocidos. El primero se basa en una técnica gráfica muy útil llamada **gráfica de probabilidad** y el segundo procedimiento se basa en la **distribución Chi-cuadrada.** En este estudio se aplicará la distribución Chi-cuadrada la que describimos a continuación.

3.2.2. Prueba de bondad de ajuste de la Chi-cuadrada

El procedimiento de prueba de la Chi-cuadrada es un método analítico, requiere una muestra aleatoria de tamaño *n* de la variable aleatoria *X.* Estas *n* observaciones se arreglan en histogramas de frecuencias, teniendo *k* intervalos de clase (donde $k = \sqrt{n}$). Sea n_i la frecuencia observada en el *i*-ésimo intervalo de clase. De la distribución de probabilidad hipotética, se calcula la frecuencia esperada en el *i*-ésimo intervalo de clase, identificada como p_i.

H_0 : La distribución es $F_0(X)$ (Distribución Teórica dada)

H_a : La distribución no puede ser $F_0(X)$

EP: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$

RR: $\chi^2 > \chi^2_{\alpha, k-r-1} gl$

[7] Tomado de: http://ininweb.uprm.edu/cc/PRUEBA%20DE%20BONDAD%20Y%20AJUSTE%20.doc

3.2.3. Grados de Libertad [8] es un estimador del número de categorías independientes en una prueba particular o experimento estadístico. Se encuentran mediante la fórmula $k - r$, donde k=número de grupos, cuando se realizan operaciones con grupos y r es el número de sujetos o grupos estadísticamente dependientes.

3.3 Aplicación de Árboles de Decisiones para estimar el nivel de deterioro de las haciendas.

Para la realización de los árboles de decisión, en este estudio se han utilizado tres de los algoritmos que presenta el software, las variables utilizadas para este estudio son las siguientes:

Variable dependiente: “id_hacienda”

Variables predictoras o explicativas:

- “total_de_pies_rechazados”
- “total_de_pies_muestreados”
- “id_satelites”

[8] Tomado de: http://es.wikipedia.org/wiki/Grados_de_libertad_(estadÃ-stica)

Algoritmo C&RT [9]

El algoritmo de Árboles de Clasificación y Regresión, más conocido por sus siglas en inglés C&RT (Classification And Regression Tree), es un algoritmo de regresión y clasificación jerárquica que constituye un método parecido al algoritmo AID que dicotomiza la muestra según el modelo de Breiman y genera árboles de decisiones binarios. El propósito es realizar una partición de los datos en subconjuntos que sean homogéneos con respecto de la variable dependiente, a cada partición de los datos se selecciona la mejor variable independiente predictora (explicativa) que presentará la mejor dicotomización (para variables independientes continuas) o el mejor ajuste de categorías (para variables nominales u ordinales) en base a la mayor reducción de impurezas, es decir se crean dos nodos con el mejor predictor. El proceso se repetirá hasta que se alcancen los parámetros de paro.

En la Figura 3.2 se puede observar que ha existido una mejora del 9.63% para la variable predictora Total de pies muestreados, de donde se puede verificar que la balsa que se ha muestreado en las satélites proviene en un 47.06% de la hacienda PB08, cuando el total de pies muestreados es inferior o igual a 250. Sin embargo si los pies muestreados son mayores a 250, la plantación donde proviene la mayor cantidad de pies que intervienen en la muestra es la hacienda denominada MS09 con un 74.29%, como se observa en el nodo 2.

