ESCUELA SUPERIOR POLITECNICA DEL LITORAL

INSTITUTO DE CIENCIAS MATEMATICAS
INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

"CREACION E IMPLEMENTACION DE UN CLASIFICADOR SUAVE PARA ESTIMAR LA APROBACION DE MATERIAS DE LOS ESTUDIANTES DEL INSTITUTO DE CIENCIAS MATEMATICAS DE LA ESPOL."

TESIS DE GRADO
PREVIA LA OBTENCIÓN DEL TÍTULO DE:

INGENIERO EN ESTADISTICA INFORMATICA

Presentada por:

CHANG AGUILAR MIGUEL ANGEL

GUAYAQUIL – ECUADOR
2007

AGRADECIMIENTO

A todas las personas que de uno u otro modo colaboraron en la realización de este trabajo.

DEDICATORIA

A DIOS.

A GRACE AGUILAR.

TRIBUNAL DE GRADUACIÓN

Ing. Robert Toledo E.
PRESIDENTE DEL TRIBUNAL

Ing. Juan Alvarado
DIRECTOR DE TESIS

Ing. Luis Rodríguez
VOCAL

Mat. Eduardo Rivadeneira
VOCAL

DECLARACIÓN EXPRESA

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

Miguel Angel Chang Aguilar

RESUMEN

La conjunción de estadística e informática obtiene como resultado conocimiento en estado puro, es muy difícil imaginar de forma independiente que avance existiera en cada rama de forma individual si no existiera la otra, aunque a simple vista podríamos definir como la computación como un área independiente, el desarrollo sostenido en los últimos tiempos de la misma no sería posible si las ciencias numéricas no existieran.

El presente estudio contiene una gran parte de ambas, logrando de cierta forma hacer una interacción natural entre las mismas. La clasificación bayesiana aplicada a la resolución del problema de estimación de aprobación de materias para estudiantes del Instituto de Ciencias Matemáticas de la ESPOL, responde en gran medida a que podamos analizar tan rápidamente la totalidad de registros disponibles para el semestre específico objeto de estimación sobre el modelo definido de sobre como vamos a clasificar, y aplicarlo a una cantidad igual o menor de datos adicionales o extenderlo ha las demás unidades académicas, teniendo resultados que comparados a la realidad, se aproximan de forma bastante confiable como se explica a lo largo del estudio.

El presente documento muestra primero de forma independiente cada disciplina y los detalles relevantes de las mismas, a efectos del estudio

realizado, y luego une las dos áreas para crear el conocimiento y la información relacionada al mismo.

ÍNDICE GENERAL

RESUMEN **I**

ÍNDICE GENERAL **II**

ÍNDICE DE TABLAS **III**

ÍNDICE DE GRÁFICOS **IV**

ÍNDICE DE ABREVIATURAS **V**

SIMBOLOGÍA **VI**

INTRODUCCIÓN **VII**

CAPITULO I

1. CONCEPTOS GENERALES SOBRE SISTEMAS DE INFORMACIÓN Y ESTADÍSTICA----------1

1.1 Antecedentes ----------2

1.2 Procesamiento de la información mediante “Sistemas tradicionales de ficheros” ----------5

1.2.1 Requerimientos de procesamiento de la información de manera más eficiente----------8

1.3 Bases de datos----------9

1.3.1 Bases de datos relacionales----------9

1.3.2 Arquitectura de las bases de datos ---------- 19

1.3.3 Sistemas gestores de bases de datos (SGBD)---------- 20
1.3.4 Bases de datos distribuidas---------- 21
1.3.5 Conceptos en bases de datos---------- 23
1.3.6 Aplicaciones avanzadas en las bases de datos ---------- 28
1.3.6.1Sistemas de Soporte de Decisiones - DSS---------- 29
1.3.6.2Estructura de un DSS ---------- 31
1.3.6.3Diseño y desarrollo de un DSS ---------- 32
1.3.6.4Implementación y uso de un DSS---------- 33
1.3.6.5Factores de Riesgo de un DSS---------- 34
1.3.6.6Estrategias de implementación de un DSS---------- 34
1.3.6.7Análisis y evaluación de un DSS ---------- 35
1.3.6.8Tendencias de los DSS ---------- 36
1.4 Minería de datos ---------- 37
1.4.1 ¿Qué es minería de datos? ---------- 37
1.4.2 Otras definiciones de minería de datos ---------- 39
1.4.3 Como trabaja la minería de datos ---------- 40
1.4.4 Técnicas de Minería de Datos ---------- 41
1.4.5 Metodologías de Minería de Datos ---------- 43
1.5 Clasificación bayesiana ---------- 52
1.5.1 Marco Teórico de Naive Bayes (Bayesiano ingenuo).---------- 52
1.5.2 Clasificador Bayesiano Simplificado.---------- 56
1.5.3 Aplicación de Clasificador Bayesiano Simplificado.---------- 57

CAPITULO II

2. DESCRIPCIÓN DEL PROBLEMA Y PROPUESTA DE SOLUCIÓN

2.2 Planificación académica ---------- 65

2.2.1 Introducción ---------- 65

2.2.2 Requisitos de planificación académica en la ESPOL ---------- 67

2.2.3 El proceso de Planificación académica ---------- 68

2.2.4 Problemas generales en la planificación académica en la ESPOL69

2.2.5 Problemas específicos en la planificación académica en el ICM-ESPOL ---------- 70

2.2.6 Soporte informático a la planificación académica ---------- 70

2.2.6.1 Sistema Académico de la ESPOL ---------- 71

2.3 Contexto especifico del Problema de Negocio ---------- 74

2.4 Propuesta de la Solución ---------- 74

CAPITULO III

3. IMPLEMENTACIÓN Y RESULTADOS

3.1. Desarrollo de la solución ---------- 77

3.1.1. Definición específica del Problema de minería de datos ---------- 77

3.1.2. Diagrama de flujo de la solución ---------- 78

3.1.3. Definición de Variables de Clasificación ---------- 81

3.1.4. Obtención de los datos ---------- 83

3.2. Preparando los datos ---------- 83

3.2.1. Selección de los datos ---------- 83
3.2.2. Limpieza de los datos (Data cleaning) ---------- 85
3.2.3. Transformación de los datos ---------- 86
3.3. Construyendo el modelo ---------- 87
3.3.1. Análisis exploratorio de los datos ---------- 87
3.3.2. Creación de grupos de datos necesarios ---------- 87
3.3.2.1.Datos de entrenamiento ---------- 87
3.3.2.2.Datos de prueba ---------- 88
3.3.2.3.Construcción del Modelo ---------- 88
3.4. Validación del modelo ---------- 89
3.4.1. Resultados de clasificación ---------- 89
3.4.2. Prueba de precisión versus la data real ---------- 91
3.5. Despliegue del modelo ---------- 92
3.5.1. Creación del modelo de despliegue ---------- 92
3.5.2. Evaluar el modelo en el ambiente de producción ---------- 92

CAPITULO IV

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones ---------- 95
4.2 Recomendaciones ---------- 98
BIBLIOGRAFIA ---------- 100

ÍNDICE DE TABLAS

Tabla 1.1.....60
SOLICITUD DE TARJETA DE CRÉDITO

Tabla 1.2.....60
SOLICITUD DE TARJETA DE CRÉDITO SIN RESULTADO

Tabla 1.3.....61
SOLICITUD DE TARJETA DE CRÉDITO CON RESULTADO

Tabla 1.4.....62
SOLICITUD DE TARJETA DE CRÉDITO

Tabla 1.5.....63
VALORES A PRIORI

Tabla 1.6.....63
TABLA DE ENTRENAMIENTO

Tabla 1.7.....64
SOLICITUD DE TARJETA DE CRÉDITO CON RESULTADOS CLASIFICADOS

Tabla 3.1.....82
MATRIZ DE CORRELACIONES DE LAS VARIABLES DEL ANÁLISIS

ÍNDICE DE GRÁFICOS

Gráfico 1.1.....46
De los datos al conocimiento

Gráfico 1.2.....48
Pasos en el Proceso de Minería de Datos

Gráfico 1.3.....54
Evento E en Marco Muestral S

Gráfico 2.1.....70
Estructuración de un Sistema Informático Gerencial

Gráfico 3.1.....81
Diagrama de Flujo de la Solución

Gráfico 3.2.....82
Esquema General de Clasificación

INTRODUCCIÓN

La planificación de actividades es una disciplina que une lo estratégico a lo operativo ó una herramienta de conjunción en la operación que podría ubicarse en el nivel táctico de trabajo, por lo tanto un tema muy importante y aplicable a la mayoría de instituciones educativas debiera ser determinar la demanda para los cursos que se van a planificar en un periodo futuro de actividades, siempre y cuando se trabaje bajo el esquema sobre demanda y no sobre un determinado grupo fijo de materias para aprobar el periodo académico.

CAPITULO I

CONCEPTOS GENERALES SOBRE SISTEMAS DE INFORMACIÓN Y ESTADÍSTICA

INTRODUCCIÓN

El presente capítulo nos permite introducirnos sobre los conceptos generales de informática y estadística relevantes para este estudio, permitiendo una compresión focalizada sobre lo que se pretende aplicar de forma conjunta a lo largo del presente estudio.

1.1 Antecedentes

Siendo lo anterior el único factor antes de aplicar como se define el presente análisis y solución, de otra forma habrá que hacer los ajustes respectivos antes de aplicarlos; podríamos llegar a la pregunta básica que permitirá entender de manera clara lo que persigue el presente trabajo es determinar: ¿Es posible inferir en base a determinados factores si un estudiante aprueba o reprueba en una determinada materia en un ciclo académico contando con información relevante del mismo?, sin tener en cuenta los diferentes enfoques que se pueden tener del termino 'completo' en un tema bastante amplio, el objetivo general del presente trabajo es crear un cimiento para dicho análisis.

Es importante considerar que dada la amplitud del tema se deben tener en cuenta los factores que escapan a cualquier análisis causal más elaborado o 'inferenciable' y sólo considerar aquellos que puedan ser incluidos en el mismo. Las delimitaciones del trabajo pueden originarse de diversas fuentes como por ejemplo: ¿Es más probable que un estudiante apruebe una materia del área computación siendo que este posee un computador personal ó no?, de manera subjetiva podríamos decir que la repuesta a la pregunta es ***sí***, pero para lograr resultados objetivos se trabajará de manera concreta sobre variables de las cuales se tiene alguna fuente confiable y consistente en el tiempo relacionada a los otros

diversos factores a analizar. Implica entonces que nuestro referencial de variables predictoras será el sistema académico de la ESPOL cuyo detalle se define en secciones posteriores del presente documento.

Dentro de las instituciones educativas el proceso de planificación del próximo ciclo debe ser llevado de manera científica mas que improvisada, la administración de dicho proceso requiere que los diferentes directivos de cada unidad establezcan en base a sus normas, resoluciones o demanda los paralelos que van a ser dictados en el termino siguiente inmediato. Este proceso de planificación muchas veces se convierte en una tarea difícil dependiendo de la antigüedad del directivo a cargo de la realización de la misma y con muchas implicaciones dado las actividades que del mismo se desencadenan.