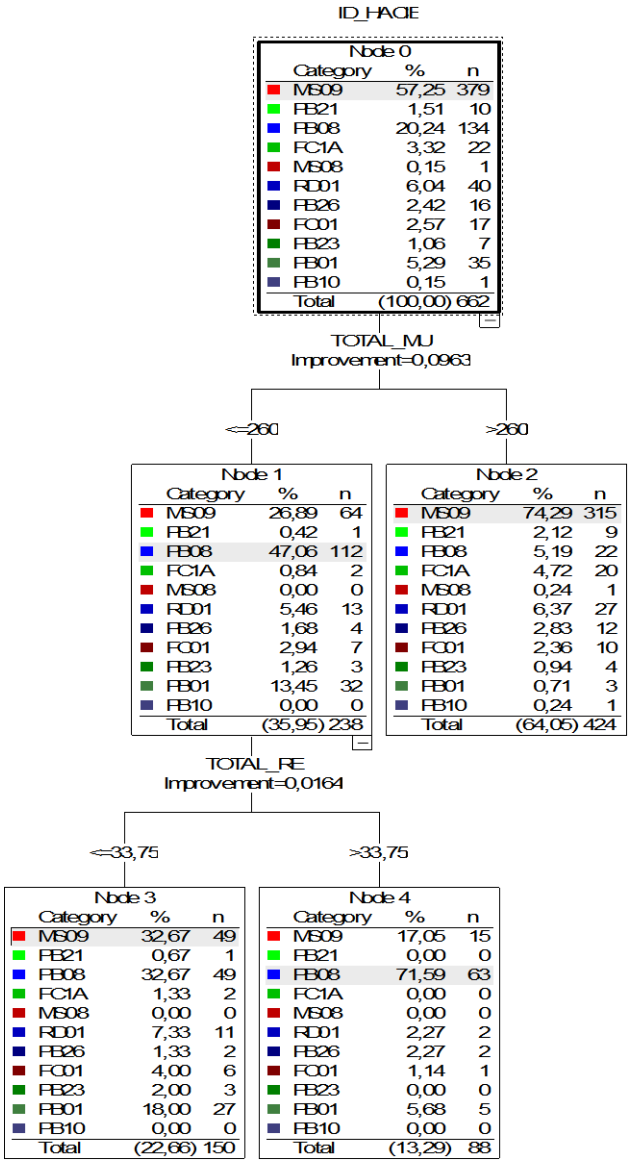
[9] Tomado de:
http://www.google.com.ec/search?hl=es&source=hp&q=prediccion+y+clasificacion+de+datos+en+marketing&meta=&aq=o&aqi=&aql=&oq=&gs_rfai=

Así mismo, se puede verificar que al incluir la variable predictora Total de pies rechazados, la mejora obtenida con el algoritmo C&RT es de 1.64% y clasifica a los nodos 3 y 4 en total de pies rechazados menores o iguales a 33.75 y mayores a 33.75 respectivamente. Determinando así que cuando los pies rechazados son menores iguales a 33.75 la madera de balsa proviene principalmente de la hacienda MS09 en un 32.67% de los casos. Sin embargo si al muestrear los pies rechazados superan 33.75, la probabilidad que la madera de balsa provenga de la hacienda PB08 es de 0.7159.

En el nodo 3 se puede confirmar que el 22.66% de las observaciones presentaron un deterioro en la balsa menor o igual a 33.75.

Podemos concluir el análisis de este árbol de decisión con este algoritmo mencionando que aún cuando la hacienda MS09 es la que tiene mayor cantidad de pies muestreados, no es ésta la que tiene mayor deterioro en la balsa, sino la plantación PB08.

Figura 3.2
Árbol de decisiones – Algoritmo C&RT



Algoritmo CHAID [10]

El algoritmo Detección de Interacción Automática de Chi Cuadrado más conocido por sus siglas en inglés CHAID (Chi Square Automatic Interaction Detection), contrariamente al algoritmo CRT éste actúa por iteración politómica de las categorías o clases. De manera general permite descomponer una base de datos en varios grupos en base al mejor predictor de la variable dependiente (en función de la prueba de la Chi-Cuadrada más significativa).

CHAID permite determinar la mejor partición de cada nodo y aparejar clases o categorías de las variables dependientes o predictoras si no existen diferencias estadísticamente significativas de las clases en relación con la variable dependiente. Se repite el proceso hasta que no queden más parejas significativas. Seguidamente se vuelve a seleccionar en nuevo mejor predictor y el grupo (nodo) será dividido en dos o más clases. Se repetirá el proceso hasta que se llegue a la regla de fin.

Observando la Figura 3.3, se puede verificar que el algoritmo CHAID clasifica a la variable Total de pies rechazados en seis intervalos (nodos) desde menores o iguales a 29.5 hasta mayores a 64 pies rechazados. De esta manera se obtiene con este algoritmo que la hacienda con mayores niveles de deterioro es la hacienda MS09, como se observa en la mayoría de los nodos; diferenciándose el nodo 3 que muestra que cuando los pies rechazados se encuentran en el intervalo (33,5-35,5] la plantación PB08 es de donde proviene esa madera con deterioro.

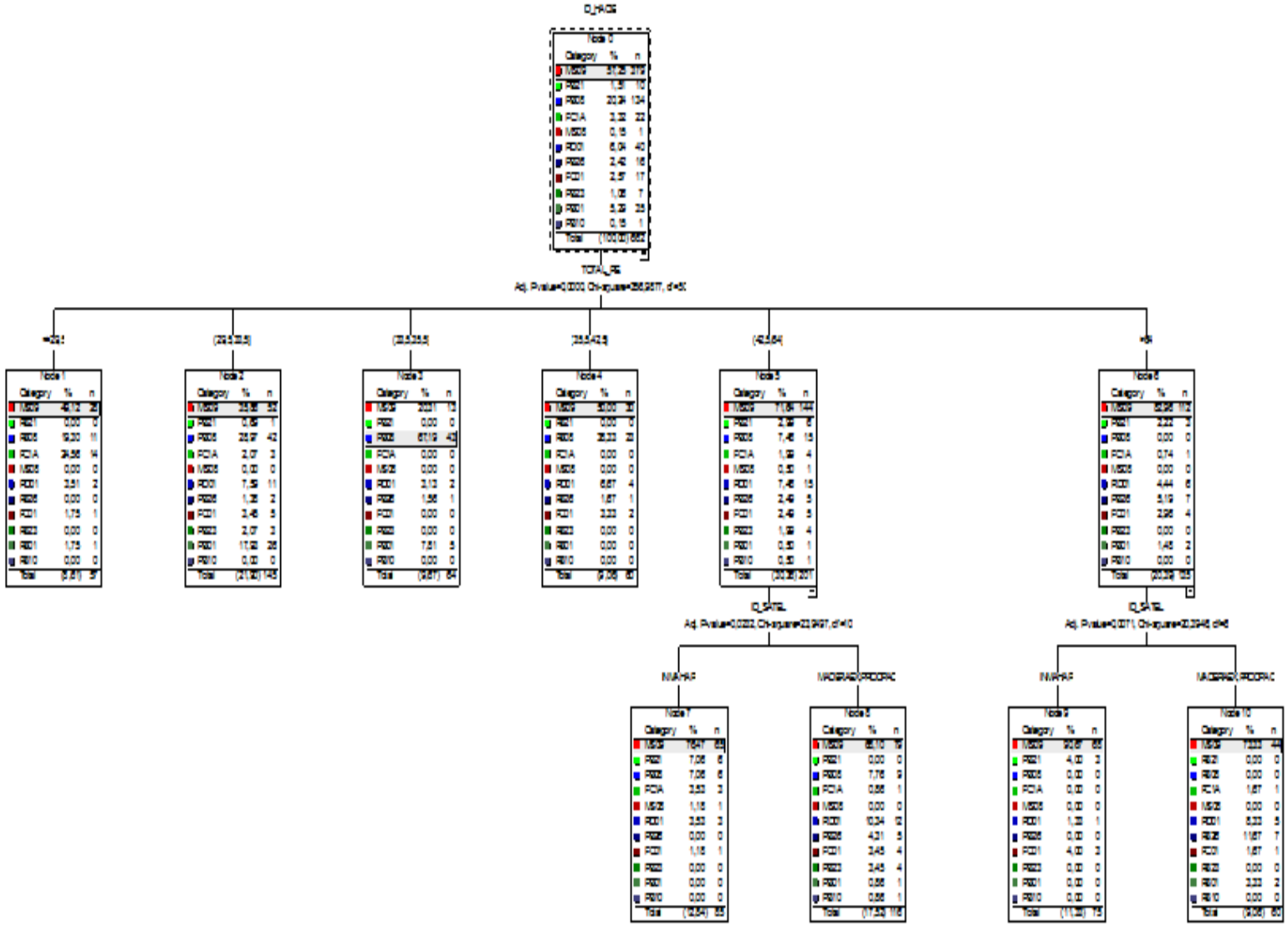
---

[10] Tomado de:
http://www.google.com.ec/search?hl=es&source=hp&q=prediccion+y+clasificacion+de+datos+en+marketing&meta=&aq=o&aqi=&aql=&oq=&gs_rfai=