Vamos a definir uno de los objetivos generales que se busca lograr, *la planificación académica debe ser planificada en base a la demanda esperada por materia que se va a dictar y no de manera improvisada o anti-técnica*, en el presente estudio vamos a dar un paso de automatización y especialización en soporte a tan importante proceso. Se cubrirán dos niveles de reportes o dos dimensiones de reporte siendo estos no los únicos posibles de obtener pero sí el objetivo del presente análisis:

1.- Estimar cuan posible es que un estudiante determinado apruebe una materia (en base a características preestablecidas resultantes del modelo aplicado)

2.- Cantidad estimada de estudiantes que aprueben una determinada materia, como información relevante y de valor para el proceso de Planificación académica

Es importante resaltar que la aclaración definida en el literal anterior por cuestiones de delimitación del alcance, ya que el soporte del proceso completo de planificación implica más de una técnica avanzada de tratamiento de datos, que el objetivo del presente trabajo es cubrir la etapa inicial de determinar de manera estadística si un estudiante aprueba o reprueba una materia.

Por lo que es necesario antes de revisar como se desarrollara el presente trabajo, revisar como ha evolucionado el proceso de obtención de información a partir de repositorios de datos o bases de datos y la tecnología que se utiliza para lograrlo.

1.2 Procesamiento de la información mediante “Sistemas tradicionales de ficheros”

Es un hecho fehaciente y para nadie desconocido que la información a través del tiempo ha tomado la relevancia que merece dentro del desarrollo organizacional, institucional y en casi todas las áreas del conocimiento, por tanto es imprescindible en la actualidad dar un tratamiento adecuado a la información. Las organizaciones que no reconocen el valor de la información y que no basan sus decisiones en función al análisis de datos están condenadas en el peor de los casos a no mejorar a permanecer estancadas a través del tiempo.

Los sistemas computacionales se utilizaron inicialmente en las organizaciones para funciones de contabilidad y, como eran funciones imprescindibles, el alto costo de los computadores era fácil de justificar.

A estos primeros sistemas se les llamó sistemas de procesamiento de datos y trataban de imitar los procedimientos manuales existentes. Al principio, la mayoría de los archivos se almacenaban en cinta magnética, ya que el almacenamiento en disco era todavía un mecanismo de alto costo, y se accedía a los datos de forma secuencial, lo que significa que cada registro puede leerse únicamente después de haber sido leídos todos los que le preceden. Estos

archivos se procesaban por lotes, es decir, todos los registros de un archivo se procesaban al mismo tiempo.

Los archivos se empleaban en distintas aplicaciones. Un programa que realiza una tarea específica es un programa de aplicación y un conjunto de programas que trabajan en tareas relacionadas entre sí se llama sistema de aplicación.

Los archivos secuenciales servían para producir facturas e informes una o dos veces al mes pero para tareas rutinarias se necesitaba acceso directo a los datos (procesar directamente un registro dado). Los operadores debían introducir datos redundantes, lo que requería esfuerzo adicional y aumentaba la probabilidad de error.

Estos problemas se resolvieron parcialmente con la introducción de los archivos de acceso directo que permitían la recuperación de registros aleatoriamente. Este tipo de archivos permiten utilizar uno o más campos para identificar un registro.

Desde que se produjo la transición del procesamiento de los datos al procesamiento de la información. Se hace una distinción entre datos e información. Por datos se entienden hechos aislados, mientras que

información son datos procesados, el producto de un proceso de transformación sobre dichos datos.

Los archivos de acceso directo también tenían una serie de deficiencias:

- *Redundancia de datos*. Como muchas aplicaciones utilizaban sus propios archivos, había algunos datos redundantes o repetidos, lo que ocasionaba el aumento de introducción de datos y las probabilidades de inconsistencia entre diversas versiones de los mismos.

- *Pobre control de datos*. El mismo elemento de los datos podía tener diversos nombres según el archivo en que se encontrara, lo cual creaba confusiones.

- *Capacidades inadecuadas de manipulación de datos*. Los archivos secuenciales indexados permitían tener acceso a un registro particular pero no a un conjunto de registros interrelacionados.

- *Esfuerzo de programación excesivo*. Frecuentemente, un nuevo programa requería nuevas definiciones de los archivos que el

programador tenía que ‘recodificar’, creando así una interdependencia muy fuerte entre los programas y los datos.

1.2.1 Requerimientos de procesamiento de la información de manera más eficiente

El paso de manejar la información desde ficheros o archivos planos permitía a las organizaciones que optaban por ese cambio a incrementar su productividad, pasar ahora a trabajar bajo el concepto de bases de datos relacionales eficientes que permitían acceder a los mismos datos y evitaba se cometan los errores comunes implícitos de la administración de información descritos anteriormente. El aumento en los volúmenes de información generada por las organizaciones y los requerimientos gerenciales de tomar decisiones sobre esa información permite que el ingreso de los Sistemas Gestores de Bases de Datos (SGBD) o (DBMS por sus siglas en Inglés) sea bien recibido y para muchos un potente aliado en el desarrollo de sus tareas cotidianas.

1.3 Bases de datos

¿Qué son las "Bases de Datos"?.- Una base de datos es una colección de elementos de datos interrelacionados que pueden procesarse por uno o más sistemas de aplicación. Un sistema de base de datos está formado por una base de datos, un sistema de gestión de bases de datos (SGBD), así como por el hardware y personal apropiado que se encargan de que el esquema de trabajo funcione tal cual como ha sido planificado. Los sistemas de bases de datos superan estas limitaciones de los sistemas orientados a los archivos. Los datos se controlan por medio de un diccionario de datos/directorio, que está controlado por los administradores de las base de datos.

1.3.1 Bases de datos relacionales

Éste es el modelo más utilizado en la actualidad para modelar problemas reales y administrar datos dinámicamente. Tras ser postulados sus fundamentos en 1970 por Edgar Frank Codd, de los laboratorios IBM en San José (California), rápidamente se consolida como un nuevo paradigma en los modelos de base de datos. Su idea fundamental es el uso de "relaciones". Estas relaciones podrían considerarse en forma lógica como conjuntos de datos llamados "tablas". Pese a que ésta es la teoría de las bases de datos relacionales creadas por Edgar Frank Codd, la mayoría de las veces

se conceptualiza de una manera más fácil de imaginar. Esto es pensando en cada relación como si fuese una tabla que está compuesta por registros (las filas de una tabla), que representarían las tablas, y campos (las columnas de una tabla).

En este modelo, el lugar y la forma en que se almacenen los datos no tienen relevancia. Esto tiene la considerable ventaja de que es más fácil de entender y de utilizar para un usuario esporádico de la base de datos. La información puede ser recuperada o almacenada mediante "consultas" que ofrecen una amplia flexibilidad y poder para administrar la información.

El lenguaje más habitual para construir las consultas a bases de datos relacionales es SQL, Structured Query Language o Lenguaje Estructurado de Consultas, un estándar implementado por los principales motores o sistemas de gestión de bases de datos relacionales.

Durante su diseño, una base de datos relacional pasa por un proceso al que se le conoce como normalización de una base de datos, lo cual permite optimizar espacio en disco y “eficienciar” las consultas sobre los datos.

OBJETIVOS GENERALES DE LAS BASES DE DATOS.-

Los objetivos fundamentales de una base de datos son:

- Permitir la disponibilidad de los datos,
- El uso de los datos debe ser controlado. De esta tarea se encarga el sistema de gestión de base de datos (SGBD),
- Los datos se integran de una forma lógica, eliminando redundancias, resolviendo ambigüedades en la definición y manteniendo la consistencia interna entre los mismos, y
- Permitir obtener información sobre los datos almacenados en las tablas.

MODELOS DE BASES DE DATOS.-

Hay 3 modelos fundamentales:

- ***Jerárquico.*** Este modelo presume de que todas las interrelaciones entre los datos pueden estructurarse como jerarquías. Los archivos se conectan entre sí mediante punteros físicos (direcciones físicas que identifican dónde se puede encontrar un registro en disco) o campos de datos añadidos a los registros individuales. Tiene algunas

limitaciones, ya que no todas las relaciones se pueden expresar de forma jerárquica

En red. Debido a la necesidad de manipular las interrelaciones, se desarrolló este modelo de base de datos que maneja relaciones en forma de red en lugar de jerárquicas. También utiliza punteros físicos.

- ***Relacional.*** Descrito anteriormente, tiene la debilidad que tenían los punteros físicos era que había que definir las interrelaciones antes de que el sistema fuera puesto en explotación. Un modelo en el que los datos se representarían en tablas constituidas por filas y columnas, llamadas relaciones. También dos lenguajes para manipular los datos en las tablas: el álgebra relacional y el cálculo relacional. En los sistemas de bases de datos relacionales, los archivos se pueden procesar con instrucciones sencillas, sin embargo, en los sistemas tradicionales se deben procesar de registro en registro.

COMPONENTES DE UNA BASE DE DATOS.-

Hardware. Es el conjunto de dispositivos físicos sobre los que reside una base de datos. Pueden usarse mainframes (Equipos

centralizados con gran capacidad de procesamiento accedidos generalmente desde centrales tontas únicamente para consulta puesto que todo el trabajo se realiza en este equipo) o computadoras PCs para soportar acceso a varios usuarios, o computadoras personales que se utilizan con bases de datos autónomas controladas por un usuario único. Hay que señalar también que las unidades de disco son el mecanismo de almacenamiento principal para las bases de datos.

Debido al avance y el abaratamiento de la alta tecnología, los sistemas de bases de datos se han difundido considerablemente.

- **Software.** Hay dos tipos de software: el sistema de gestión de bases de datos (SGBD) y el software de aplicación (que usa las facilidades del SGBD para manipular las bases de datos). Este último suele ser desarrollado por los empleados de la compañía para resolver un problema específico.

- **Datos.** Los datos tienen que ser cuidadosa y lógicamente estructurados y deben almacenarse de manera precisa en el diccionario de datos.

- **Personas.** Quienes utilizan la información de las Bases de Datos y pueden ser: usuarios (que necesitan información de la base de datos para desarrollar su responsabilidad en el negocio) o

profesionales de la computación (que su responsabilidad reside en el diseño y mantenimiento del sistema de la base de datos), u otro tipo de usuarios para funciones especificadas de información o de validación (control) u otros. [8]

INDEPENDENCIA FÍSICA Y LÓGICA DE LOS DATOS.-

En una base de datos hay que lograr la independencia entre las estructuras lógica y física de los datos, lo que significa distinguir entre datos y aplicaciones.

El concepto de independencia de los datos implica la separación entre el almacenamiento y la organización lógica de los datos tal como éstos se contemplan por los distintos programas de aplicación que hacen uso de la base, con lo que se consigue que unos mismos datos se puedan presentar de distintas formas según las necesidades y, por otra parte, que el almacenamiento de los datos, su estructura lógica y los programas de aplicación sean independientes unos de otros.

INTEGRIDAD.-

La integridad de los datos consiste en mantener la precisión y consistencia de los valores de los datos. Los mecanismos de seguridad protegen la integridad de los datos. También se pueden

mantener en el diccionario de datos restricciones sobre los valores, aunque es una tarea que resulta complicada.

Por último, resaltar que los mecanismos de copias de seguridad y restauración soportados por el SGBD deben preservar los datos de cualquier fallo del sistema.

SEGURIDAD.-

Los DBAs (Administradores de la Base de Datos por sus siglas en inglés Data Base Administrator) pueden restringir el acceso a los usuarios sólo para recuperación o permitir acceso y actualización. La información relativa a los derechos de acceso se almacena en el diccionario de datos.

El acceso a la base de datos también es controlado por un mecanismo de contraseñas; un usuario que quiera acceder al sistema debe dar una contraseña y que el sistema la valide. El encargado de la asignación de contraseñas generalmente el DBA.

Actualmente muchas organizaciones están incorporando a sus estructuras profesionales encargados de controlar accesos a las bases de datos, eventos sobre las bases de datos; funciones necesarias para preservar la seguridad sobra las mismas.

REDUNDANCIA MÍNIMA.-

Para que una base de datos sea efectiva hace falta eliminar en la medida de lo posible las redundancias, es decir, las repeticiones que puedan llevar a error, como el llamar a un mismo campo de distinta manera en varios archivos, ya que si no existe el riesgo de inconsistencia entre las distintas versiones de los mismos datos.

COMPARTIR DATOS.-

Quizás la diferencia más importante entre un sistema basado en archivos y un sistema de base de datos es que los datos se comparten. Hay 3 formas de compartir:

- *Entre unidades funcionales.* El combinar los datos en una base de datos produce que los datos combinados tengan más valor que la suma de los datos en los archivos por separado. A este concepto de combinar los datos para un uso común se le llama integración de datos.
- *Entre diferentes niveles de usuarios.* Se pueden distinguir 3 niveles de usuarios: personal, mandos intermedios y ejecutivos. Estos niveles se corresponden con los 3 diferentes tipos de automatización de los sistemas de negocios: Procesamiento Electrónico de Datos (**PED**), Sistemas de Información de Gestión (**MIS** – Management Information Systems por sus

siglas en ingles) y Sistemas de Apoyo a la Toma de Decisiones (**DSS** – Decision Support Systems por sus siglas en ingles).

Los PED se caracterizan por tener el foco de atención en el nivel operativo del almacenamiento, procesamiento y flujo de los datos, así como procesar eficientemente las transacciones y realizar informes resúmenes para los dirigentes.

Los MIS se caracterizan porque su foco de atención está en la información orientada a mandos intermedios, por la integración de las tareas de PED por sus funciones en los negocios y por la generación de encuestas e informes.

Un STD está más centrado en la decisión y orientado hacia altos ejecutivos.

- *Entre diferentes localidades.* Una base de datos centralizada es una base de datos que está físicamente situada en un único lugar, controlado por una sola computadora. La mayoría de las funciones se llevan a cabo más fácilmente si la base de datos está centralizada. Sin embargo, un sistema de base de datos distribuida (compuesto de varios sistemas de bases de datos operando en los sitios locales y conectados por líneas de comunicación), hace posible que los datos residan donde

se necesitan con más frecuencia, mientras que al mismo tiempo puedan acceder a los mismos otros usuarios no locales.

CONCURRENCIA.-

Gracias al SGBD existe la posibilidad de que varios usuarios tengan acceso de forma rápida y eficiente a los datos de la base. Al centralizar los datos en una base de datos, aumentan las probabilidades de que se dé este caso. Si el SGBD permite esto, seguramente el trabajo realizado por los usuarios se vería dañado, por eso el SGBD debe proteger los datos de la actualización simultánea por otro usuario; para ello utiliza mecanismos sofisticados de bloqueo.

Los conceptos anteriormente descritos pueden ser agrupados en los 3 requerimientos básicos de la información en una empresa o institución con grandes volúmenes de información y que toma decisiones sobre la misma y son:

INTEGRIDAD.- garantía de la exactitud y completitud de la información y los métodos de su procesamiento.

DISPONIBILIDAD.- aseguramiento de que los usuarios autorizados tienen acceso cuando lo requieran a la información y sus activos asociados.

CONFIDENCIALIDAD.- aseguramiento de que la información es accesible solo para aquellos autorizados a tener acceso.

1.3.2 Arquitectura de las bases de datos

Existen 3 niveles de abstracción distintos en los que se podría dividir una base de datos:

Nivel conceptual: consiste en el análisis de las necesidades de los usuarios y la definición de las clases de los datos. Como resultado se obtiene un esquema conceptual con todos los elementos de los datos y sus relaciones.

Nivel externo: es la colección de las vistas de distintos grupos de usuarios sobre la base de datos, las cuales describen los elementos de los datos y sus relaciones.

Nivel interno: está compuesto por la vista física de la base de datos (discos, direcciones, punteros...). Este nivel es responsabilidad de los diseñadores de la base de datos y no de los usuarios.

La implementación de estos 3 niveles requiere que el SGBD haga corresponder cada nivel con el otro.

1.3.3 Sistemas gestores de bases de datos (SGBD)

Un SGBD es un sistema computacional de propósito general que manipula la base de datos. El diccionario de datos/directorio (DD/D) almacena las definiciones de todos los elementos de los datos en la base de datos, así como las interrelaciones que existen entre las diversas estructuras de datos. A esto se le llaman **metadatos** o "datos sobre los datos".

Mediante mecanismos de seguridad, el SGBD limita el acceso al personal autorizado y también lo restringe a ciertos datos. La integridad y la consistencia de la base de datos se protegen por medio de restricciones sobre los valores que pueden tomar los elementos de los datos y por las capacidades de recuperación y respaldo suministradas por el SGBD.

El SGBD proporciona los mecanismos físicos que permiten a varios usuarios tener acceso de forma rápida y eficiente a diferentes datos relacionados. También utiliza mecanismos de

bloqueo para que la actualización de más de un usuario simultáneamente no afecte a los datos.

Se debe permitir a los usuarios formular sus consultas y pedir informes únicos directamente de la base de datos. Por último, el SGBD ofrece al programador una serie de herramientas que facilitan la creación de software de aplicación.

1.3.4 Bases de datos distribuidas

CONCEPTO DE DISTRIBUCIÓN.-

Un sistema de base de datos distribuida consiste en varios sistemas de bases de datos operando en los sitios locales y conectados por líneas de comunicación.

PROCESAMIENTO DISTRIBUIDO.-

Una consulta o una actualización deja de ser un proceso simple controlado por un único módulo de software, se convierte en varios procesos cooperando entre sí controlado por varios módulos independientes. Pero para que funcione con efectividad, deben estar disponibles tecnologías adecuadas de comunicación y los SGBDs deben poder comunicarse entre sí.

VENTAJAS E INCONVENIENTES.-

Una clara ventaja es que es posible ubicar los datos en lugares donde se necesitan con más frecuencia, aunque también al mismo tiempo se permita a usuarios no locales acceder a los datos según sus necesidades. Esto mejora la relación costo-efectividad y la autonomía local.

PLATAFORMAS CLIENTE-SERVIDOR.-

Las plataformas cliente/servidor son sistemas abiertos, lo que significa que tratan de lograr la interoperabilidad entre dos o más sistemas, es decir que se comuniquen y contribuyan cada uno a alguna parte del trabajo común.

Los ordenadores clientes están interconectados a un servidor, así, un cliente que necesite hacer una consulta o actualización en la base de datos, envía una petición al servidor de la base de datos y este le devuelve los datos solicitados.

Al principio los servidores se instalaron para controlar la impresión y el acceso a los archivos, pero hoy la mayoría son servidores de base de datos, mientras los clientes son los que manipulan al Interfaz Gráfica del Usuario (GUI).

1.3.5 Conceptos en bases de datos

La siguiente tabla muestra en la columna TERMINO, términos utilizados muy frecuentemente en el manejo de bases de datos y desarrollados o incorporados durante la evolución que ha tenido en el tiempo la tecnología desarrollada en el manejo de SGBD, y en la columna DESCRIPCION la descripción de estos:

TÉRMINO	DESCRIPCIÓN
BDR	Base de Datos Relacional. Sistema de almacenamiento de datos basado en un conjunto de tablas unidas mediante relaciones.
BDM	Base de datos Multidimensional. Base de datos de estructura basada en dimensiones orientada a consultas complejas y alto rendimiento. Puede utilizar un SGBDR en estrella (Base de datos Multidimensional a nivel lógico) o SGBDM (Base de datos Multidimensional a niveles lógico y físico o Base de datos Multidimensional Pura)
OLTP	On Line Transactional Processing. Procesamiento Transaccional En Línea. Se trata de los procesos clásicos de tratamiento automático de información, que incluyen Altas, Bajas, Modificaciones y Consultas.
OLAP	On Line Analytical Processing. Procesamiento Analítico

	En Línea. Se trata de procesos de análisis de información. Estos sistemas están orientados al acceso en modo consulta.
DW	DataWarehouse. Sistema almacén de datos que reúne la información generada por los distintos departamentos de una organización. Pretende conseguir que cualquier departamento pueda acceder a información de cualquiera de los otros mediante un único medio, así como obligar a que los mismos términos tengan el mismo significado para todos. Es un almacén de datos históricos, utilizado por una herramienta OLAP para procesar información, elaborar informes y vistas. También se define como un conjunto de datos orientados por tema, integrados, variables en el tiempo y no volátiles que se emplea como apoyo a la toma de decisiones.
Datamart	Sistema que mantiene una copia de parte de un DataWarehouse para un uso departamental. Almacén de datos históricos relativos a un departamento de una organización, utilizado por una herramienta OLAP para procesar información, elaborar informes y vistas.
EIS	Executive Information Systems. Sistemas de información para directivos.
DSS	Decision Support System. Sistema de ayuda a la toma de decisiones.

Data Mining	Proceso no trivial de análisis de grandes cantidades de datos con el objetivo de extraer información útil. Por ejemplo, se trata de aplicar algoritmos de clasificación de datos para realizar predicciones futuras, o estudios de correlación entre variables aparentemente independientes. Para ello, es común la utilización de Redes Neuronales o Algoritmos Evolutivos.
KDD	Knowdledge Discovery in Databases en ingles. Es el descubrimiento de conocimiento en las bases datos, supone de técnicas especificas, generalmente estadísticas, para la obtención de información valiosa de la base de datos que no podría ser obtenida por las consultas comunes sobre las mismas.
Drill Down	Descomponer (visualmente) en detalle un dato según una jerarquía de una dimensión.
Drill Up	Agregar (visualmente) un dato según una jerarquía de una dimensión.
Roll Up	Proceso que calcula para un indicador, y para una o más de las dimensiones por las que ese indicador se mueve, los valores agregados o padres sucesivos a partir de la suma de sus hijos, según las jerarquías especificadas, pudiendo poseer cada dimensión más de una jerarquía. Por ejemplo, es el proceso que suma los ingresos por cada provincia acumulándolos en los ingresos de la comunidad autónoma correspondiente. Se trata de una

	función que relaciona los valores de dos niveles jerárquicos distintos y adyacentes en una dimensión, transformando un grupo de datos de un nivel en un único dato asignable a otro valor del nivel superior.
Spread	Proceso que produce dentro de una dimensión una progresión o algún tipo de reparto proporcional de la cantidad asignada a un elemento entre otros de acuerdo a algún criterio.
Dimensión	Criterio de clasificación de información. Eje de análisis. Lista de valores que proporciona un índice a los datos. Por ejemplo: <Tiempo>, <Geografía>, <Producto>
Rotación	Cambio de dimensiones en un informe.
Elementos de una dimensión	Posibles valores de un eje de análisis. Por ejemplo, "Enero de 1998", "Trimestre 4 de 1998", o "1996" para la dimensión <Tiempo> y "Bilbao", "Andalucía" o "Zona Norte" para la dimensión <Geografía>
Indicador, Medida, Hipercubo, Variable, Fórmula	Objeto de estudio. Cada indicador tiene asociada una serie de dimensiones sobre las que se pueden clasificar sus valores, se dice que se mueve por un cierto número de dimensiones. Por ejemplo, algunos indicadores son: Ingresos (<Tiempo>, <Geografía>, <Producto>) Número de Empleados (<Tiempo>, <Geografía>) Si el indicador contiene datos almacenados se habla de Variable Multidimensional. Si por el contrario, lo que se