Incluyendo en este análisis la variable Satélites como predictora, se puede verificar que independientemente de la Satélite donde se haya realizado el muestreo y de la clasificación realizada, la hacienda con mayores niveles de deterioro en la balsa es la hacienda MS09, como se observa en los nodos 9 y 10, tomados por el nivel de significancia menor (0.0071) con respecto a los nodos 7 y 8.

Así también se observa que en la satélite Inmahar ingresa el 11.33% de la balsa que proviene de la hacienda MS09 que tiene niveles de deterioro mayores a 64 pies, mientras que en las otras satélites reciben el 9.06%.
De los 11.33% que clasifica Inmahar con más de 64 pies rechazados el 90.67% corresponde a MS09. Y de los 9.06% que clasifican Madera Export y las otras satélites que prestan servicios a la empresa, el 73.83% proviene de MS09 y el 11.67% de PB08.

Figura 3.3
Árbol de decisiones – Algoritmo CHAID



Algoritmo CHAID exhaustivo [11]

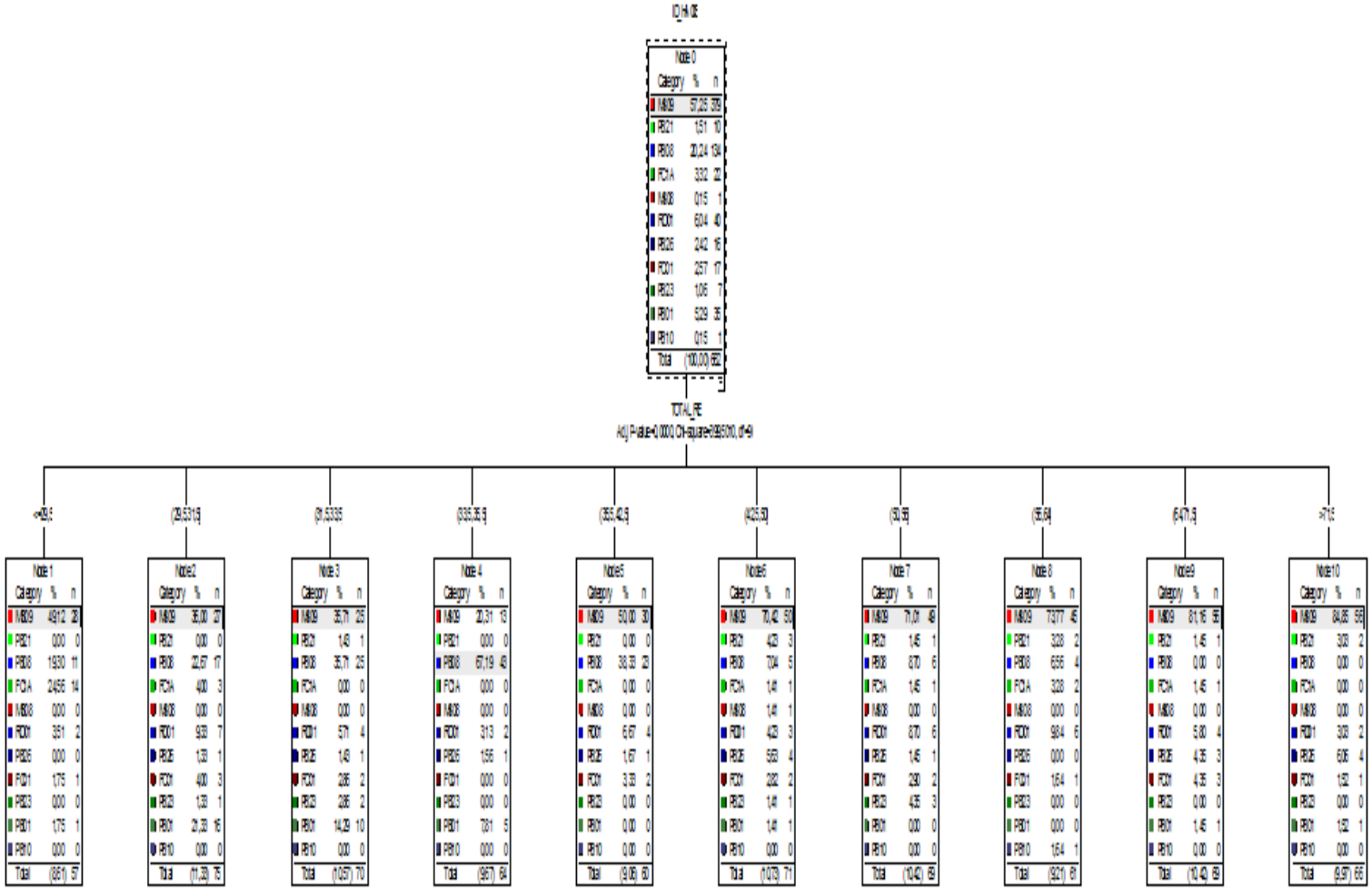
A diferencia del algoritmo CHAID que no garantiza la partición óptima de cada grupo (en cada nodo), el algoritmo CHAID Exhaustive (Biggs, 1991) realiza una investigación exhaustiva de todos los grupos posibles hasta que quede el grupo más significativo. Este algoritmo actúa de la misma manera que el CHAID pero de manera recursiva.