	almacena es la expresión para calcular esos datos a partir de otros (que puede ser una fórmula o un programa), se habla de Fórmula Multidimensional
Jerarquía	Forma de agrupar todos o sólo algunos de los elementos de una dimensión con relaciones padre-hijo. Casi siempre, pero no obligatoriamente, implican que el padre se calcula como la suma de sus hijos. Una dimensión puede tener cero, una o varias jerarquías.
Relaciones o Atributos	Definen vínculos entre valores de dos dimensiones, de forma que cada valor de una dimensión puede estar relacionado con uno o más valores de otra dimensión
Celda	Estructura mínima de almacenamiento formada por la intersección de un valor de cada una de las dimensiones que componen el cubo. Puede contener o no contener datos
SQL	Structured Query Language. Lenguaje de Consultas Estructurado. "Select Query Language". Lenguaje orientado a la creación de consultas de bases de datos relacionales.
RDBMS	Relational DataBase Management System. Sistema de gestión de bases de datos relacionales. Programa que sirve para crear, diseñar y manipular bases de datos relacionales
OLTP to OLAP	Proceso de migración de datos desde un sistema OLTP a

	uno OLAP. Esta migración es habitualmente el elemento crítico en un desarrollo OLAP
R-OLAP	Arquitectura de Base de Datos Multidimensional en la que los datos se encuentran almacenados en una Base de Datos Relacional, normalmente con en forma de estrella (copo de nieve, araña).
M-OLAP	Arquitectura de Base de Datos Multidimensional en la que los datos se encuentran almacenados en una Base de Datos Multidimensional, que mejora los tiempos de acceso a costa de mayores necesidades de almacenamiento y retardos en las modificaciones.
H-OLAP	Arquitectura que combina las tecnologías ROLAP y MOLAP. En HOLAP, el soporte de almacenamiento de datos y el motor de generación de vistas contienen elementos de ambas tecnologías. Pretende combinar las ventajas de cada una sin sus inconvenientes.

Fuente: www.microsoft.com/spain/sql/productinfo/datasheet/DataMining.mspx

1.3.6 Aplicaciones avanzadas en las bases de datos

El uso intensivo de tecnologías en los diferentes tipos de negocios, por ejemplo los Sistemas de Soporte de Decisiones (DSS por sus siglas en ingles (Decision Support Systems)), provee a las personas encargadas de tomar las decisiones más importantes dentro de la organización de los aplicativos que soporten dichas decisiones.

Desde los inicios del procesamiento de datos sobre las bases de datos se ha facilitado en gran medida los análisis de datos, el avance es tan vertiginoso que actualmente al movimiento se lo ha denominado inteligencia de negocios o BI (Business Intelligence).

1.3.6.1 Sistemas de Soporte de Decisiones - DSS

Un Sistema para el Soporte de Decisiones o DSS, es un sistema interactivo que permite al usuario un acceso fácil a los modelos de decisión y a los datos obtenidos de una amplia gama de fuentes, para respaldar las tareas semi-estructuradas de la toma de decisiones, típicamente relacionadas con fines organizacionales. Es una aplicación de información que esta diseñada para asistir a una organización en la toma de decisiones, a través de datos suministrados por herramientas inteligentes de negocio (en contraste con una aplicación operativa que recoge datos en el curso normal de las operaciones de la organización). Una tipo de predicción típica buscada por este tipo de sistemas es encontrar las consecuencias y alternativas de decisiones diferentes, dada la experiencia pasada en un contexto dado.

Un sistema de soporte de decisiones puede presentar la información gráficamente y puede incluir un sistema experto o de

inteligencia artificial. Además puede estar dirigido a ejecutivos de negocio o a otro grupo de profesionales del conocimiento.

Características de un DSS:

- Esta dirigido a resolver problemas menos estructurados y de bajo nivel de especificad, que enfrenta la alta gerencia
- Combina el uso de modelos o las técnicas de análisis con el acceso tradicional a los datos y a las funciones de recuperación
- Enfatiza la flexibilidad y la adaptabilidad para acomodarse a los cambios en el ambiente y en el enfoque que dan los usuarios a la toma de decisiones

Un principio del diseño del DDS es concentrarse menos en la **eficiencia** (realizar tareas con rapidez y reduciendo los costos) y más en la **eficacia** (realizar la tarea correcta).

Los DSS se desarrollan a menudo con una decisión específica o con una clase de decisiones bien definidas para resolver un problema.

1.3.6.2 Estructura de un DSS

La estructura es una generalización sobre un campo, que ayudan a poner en perspectiva muchos casos e ideas específicas. La estructura de Gorry-Scott-Morton es el modelo de sistema de información relacionado con conocimientos y con control de sistemas mas completo, y se basa en la clasificación de problemas en tipos estructurados y no estructurados, así como también el horizonte de tiempo de las decisiones. Esta estructura caracteriza las actividades de DSS en dos dimensiones:

1.- El grado de estructuración del proceso de decisión que es respaldado.

2.- El nivel gerencial en el cual tiene lugar la toma de decisión.

La dimensión del nivel de administración se desglosa en tres partes:

1.- Control operativo

2.- Control gerencial

3.- Plantación estratégica

La dimensión de la estructura de decisiones también se desglosa en tres partes:

1.- Estructurada

2.- Semi-estructurada

3.- No estructurada

El grado en que se estructura un problema o una decisión corresponde de manera general al grado en que el problema o decisión puede ser automatizado o programado.

1.3.6.3 Diseño y desarrollo de un DSS

La creación de prototipos es el método más popular del diseño y desarrollo de un DSS. La creación de prototipos por lo general evade la definición habitual de requerimientos. Los requerimientos de sistema evolucionan a través del proceso de aprendizaje del usuario. Los beneficios de la creación de prototipos incluyen los siguientes:

- El aprendizaje esta incorporado de manera explicita al proceso de diseño debido a la naturaleza interactiva del diseño del sistema
- La retroalimentación proveniente de las iteraciones del diseño es rápida para mantener un proceso efectivo de aprendizaje para el usuario

- La participación del usuario en el área del problema puede ayudar al usuario a sugerir mejoras al sistema
- La creación del prototipo inicial debe ser de bajo costo
- La creación del prototipo inicial debe ser de bajo costo

1.3.6.4 Implementación y uso de un DSS

Los DSS son difíciles de implementar a causa de su naturaleza discrecional. Usar un DSS para resolver un problema representa un cambio en el comportamiento del usuario. Implementar un DSS es un ejercicio de cambio en la organización. El principal desafío es lograr que los usuarios acepten el uso del software. Cambiar un comportamiento implica seguir los pasos siguientes:

- **Descongelar:** Este paso altera las fuerzas que actúan sobre las personas, de modo que los individuos son distraídos lo suficiente como para que cambien. La descongelación se logra ya sea aumentando la presión para que se cambie o reduciendo algunas de las amenazas de resistencia al cambio
- **Trasladar:** Este paso presenta una dirección de cambio y el proceso verdadero de aprender nuevas actitudes
- **Recongelar:** Este paso integra las actitudes cambiadas en la personalidad del individuo

1.3.6.5 Factores de Riesgo de un DSS

Los desarrolladores deben estar preparados para superar cualquiera de los ocho factores de riesgo de la implementación del DSS:

1. Usuario inexistente o no dispuesto
2. Numerosos usuarios o implementadores
3. Cada vez menos usuarios o personas que implementen o mantengan
4. Incapacidad para especificar el propósito o los patrones de uso de manera anticipada
5. Incapacidad para predecir y para amortiguar el impacto para todas las partes
6. Falta de soporte o pérdida del mismo
7. Falta de experiencia con sistemas similares
8. Problemas técnicos y aspectos de eficiencia de costo

1.3.6.6 Estrategias de implementación de un DSS

Las estrategias se enfocan en mayor proporción en planificar los riesgos posibles y prevenir que los mismos ocurran:

- Dividir el proyecto en piezas manejables
- Mantener la solución sencilla

- Desarrollar una base satisfactoria de apoyo
- Satisfacer las necesidades del usuario e institucionalizar el sistema

1.3.6.7 Análisis y evaluación de un DSS

La verdadera prueba de un DSS reside en si la misma mejora la toma de decisión de un gerente, que es algo que no puede ser fácilmente medido. Además de los sistemas DSS muy pocas veces tienen como consecuencia desplazamientos de costos como por ejemplo una reducción de personal o de otros gastos.

Adicionalmente, como los sistemas DSS por naturaleza evolucionan, les faltan fechas de terminación claramente definidas.

Usar un método incremental para desarrollar un DSS reduce la necesidad de evaluación. Desarrollando un paso por vez y obteniendo resultados tangibles al final de cada paso, el usuario no necesita hacer compromisos extensos de tiempo y dinero al comienzo del proceso de desarrollo.

El diseñador y el usuario del DSS deben usar criterios amplios de evaluación, los cuales deben incluir:

- El análisis tradicional de Costo/Beneficio
- Cambios en el procedimiento
- Evidencia de una mejora en la toma de decisión
- Cambios en el proceso de decisión

Algunas tendencias comunes en el uso de los DSS incluyen:

- La necesidad de los gerentes por información más exacta es un motivador importante para el desarrollo del DSS
- Cada vez menos evaluaciones de los DSS usan el tradicional análisis Costo/Beneficio
- Los usuarios finales son por lo general quienes motivan para que se desarrolle un DSS
- Los usuarios perciben la flexibilidad como el factor mas importante que influye en el éxito del sistema

1.3.6.8 Tendencias de los DSS

Las tendencias de los DSS incluyen:

- Mejora gradual y mayor especificad de las destrezas en el desarrollo e implementación de los DSS
- Adelantos en la base de datos y en las capacidades sobre la graficación de reportes. [7]

1.4 Minería de datos

1.4.1 ¿Qué es minería de datos?

Es una metodología de análisis de información que usa técnicas de varias ramas científicas como la estadística, la investigación de operaciones, entre otras; que permite disponer de información base para la toma de decisiones haciendo en la mayoría de los casos uso intensivo de las tecnologías de información disponibles permitiendo así mejorar la forma en que se fundamentan las mas importantes decisiones en diferentes organizaciones, en diversas áreas de conocimiento y habilitando las guías de mejores practicas de administración que sugieren la toma de decisiones basada en información.

De entre las muchas definiciones que pueden tenerse de lo que significa minería de datos, establecemos las tres siguientes como las que mejor cubren el concepto desde una perspectiva individual y concreta:

- “Extracción de información oculta y predecible de grandes bases de datos utilizando particulares algoritmos y presentando modelos o determinados patrones a partir de datos”

- “Es el proceso de extracción de información significativa de grandes bases de datos, información que revela inteligencia del negocio, a través de factores ocultos, tendencias y correlaciones para permitir al usuario realizar predicciones que resuelven problemas del negocio proporcionando una ventaja competitiva”

- “Se podría definir la Minería de Datos (Data Mining) como un proceso interactivo que combina la experiencia sobre un problema dado con variedad de técnicas tradicionales de análisis de datos y tecnología avanzada de aprendizaje automático, con el objetivo de estimar modelos de predicción validos”

Si bien es de gran importancia para las organizaciones enfocarse y buscar vías para la “generación de valor”, lograr diferenciarse de la competencia y tomar las decisiones adecuadas fundamentadas en procesos de soporte idóneos, esta no es un camino fácil aunque existen muchos aplicativos que cubren de amplia forma en cuanto a lo denominado “Inteligencia de Negocio” donde impera el gigante SAP compitiendo duramente con aplicaciones desarrolladas en Oracle, MS SQL Server, y otras cuantas soluciones que proveen a los administradores generalmente de multinacionales o empresas

nacionales de gran tamaño donde se manejan este tipo de herramientas tal como se describió anteriormente; cada mercado es diferente y no se puede cubrir a todos con la misma ‘receta’ por mas genérica que sea, se debe desarrollar una solución acorde a cada negocio con personal adecuado y proveyendo los suficientes recursos para dicha tarea.

La minería de datos utiliza técnicas y utilitarios diversos desde los métodos científicos estadísticos de análisis multivariado hasta el almacén de datos de algún SGBD (Sistema Gestor de Base de Datos), por lo tanto la minería de datos no es en si una rama de la estadística sino mas bien una evolución de la ciencia aplicada al uso colectivo y la generación de valor en los negocios, soportada fuertemente por la tecnología. [3]

1.4.2 Otras definiciones de minería de datos

- “Extracción de información oculta y predecible de grandes bases de datos utilizando particulares algoritmos y presentando modelos o determinados patrones a partir de datos”
- “Es el proceso de extracción de información significativa de grandes bases de datos, información que revela inteligencia del negocio, a través de factores ocultos, tendencias y correlaciones

para permitir al usuario realizar predicciones que resuelven problemas del negocio proporcionando una ventaja competitiva" [3]

1.4.3 Como trabaja la minería de datos

Crear un modelo de minería de datos puede ser de forma simple comparado a un modelo de manufactura en donde que es lo básico que se necesita para trabajar – "la materia prima" o sea "los datos"- como segundo requerimiento se necesita tener el proceso mecánico – "el algoritmo" – este a su vez produce el – "modelo de minería" – la diferencia entre el proceso de manufactura y el proceso de minería de datos esta en que en lugar de crear el producto haciendo uso de medios mecánicos, se elabora el mismo haciendo uso de medios matemáticos y algún lenguaje de programación.

La minería de datos surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos. De forma general, 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 del confronto entre la

información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.

1.4.4 Técnicas de Minería de Datos

Agrupamiento.- El propósito de estas técnicas es agrupar un conjunto de elementos relacionando aquellos que sean semejantes y al mismo tiempo que sean suficientemente diferentes de otro grupo de elementos formados. A este tipo de algoritmos se le conoce como no dirigido, pues no se conoce con antelación el grupo especifico al que pertenece una instancia, sino que de acuerdo a los datos, los grupos se van formando, según sus semejanzas y diferencias. Dentro de las aplicaciones del agrupamiento se encuentran: reducción de datos, generación de hipótesis, y predicción basada en grupos.

Análisis de series de tiempo.- El pronostico de series de tiempo estima valores aun no conocidos para guiar sus predicciones.

Asociación.- El objetivo de la asociación es encontrar aquellos artículos (sucesos) que tienden a aparecer juntos en algún momento dado.

Predicción.- Existen dos tipos de algoritmos para realizar predicciones:

Regresión.- El objetivo de este tipo de análisis es determinar, de acuerdo a un resultado dado, el valor de los parámetros que produjeron ese resultado.

Clasificación.- La clasificación trata de encontrar las características que identifican a un grupo para ser clasificado dentro de cierta clase.

Entre los algoritmos de clasificación se encuentran:

Análisis discriminante.- Busca determinar la localización optima de una línea que actúa como frontera en los diferentes casos.

K-vecinos más cercanos.- conociendo ciertos individuos similares el algoritmo forma un grupo de k-individuos, de acuerdo a sus características. Cuando aparece un nuevo individuo, este se puede clasificar en cierto grupo de acuerdo a su semejanza con los k individuos pertenecientes a ese grupo.

Redes neuronales.- Este tipo de algoritmos intenta emular el funcionamiento de los cerebros de los seres vivos mediante capas de ‘neuronas’, que son funciones matemáticas con un comportamiento determinado.

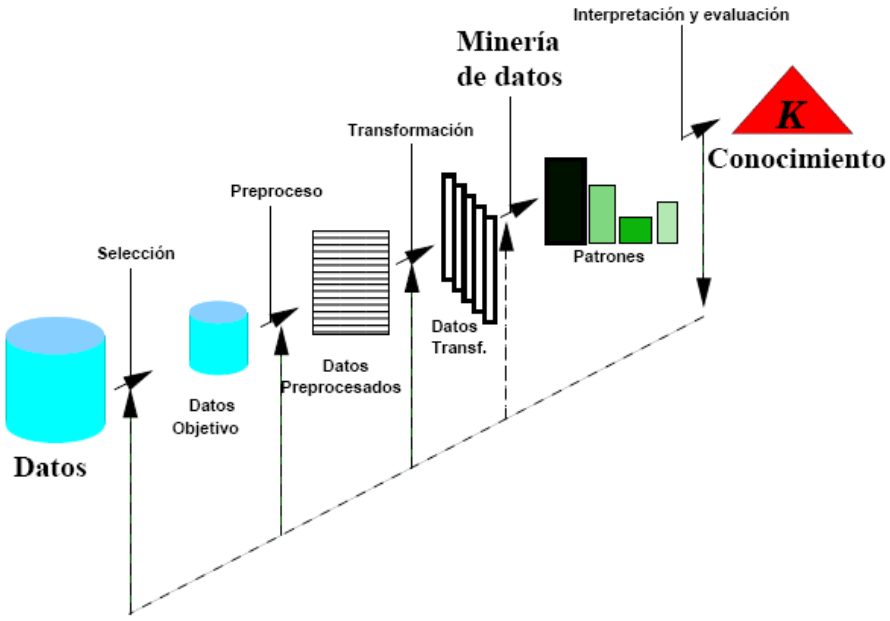
Árboles de decisión.- Estos algoritmos 'aprenden' reglas a partir de datos, tratando de obtener la descripción mas sintética que represente de forma mas cercana los datos originales.

Clasificador Bayesiano.- Método basado en la teoría de la probabilidad, que usa frecuencias para calcular probabilidades condicionales para calcular predicciones sobre nuevos casos. [5]

1.4.5 Metodologías de Minería de Datos

El proceso de minería de datos tiene como objetivo extraer conocimiento a partir de los datos. Se describe en la figura 1 de manera general el proceso de llegar de datos a conocimiento.

Gráfico 1.1
De los datos al conocimiento



Es importante definir el enfoque metodológico a usarse, para trabajar con Minería de Datos se tiene también una guía de trabajo o metodología general aplicable a cualquier modelo y que se detalla en seis pasos descritos de manera detallada en el presente capitulo, como ya hemos establecido al proceso de minería de datos este será adecuado a los fines iniciales del trabajo de investigación y cubrirá de manera exitosa la solución del problema propuesto de forma inicial.

Se presenta la siguiente metodología de trabajo para realizar una minería de datos:

PASO 1.- DEFINICION DEL PROBLEMA

PASO 2.- PREPARACION DE LOS DATOS

PASO 3.- CONSTRUCCION DEL MODELO

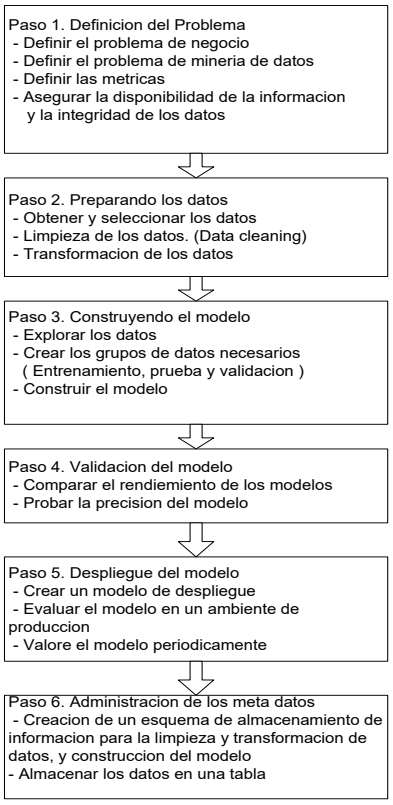
PASO 4.- VALIDACION DEL MODELO

PASO 5.- DESPLIEGUE DEL MODELO

PASO 6.- ADMINISTRACION DE LOS META DATOS

Cada paso será detallado de manera específica a continuación:

Gráfico 1.2
Pasos en el Proceso de Minería de Datos



Paso 1.- Definición del Problema

Antes de comenzar con el proceso de minería, se debe tener claro los datos de los que disponemos así como también el problema de negocio que tratamos de resolver. Esto incluye el análisis de los requerimientos de negocio, definir el alcance que se va a cubrir sobre el problema, definir las métricas o medidas con las que el modelo será evaluado, y definir el objetivo final del proceso de la minería de datos.

Nos debemos plantear las siguientes preguntas:

¿Qué es lo que se espera que hagamos?
¿Qué atributos de los datos deseamos tratar de predecir?
¿Qué tipos de relaciones entre los datos estamos tratando de encontrar?
¿Deseamos hacer predicciones o descubrir patrones de interés en los datos?
¿Cómo están distribuidos los datos?
¿Cómo están las columnas o tablas relacionadas?

Una vez que encontremos respuestas adecuadas o que justifiquen las preguntas expuestas podemos establecer el horizonte del trabajo y pasar al siguiente punto.

Paso 2.- Preparando los datos

Haber definido el problema nos pone en un siguiente nivel dentro del proceso, la preparación de los datos puede ser un proceso con atolladeros que inicia en si desde la obtención de los datos, debemos llegar a la fuente correcta de los datos y no siempre es tarea sencilla lograr obtener los datos, por ejemplo obtener la base de datos de clientes, obtener los registro de las ventas de un determinado periodo, entre otra información que es considerada un tesoro por el valor que en si se encuentra oculto entre sus registros y es muy cuidado en las organizaciones.

Generalmente los datos presentan incongruencias o errores fácilmente identificables como “ingresos erróneos de registros”, p.e. un estudiante que haya nacido en 1900 y que a la hora de tomar una materia tenga 106 años, este tipo de errores obligan a establecer dentro del proceso una actividad de corrección o de limpieza de datos conocida como el “Data Cleaning” y el problema principal de la limpieza de los datos es que nos detiene, es tiempo muerto que no nos permite avanzar en mayor o menor medida dependiendo de la cantidad de errores por corregir que existan. En realidad filtrar los problemas y corregirlos no es un trabajo tan sencillo teniendo en cuenta que generalmente tenemos una gran cantidad de datos por

revisar, es aquí donde se usan las técnicas estadísticas exploratorias como mínimo y máximo, promedio y desviación estándar, distribución de los datos entre otras para tener un adecuado panorama de los datos y su integridad.