Similar al algoritmo CHAID, el exhaustive realiza particiones de los datos en intervalos que van desde menores o iguales a 29.5 pies rechazados o deteriorados, hasta mayores a 71.5. A diferencia del CHAID, este algoritmo ha clasificado los pies rechazados en más nodos que en el caso anterior, y como se observa en este análisis los nodos resultantes son diez; en los que se puede verificar que la los intervalos resultantes, la mayoría de ellos la hacienda donde proviene la mayor cantidad de balsa con deterioro es la MS09, sin embargo cuando los pies rechazados se encuentran en el intervalo (33,5-35,5] la plantación PB08 presenta un 67.19% del 9.67% de los pies rechazados que se presentan en el nodo 4.

---

[11] Tomado de: http://www.google.com.ec/search?hl=es&source=hp&q=prediccion+y+clasificacion+de+datos+en+marketing&meta=&aq=o&aqi=&aql=&oq=&gs_rfai=

Figura 3.4
Arbol de decisiones – Algoritmo CHAID exhaustive



CONCLUSIONES

Con la realización del presente trabajo se implementó una de las técnicas de minería de datos como son los Árboles de decisión partiendo de la elaboración de un Data Mart que será de utilidad en las satélites; que es el área del negocio donde se obtienen los datos para determinar la relación existente entre la procedencia de la madera y los defectos encontrados en la misma. Considerando este objetivo se construyeron y evaluaron los Árboles de Decisión para obtener los patrones en los datos. Se obtuvieron modelos de predicción basados en la variable dependiente, las haciendas, y las variables predictoras; tomadas de diferente forma en pruebas realizadas a fin de comprobar los resultados.

1. En el árbol elaborado con el algoritmo CHAID, con el total de pies rechazados y satélites como variables predictoras, muestra que la hacienda con mayor nivel de deterioro es la denominada MS09, siguiéndole la hacienda PB08 cuando el total de pies rechazados son mayores a 33,5 y menores o iguales a 35,5.

2. Elaborado este árbol, se pudo constatar que independientemente de la forma como sea calificada la madera en las satélites, la hacienda con mayor nivel de deterioro sigue siendo la MS09. Sin embargo se utilizó en

los siguientes árboles otros algoritmos para confirmar los resultados antes mencionados o verificar si podían diferenciarse de alguna manera.

3. Tomando para el análisis el algoritmo C&RT se pudo concluir que así que cuando los pies rechazados son menores iguales a 33.75 la madera de balsa proviene principalmente de la hacienda MS09 en un 32.67% de los casos. Sin embargo si al muestrear los pies rechazados superan 33.75, la probabilidad que la madera de balsa provenga de la hacienda PB08 es de 0.7159. Además es importante mencionar que en este algoritmo se pudo verificar que aún cuando la hacienda MS09 es la que tiene mayor cantidad de pies muestreados, no es ésta la que tiene mayor deterioro en la balsa, sino la plantación PB08.