En la preparación de datos se debe señalar además otro factor importante antes de construir el modelo de minería, **discretizar** los datos que en resumen es la creación de nuevas columnas de datos establecidas de las columnas originales asignando valores de clases a los valores originales para obtener un mejor nivel de clasificación. Se debe justificar por que discretizar, p.e. el factor socioeconómico de un estudiante en la ESPOL o factor p que se encuentra en una escala de 3 (mínimo) hasta 50 (máximo) y que describe la situación socioeconómica del entorno de cada estudiante y que potencialmente puede influir dentro de los parámetros a investigar o sea si contribuye o no a que un estudiante apruebe una materia especifica.

Paso 3.- Construyendo el modelo

El mas importante concepto en minería de datos es precisamente conocer los datos, no entender la estructura de los datos es similar a no saber que tarea realizar o que columnas se deben incluir en la construcción del modelo. Al igual que ir a una reunión de negocios

sin estar preparado o dictar un seminario sin preparar el material y enfocar los tipos de asistentes al mismo, se pierde efectividad y enfoque, de la misma forma en minería de datos no conocer los datos antes de construir el modelo de minería. Antes de la construcción del modelo es necesario separar los datos de manera aleatoria en dos grupos:

Grupo de entrenamiento – Para construir el modelo.
Grupo de pruebas – Para validar el modelo.

En donde se tiene que la construcción del modelo se la realiza con los datos de entrenamiento y la validación con los datos de prueba, aunque exista una regla especifica sobre que porcentaje de los datos haya que asignar a cada grupo una distribución adecuada podría ser 60% para datos de entrenamiento y 40% para datos de prueba con los cuales de deberá calcular el poder de clasificación del modelo y ajustar el modelo o confirmar la eficiencia del mismo.

Después de explorar los datos y haber realizado un Data Clearing adecuado para los fines perseguidos, se debe proceder a construir el modelo con los datos de entrenamiento. Este procedimiento se ejecuta tal como suena, se debe ingresar los datos como parámetros de entrada para el modelo y modificar en caso de ser necesario los parámetros del mismo para entrenar el modelo y corregir alguna

desviación que pueda afectar los resultados. Se detalla a continuación los pasos sugeridos para construir el modelo de minería:

Seleccionar las columnas que intervendrán en el modelo.

Seleccionar el modelo a construir

Ajustar los parámetros

Entrenar el modelo

Paso 4.- Validar el modelo

Después de haber construido el modelo, es necesario determinar cuan bien este se desempeña. No es adecuado pasar un sistema a producción mientras no se a comprobado cuan bien trabaja y si las especificaciones con adecuadas y de igual forma no seria adecuado ejecutar el modelo en el ambiente de usuario final sin verificar previamente cuan bien predice o clasifica el mismo. Generalmente se construyen varios modelos y se realizan pruebas de cada uno, en este punto intervienen los datos de prueba con los cuales podre validar la efectividad del modelo construido.

Paso 5.- Despliegue del modelo

Este es el paso donde se comienzan a ver los resultados, es el primer gran examen que se le toma al modelo y se debe tener claro

que se tomaran decisiones sobre los resultados del mismo. No es permitido descubrir fallas con el modelo ya funcionando en un ambiente de producción por lo cual es de vital importancia realizar una minuciosa validación del modelo tal como se describe en el paso anterior. Generalmente es necesario actualizar o ajustar el modelo a las necesidades del negocio sean estas regulaciones o ingreso de nuevas variables que se desea se incluyan dentro del modelo por que se considera ayudaran a soportar mejor las decisiones.

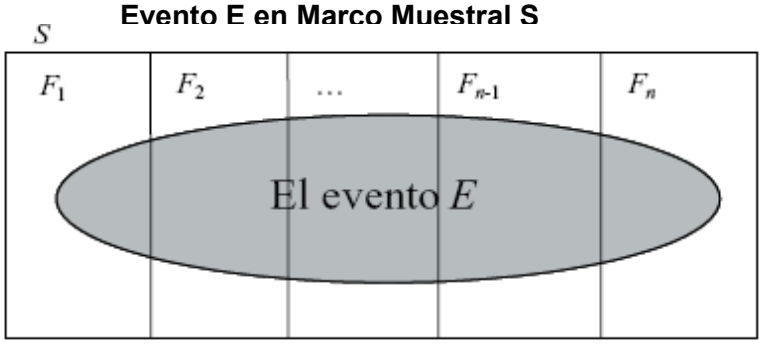
Paso 6.- Administración de los Meta Datos

Es importante almacenar la información del proceso de construcción del modelo como p.e. información sobre columnas transformadas, modelos previos, efectividad de los mismos ya que esta información es muy útil al momento de ajustar o crear un modelo nuevo partiendo de los mismos datos total o parcialmente.

1.5 Clasificación bayesiana

Es un método basado en la teoría de la probabilidad, usa frecuencias para calcular probabilidades condicionales para calcular predicciones sobre nuevos casos. Naive Bayes es una técnica tanto predictiva como descriptiva.

Gráfico 1.3
Evento E en Marco Muestral S



1.5.1 Marco Teórico de Naive Bayes (Bayesiano ingenuo).-

Sean E y F eventos, podemos expresar a E como: E= EF ∪ EF´, es decir, para que ocurra un evento E, deben suceder E y F, o bien suceder E y no suceder F.

Debido a que E y F son mutuamente excluyentes, tenemos que:

P(E)= P(EF) + P(EF´)

= P(E|F)P(F) + P(E|F')P(F')

= P(E|F)P(F) + P(E|F')(1-P(F))

Donde se determina la ponderación relacionada a la probabilidad condicional del evento E dada la ocurrencia de F y la probabilidad condicional de E dado que F no ha ocurrido. Cada probabilidad condicional proporciona tanta ponderación como el evento condicional tiende a ocurrir.

Cuando se usa la clasificación bayesiana?.- El problema o planteamiento más general para utilizar este algoritmo o modelo de minería de datos es la necesidad de separar o clasificar n observaciones pertenecientes a una población definida X en K categorías establecidas mediante la discriminación por los parámetros de entrada a la función de clasificación desarrollada a través de observaciones generalmente denominadas de entrenamiento y validadas con datos de prueba.

La entrada que recibe la función de clasificación es una cantidad P de parámetros o características de cada observación conocida como vector de parámetros. El clasificador entrega una de las K clases establecidas que

corresponda a la mas adecuada respuesta en base a los parámetros ingresados.

Este tipo de clasificador esta basado en la regla de Bayes:

$$p(y_j|x) = \frac{p(x|y_j)p(y_j)}{p(x)}$$

Donde:

$p(y_j|x)$ = probabilidad que ocurra la clase y_j dado el dato x (distribución a posteriori)

$p(x|y_j)$ = probabilidad que el dato x ocurra dado que venga de la clase y_j (distribución a priori)

$p(y_j)$ = probabilidad de ocurrencia de la clase y_j,

$p(x)$ = probabilidad de ocurrencia del dato x

El valor y_j (y = j) dado el valor x se considera que es una variable aleatoria que tiene una distribución condicional $p(y_j|x)$, pero esta distribución tiene valores de $p(y_1|x)$, $p(y_2|x)$ ... $p(y_n|x)$ tanto como posibles resultado tenga la variable y. Entonces la predicción de un valor y_j dado el valor x se lo obtiene seleccionado el índice i que de el mayor

posible para la probabilidad $p(y_j \mid x)$ es decir se selecciona la clase y_j más probable dado un valor de x

$$y_j = f(x) \equiv x \in y_j \equiv p(y_j|x) \geq p(y_j|x)$$

Para todo i , j

Teóricamente dada cualquier distribución conjunta de X y Y en la cual existe una dependencia desconocida entre las variables X y Y, el clasificador bayesiano es el clasificador que va tener el menor error en la predicción de Y dado X. Y=f(X).

El error del clasificador bayesiano viene dado por la expresión:

$$EB = \int (1 - \max(p(y_k|x))p(x))dx$$

Este error viene a ser una cota inferior para los errores de predicción de cualquier clasificador. El clasificador bayesiano se considera entonces el mejor clasificador posible.

Inclusive se ha llegado a demostrar que el clasificador de los k-vecinos más cercanos se aproxima asintóticamente al clasificador bayesiano cuando el número de datos N tiende a

infinito, igual resultado se ha demostrado para las redes neuronales.

Trabajar con clasificadores bayesiano sería lo ideal pero el problema que es difícil estimar adecuadamente las probabilidades a priori ($p(x|y_j)$) a partir solo de los datos de prueba. Esto hace que se tenga que o usar métodos clasificatorios alternativos tales como redes neuronales, árboles de decisión, etc. o asumir ciertos supuestos sobre la distribución a priori que simplifica su estimación. El segundo caso es lo que conoce como clasificador bayesiano simplificado (Naive Bayes). [3]

1.5.2 Clasificador Bayesiano Simplificado.-

Este clasificador asume que los atributos de X (X1, X2, …, Xn) son variables aleatorias independientes por lo tanto se puede estimar la distribución a priori p(x | yj) de la siguiente forma:

$$p(x|y_j) = p(x_1|y_j) * p(x_2|y_j) * \ldots * p(x_n|y_j)$$

Cada uno de los $p(x_j | y_j)$ puede ser estimado por medio de un histograma de los valores x_i para cada clase y_i, y este histograma puede ser fácilmente calculado a partir de los datos de prueba, o definido mediante un procedimiento alternativo de discretización de datos, ya explicado, que se aproxima de forma consistente a la definición de los histogramas pero que es mucho más flexible cuando la cantidad de datos es amplia y la graficación es una alternativa menos flexible. [3]

1.5.3 Aplicación de Clasificador Bayesiano Simplificado.-

El siguiente ejemplo considera que para la aprobación de solicitudes de tarjetas de crédito se consideran únicamente los factores: el Sueldo del solicitante, la Edad del solicitante, si el solicitante posee Título Académico de tercer nivel, Si consta o no con Calificación E (incobrable) en la Central de Riesgos; como factores únicos a revisar previa la aprobación o rechazo de una solicitud, además se posee la información de las últimas 10 solicitudes analizadas descritos en la siguiente tabla:

Tabla 1.1

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

SOLICITUD DE TARJETA DE CRÉDITO

SUELDO	EDAD	TITULADO	Calificación 'E' en Central de riesgos	RESULTADO
1800	40	si	no	si
500	32	no	no	no
2500	68	no	si	si
800	28	no	si	no
1000	27	si	no	si
900	26	si	no	si
3000	56	no	no	si
950	24	no	no	no
1200	26	si	no	si
2000	29	no	no	si

Elaborado por: Miguel A. Chang

que adicionalmente contiene el resultado de la solicitud (aprobada o rechazada), en base a la información descrita como se clasificaría una solicitud con los siguientes datos:

Tabla 1.2

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

SOLICITUD DE TARJETA DE CRÉDITO SIN RESULTADO

SUELDO	EDAD	TITULADO	Calificación 'E' en Central de riesgos	RESULTADO
1200	32	si	No	? =
500	26	no	Si	? =

Elaborado por: Miguel A. Chang

Los datos del ejemplo, procedemos a definir el modelo de estimación de aprobación de solicitudes de crédito siguiendo los siguientes pasos:

Separar la muestra:

Tabla 1.3

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

SOLICITUD DE TARJETA DE CRÉDITO CON RESULTADO

SUELDO	EDAD	TITULADO	Calificación 'E' en Central de riesgos	RESULTADO
500	32	No	No	No
800	28	No	Si	No
950	24	No	No	No
1800	40	Si	No	Si
2500	68	No	Si	Si
1000	27	Si	No	Si
900	26	Si	No	Si
3000	56	No	No	Si
1200	26	Si	No	Si
2000	29	No	No	Si

Elaborado por: Miguel A. Chang

Un paso previo de extrema importancia es la observancia que se debe realizar sobre los datos, antes de comenzar a diseñar el modelo. Generalmente se debe hacer un recorrido y entender los datos, captar de forma inherente como definir el modelo a partir de la información, ayudándonos a lograr un modelo eficiente y confiable.