4. En el último algoritmo analizado el CHAID exhaustive, se pudo constatar que tiene similitud al CHAID, pero como su nombre lo indica este análisis es más exhaustivo y presentó 10 nodos o clasificaciones para la intervención de la variable predictora total de pies rechazados. Dando como resultado que las haciendas con mayores índices de deterioro son las MS09 y la PB08, especialmente la última para cuando los niveles de deterioro se encuentran entre (33,5-35,5].

5. Finalmente se puede concluir que para mejorar los índices de deterioro actuales, es importante realizar un análisis de las dos plantaciones mencionadas en el numeral 4, a fin de establecer las razones por las cuales estás dos haciendas están teniendo madera defectuosa, y poder determinar si de alguna manera la forma de aserrar la balsa, esta influyendo en los resultados encontrados en las haciendas y en este estudio.

RECOMENDACIONES

1. Utilizar los resultados del proyecto en aplicaciones que permitan mejorar y controlar el proceso de aserrado de la madera en las haciendas para disminuir el nivel de deterioro.

2. Continuar la investigación a partir de los resultados obtenidos, y efectuar otros análisis que se consideren necesarios en lo posterior, tales como efectuar tablas de contingencia, con el fin de mantener controlado el nivel de deterioro existente en la balsa; siguiendo las orientaciones de la fase de evaluación, guiadas por las técnicas de minería de datos.

3. Fomentar el desarrollo de proyectos de Data Marts en las áreas de la organización e implementación de Data Warehouse a escala global para realizar futuros análisis.

GLOSARIO

ASERRADO	Acción de aserrar.
ASERRAR	Cortar las ramas para hacer tablas.
BULTO	Conjunto de piezas de madera de un mismo largo y espesor.
DETERIORO	Inconformidad en la madera, también denominado defecto.
ESPESOR	Grueso o ancho de una pieza de madera.
HACIENDA	Hectáreas de terreno con sembríos de árboles de balsa, también denominada plantaciones.
SATELITES	Denominación para las fábricas que le proveen bloques a Corporación Balsa.
SECADORAS	Cuarto destinado para el secado de la madera.
TROZA	Tronco aserrado por los extremos.
TUMBA	Proceso de tala de árboles.
DICOTOMIA	División de un concepto o una materia teórica en dos aspectos, especialmente cuando son opuestos o están diferenciados entre sí.

BIBLIOGRAFÍA

1. Manual de procedimiento de la empresa.

2. SPSS (2002). Answer Tree 2.0. User‘s Guide. Chicago: SPSS. Statistics Package for the Social Sciences (2002). Answer Tree 2.0. User‘s Guide. Chicago: SPSS.

3. Statistics Package for the Social Sciences (2002). Answer Tree 2.0. User‘s Guide. Chicago: SPSS.

4. 2008, http://www.telematica.com.pe/Product/spss_AnsTree.htm

5. 2008, http://www.spss.com/la/soluciones/data-mining2.htm

6. 2009, http://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos

7. 2010, http://www.diccionarios.com/consultas.php

8. 2010,http://empresas.telefonica.es/asp/catalogo_servicios/aplicaciones_negocio/aplicaciones_soporte/serv_gestion/data.htm

9. 2010, http://www.eustat.es/document/datos/ct_04_c.pdf

10. 2010, http://ininweb.uprm.edu/cc/PRUEBA%20DE%20BONDAD%20Y%20AJUSTE%20.doc

11. 2010, http://es.wikipedia.org/wiki/Grados_de_libertad_(estadÃ-stica)

12. 2010, http://www.google.com.ec/search?hl=es&source=hp&q=prediccion+y+clasificacion+de+datos+en+marketing&meta=&aq=o&aqi=&aql=&oq=&gs_rfai=

13. 2010,http://www.google.com.ec/search?hl=es&source=hp&q=prediccion+y+clasificacion+de+datos+en+marketing&meta=&aq=o&aqi=&aql=&oq=&gs_rfai=