Discretizar la información en este caso es conveniente únicamente a dos niveles para las variables Sueldo (>=900 y <900) y Edad (>=27 y < 27). Por tanto la información queda de la siguiente forma:

Tabla 1.4

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

SOLICITUD DE TARJETA DE CRÉDITO

SUELDO	Clase Sueldo	EDAD	Clase Edad	TITULADO	Calificación 'E' en Central de riesgos	RESULTADO
500	1	32	2	No	no	no
800	1	28	2	No	si	no
950	2	24	1	No	no	no
1800	2	40	2	Si	no	si
2500	2	68	2	No	si	si
1000	2	27	2	Si	no	si
900	2	26	1	Si	no	si
3000	2	56	2	No	no	si
850	1	32	2	Si	no	si
2000	2	29	2	No	no	si

Elaborado por: Miguel A. Chang

Los resultados posibles se representan como una combinación de las opciones de cada variable, teniendo: 2*2*2*2 = 16, de los cuales diferenciaremos los no repetidos. En la siguiente tabla se incluyen la información a partir de los datos de estimación, los que incluyen para cada categoría definida, las clases y la probabilidad independiente de aprobar para cada una:

Tabla 1.5

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

VALORES A PRIORI

Característica	Categoría	Valor
Sueldo	1	0.1
	2	0.6
Edad	1	0.1
	2	0.6
Posee Título???	Si	0.3333
	No	0.3333
Consta en Central riesgo con 'E'	Si	0.1
	No	0.6

Elaborado por: Miguel A. Chang

Tabla 1.6

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

TABLA DE ENTRENAMIENTO

Priori 1	Priori 2	Priori 3	Priori 4	Resultado	Decisión
(1/10)	(1/10)	(3/10)	(1/10)	0.000333	Negado
(1/10)	(6/10)	(3/10)	(1/10)	0.002	Negado
(1/10)	(1/10)	(3/10)	(1/10)	0.000333	Negado
(1/10)	(1/10)	(3/10)	(6/10)	0.002	Negado
(1/10)	(6/10)	(3/10)	(6/10)	0.012	Aprobado
(6/10)	(1/10)	(3/10)	(1/10)	0.002	Negado
(6/10)	(6/10)	(3/10)	(1/10)	0.012	Aprobado
(6/10)	(1/10)	(3/10)	(6/10)	0.012	Aprobado
(6/10)	(6/10)	(3/10)	(6/10)	0.071999	Aprobado

Elaborado por: Miguel A. Chang

El cálculo de frecuencias a priori se realiza en base a la frecuencia por clase (teoría clásica de probabilidades (resultados observados / resultados totales; para cada clase)), como se había descrito anteriormente, que es el equivalente numérico al método gráfico del histograma. Por ejemplo, la probabilidad a priori de que se apruebe la solicitud dado que

el sueldo del aplicante es <900 (clase 1), es (1/10) = 0.1 puesto que únicamente 1 de los 10 aplicantes, obtuvo la aprobación de la solicitud siendo su sueldo menor que <900 ó de clase 1.

De la tabla anterior valores a priori, podemos inferir cuales son los valores que resultarán favorables a la hora de aplicar una solicitud para cualquier caso que se ajuste a las variables definidas, por tanto procedemos a completar el ejercicio para los datos complementarios:

Tabla 1.7

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

SOLICITUD DE TARJETA DE CRÉDITO CON RESULTADOS CLASIFICADOS

SUELDO	EDAD	TITULADO	Calificación 'E' en Central de riesgos	valor	RESULTADO
1200	32	si	No	0.071999	? = APROBADO
500	26	no	Si	0.000333	? = NEGADO

Elaborado por: Miguel A. Chang

Detallando el ejemplo aplicado y refiriéndonos a los valores por clase, para el individuo 1 tenemos:

Prob = (6/10) * (6/10) * (3/10) * (6/10) = 0.071999, con lo cual se aprueba su solicitud.

Mientras para el individuo 2:

Prob = (1/10) * (1/10) * (3/10) * (1/10) = 0.000333, con lo cual se rechaza su solicitud.

Es necesario indicar que el ejemplo está únicamente a guiar hacia los aspectos generales de la clasificación, y del uso y aplicación del clasificador bayesiano; la aplicación de Naive Bayes o Bayes simplificado implica técnicas de carácter objetivo y se basa fuertemente en la teoría de la probabilidad siendo este el factor que a la vez de ser una ventaja para clasificar casos tipo, es una seria desventaja para los datos atípicos o que no se ajustan a un patrón normal de resultado.

CAPITULO II

DESCRIPCIÓN DEL PROBLEMA Y PROPUESTA DE SOLUCIÓN

INTRODUCCIÓN

Presentamos a continuación la descripción detallada del problema de negocio generalizada para ESPOL y puntualizada para el ICM, identificando los puntos que se desean corregir y/o soportar para luego definir la propuesta de solución a nivel de detalle, delineando las definiciones del trabajo a desarrollarse con ese fin.

2.1 Descripción del problema y propuesta de solución

Abordar un problema implica un análisis detallado del mismo, contar con las herramientas necesarias y el planteamiento de la solución optima. El problema en este caso es estimar la probabilidad de que un estudiante apruebe una materia como punto general de partida. En base al enunciado del problema definimos el enunciado de la solución como sigue: "*Diseñar un sistema semi-automatizado que permita realizar la estimación de que un estudiante apruebe una materia en base a varios factores a los que llamaremos factores predictores*".

Antes de completar y detallar la definición del problema, es importante revisar el escenario de desarrollo del proceso que se desea soportar, de manera general el bosquejo de cómo funciona dicho proceso y las debilidades del mismo en cuanto a lo que se desea solucionar.

2.2 Planificación académica

2.2.1 Introducción

Dentro de las funciones de los directivos de las unidades académicas de la ESPOL se encuentra la elaboración de la planificación académica que constituye en breves rasgos: "el establecimiento de las materias y cantidad de paralelos a dictarse

durante el termino académico siguiente inmediato, de acuerdo al ordenado flujo de las materias en los programas de las carreras y teniendo en cuenta además la demanda estimada para cada una de las mismas". El procedimiento es un detalle puntualizado del proceso general de planificación académica institucional, que no entra en detalles puntuales de planificación y por lo tanto no será referenciado en el presente estudio.

Teniendo en cuenta la definición de planificación académica descrita anteriormente podemos establecer un punto que dará utilidad y sentido a la realización del presente trabajo. Si bien tenemos una actividad que han enfrentado los directivos durante mucho tiempo que es tratar de satisfacer la demanda en base a ¿información?, ¿experiencia?, ¿buen criterio?; encontramos una gran debilidad que debe ser cubierta y es el uso de la información disponible en las grandes bases de datos disponibles, por lo tanto nos centraremos en la forma en como los directivos del ICM en la ESPOL determinan la planificación académica que siempre será un proceso estimativo y de gran complejidad al tener cada vez una mas numerosa población estudiantil.

"La planificación académica es de naturaleza dinámica, es un plan que en su etapa de formación puede sufrir muchos cambios, antes

de que sea aprobada, en su ‘versión definitiva’, por la Comisión Académica. Por su carácter dinámico y de cambios que en la mayoría de los casos de carácter urgente, ya que los directivos encargados de hacerla disponen de poco tiempo para su elaboración, necesita del soporte adecuado informático para poder aumentar la efectividad y eficiencia.”

2.2.2 Requisitos de planificación académica en la ESPOL

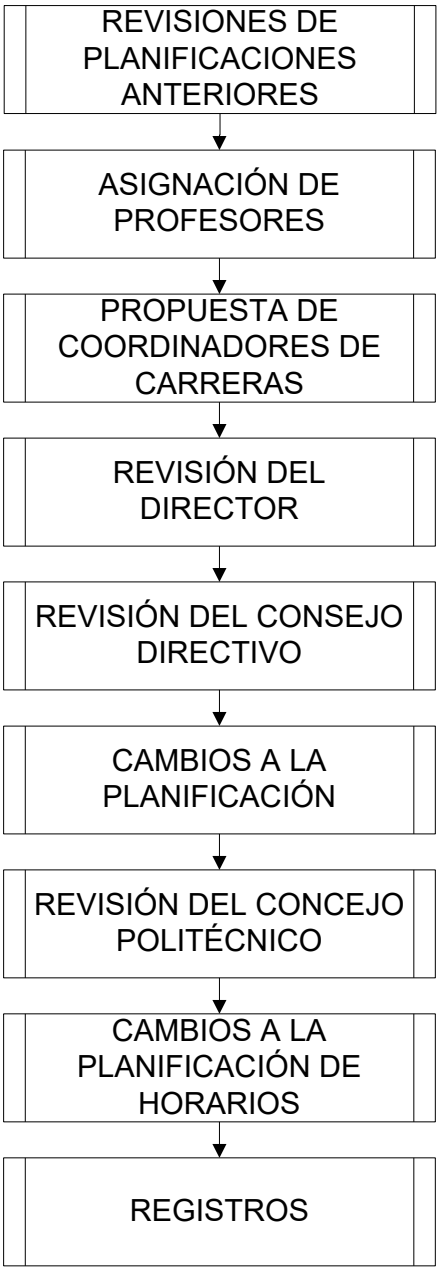
Los eficacia varía de planificación en planificación eso es seguro pero el producto de varios días de trabajo será tener una planificación precisa que cubra los requerimientos de los clientes internos, los estudiantes, y que reciba el mínimo de modificaciones durante todo el proceso desde su formulación inicial hasta su aprobación final, o en términos generales, que se ajuste a la realidad anticipada del nuevo ciclo académico.

Tener disponible la información de términos académicos anteriores es una de los guías mas comunes a seguir y hacer una copia del mismo para aplicarlo al presente término que se avecina es la solución más rápida y no siempre mas efectiva que se dispone. Centrarnos en la información será el objetivo para proporcionar a las cabezas de las diferentes unidades académicas una guía, una herramienta y un soporte al momento de elaborar su planificación de una mejor forma.

2.2.3 El proceso de Planificación académica

Se describe en la siguiente figura el flujo de la planificación académica de la ESPOL.

Gráfico 2.1
Estructuración de un Sistema Informático Gerencial



Elaborado por: Miguel A. Chang

El presente trabajo agregaría una actividad inicial al proceso descrito anteriormente, la estimación de alumnos que estarían aptos para tomar determinada materia que cambiaría de cierta forma el arranque de esta planificación dejando de ser la misma como una copia de la programación del periodo anterior, y tener como información de entrada los valores esperados de aprobación y reprobación por materia para cada una de las mismas que dicta la institución.

2.2.4 Problemas generales en la planificación académica en la ESPOL

Una gran cantidad de cambios se realizan al documento inicial de planificación de materias desde su creación o versión 0 hasta que este se implementa en los registros de los estudiantes previo al inicio de un periodo académico. Si tenemos disponible la información ¿por que no usarla?, cubrir los requerimientos de los estudiantes, migrar hacia la toma de decisiones estocástica y no determinística, establecer lo que percibimos con fundamentos y sobre todo darle un trato adecuado al cliente que bajo una metodología que soporte criticas y evite cierres o aperturas de paralelos será muy adecuado teniendo en cuenta que nos encontramos en la “era de la información”.

2.2.5 Problemas específicos en la planificación académica en el ICM-ESPOL

Adicionalmente a lo ya expuesto sobre ESPOL, se acumulan los factores internos de cada unidad, y en este caso del ICM se presentan de forma general: primero la complejidad de dar una respuesta a los requerimientos de sus clientes, donde ya el tema de confirmación de profesores se vuelve difícil y complicado por la premura con la que se deben manejar estas actividades y por los muchos datos pero poca información con la que se cuenta al momento de tomar decisiones, segundo el no saber cuantos estudiantes espero tener o listos para registrarse en una materia específica. Factores adicionales que inciden en el ICM y que dejan menor cantidad de tiempo a las tareas de planificación son: administrar el dictado de materias para otras unidades en cuanto a asignación de aulas y profesores, cuadre de horarios, selección de ayudantes, etc.

2.2.6 Soporte informático a la planificación académica

Teniendo en cuenta que el proceso de planificación académica es: El Centro de Servicios Informáticos (CSI) de la ESPOL provee a las diferentes unidades académicas de un sistema llamado: “Sistema Académico” encontrándose operativa en la actualidad la versión 2.0. En el ICM se ha desarrollado un sistema que cubre y administra varias etapas de

planificación académica y que no son cubiertas por el sistema provisto por el CSI.

Adicional al Sistema Académico de la ESPOL, existe un software adicional de soporte a la planificación desarrollado por el Ing. Vicente Jama se encarga de administrar horarios, profesores y carga politécnica de profesores con nombramiento, entre las más importantes de sus funciones; pero no cubre el punto de partida, la estimación de demanda, que consideramos un punto importante a considerar dentro de la planificación. Por lo que no se tiene un aplicativo que ayude a realizar dicha estimación de demanda requerida como una entrada importante al proceso de planificación académica.

A continuación describimos un poco la herramienta actual:

2.2.6.1 Sistema Académico de la ESPOL

El Sistema Académico se puso en producción en el año 1999. El sistema comprende todo lo relacionado al manejo de información académica de pregrado.

Sus funcionalidades son:

- Registros estudiantes

- Migración de la información del sistema de admisiones con los estudiantes que aprobaron el proceso de admisión
- Planificación de cursos
- Control de los flujos académicos por estudiante
- Control de pre-requisitos, co-requisitos y materias de arrastre
- Pre-registro y control de pagos de estudiantes
- Matrícula y Registro de estudiantes
- Registros estudiantes a través del web (proyecto piloto en octubre de 2004)
- Convalidaciones y equivalencias
- Emisión de listas de asistencia

- Control financiero
  - Cálculo del valor del registro para los estudiantes
  - Control de descuentos y exoneraciones
  - Control de valores cancelados en el Banco del Pacífico y la tesorería de la ESPOL
  - Manejo de deudas del estudiantes
  - Control de saldos a favor del estudiante
- Control de ingreso de calificaciones y faltas
  - Manejo de períodos por semestres, bimestres

- Ingreso de Calificaciones, rectificaciones a través del Web
- Ingreso de estudiantes que pierden materias por falta a través del web
- Padrones electorales y control de estudiantes que no votaron para aplicar las sanciones establecidas en el reglamento

- Consultas a través del web
  - Historia académica
  - Calificaciones del semestre
  - Deudas
  - Impresión del recibo de pago por concepto de deudas y del registro
  - Actualización de datos personales
- Interfases de información para:
  - Kiosko electrónico
  - Sistema de biblioteca
  - CEPROEM
  - Emisión de carnets
  - Consultas a través del Sistema de Audiorespuesta
  - Software para emitir reportes y obtener información de la base de datos (utilizado exclusivamente por el CRECE)

- CENACAD, evaluación del personal docente por parte de los estudiantes

2.3 Contexto especifico del Problema de Negocio

En el siguiente capítulo se cubrirá la etapa de creación un clasificador y su implementación, según las indicaciones generales expuestas en 2.3 y las específicas detalladas en el capítulo 3. Se desarrollaran de manera puntual las actividades de selección variables de clasificación, se analizara la factibilidad y conveniencia de realizar una reducción al espacio característico con las variables seleccionadas como idóneas para la creación del estimador y luego se desarrollara la metodología descrita en el capitulo 1 para la creación de un modelo de minería de datos.

2.4 Propuesta de la Solución

La propuesta de solución implica el desarrollo de un clasificador mediante la utilización de Visual Basic 6.0 y como motor de Base de Datos MS SQL Server 2000. La propuesta por lo tanto se delinea de forma macro según lo que se ha venido definiendo a lo largo del documento y pretende: "*Diseñar un sistema semi automatizado que permita realizar la estimación de que un estudiante apruebe una materia en base a varios factores a los que llamaremos factores*

predoctores”, el hecho de tener una solución semi-automática y no una completamente automatizada responde a que el modelo no es inmutable en el tiempo, siendo todo lo contrario, un esquema de afinamiento y perfeccionamiento estimativo que debe ser utilizado sin perder la esencia de su creación, como un soporte de información en base a estimación por criterios.

El clasificador a breves rasgos podrá:

- Proveer información confiable sobre la demanda esperada del siguiente periodo académico.
- Valorar en base a las características base, si un estudiante aprueba o reprueba una materia.

CAPITULO III

IMPLEMENTACIÓN Y RESULTADOS

INTRODUCCIÓN

El presente capítulo desarrolla todos los conceptos aplicables, descritos en el Capítulo I, a fin de lograr sobre la metodología generalizada de minería de datos expuesta, la construcción del modelo de minería y su implementación a través de un sistema informático probado y validado sobre datos reales del negocio con un alto nivel del eficacia o poder estimado.

3.1. Desarrollo de la solución

Se desarrollará la metodología descrita en el capítulo I para ir diseñando la solución descrita en el capítulo II, siguiendo de la siguiente forma:

3.1.1. Definición específica del Problema de minería de datos

Ya definido el Problema de negocio en el capítulo 2, iniciamos con la definición del proceso de Problema de minería, ya que en el presente trabajo vamos a obtener un modelo que permitirá estimar la probabilidad de que un determinado estudiante apruebe o no una materia tomada dentro de un ciclo académico normal basándonos en la información propia de dicho estudiante y de sus registros académicos en la ESPOL.

Las variables de entrada del modelo son variables obtenidas de los registros académicos y de identificación del estudiante en los sistemas de la ESPOL, datos proporcionados en su totalidad por el CSI, Centro de Servicios Informáticos, de la ESPOL.

La técnica a utilizar en el presente estudio es el análisis discriminante de dos grupos utilizado el clasificador Naive Bayes, donde como se ha revisado anteriormente se pretende estimar si

un estudiante determinado aprueba o reprueba una determinada materia en base a varios aspectos relevantes del estudiante.

El método de clasificación utilizado corresponde al Clasificador bayesiano ingenuo. Los datos utilizados para el presente análisis corresponden al periodo 2000 – 2004 primer y segundo término, divididos en los siguientes grupos:

Como grupo de entrenamiento: 2000 – 2003 (37750 registros)

Como grupo de pruebas: 2004 (5608 registros)

3.1.2. Diagrama de flujo de la solución

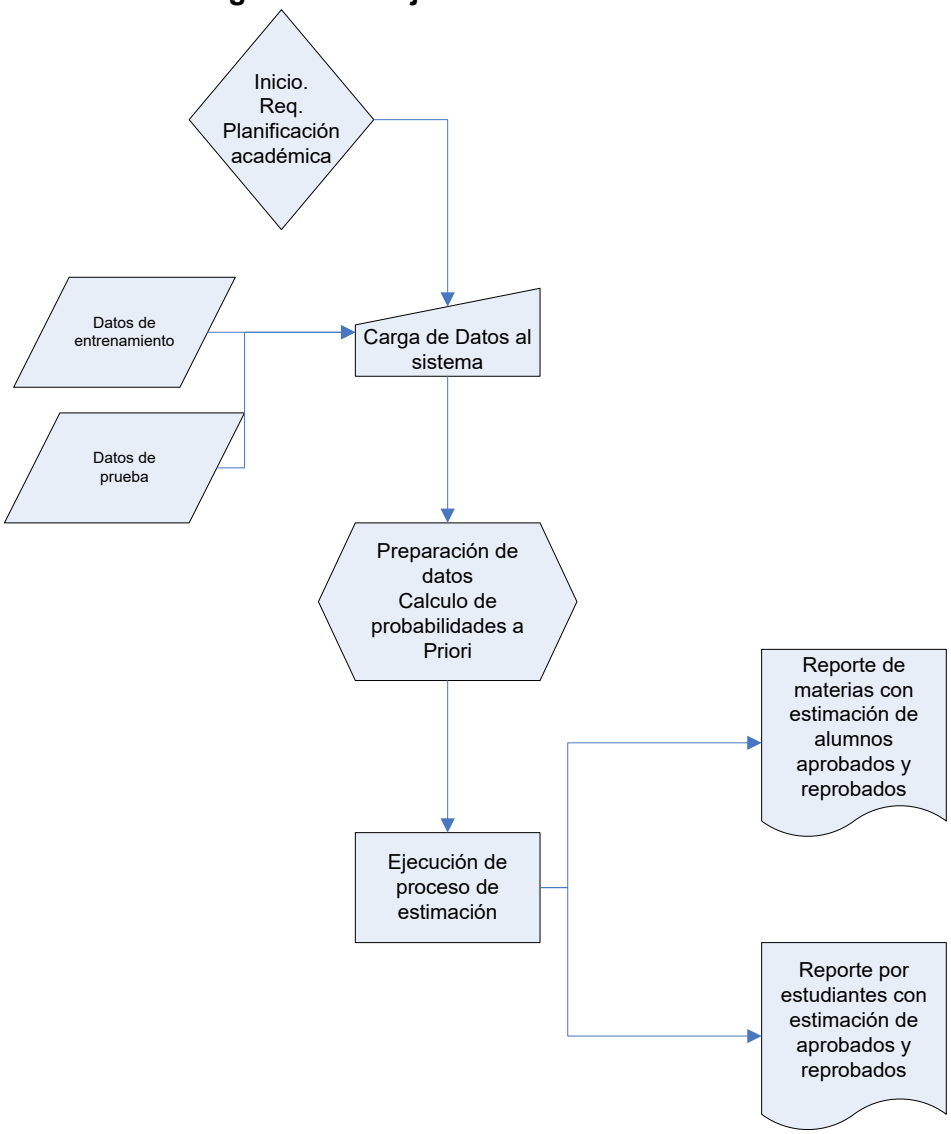
El esquema de solución presentado ya en el documento incluye:

1. Se requiere para la planificación académica la información
2. Se ingresan los datos de entrenamiento y prueba al sistema
    - Datos de entrenamiento.- Información de los registros inmediatos anteriores al periodo que se desea estimar, se considera adecuado utilizar al menos 3 años de información por cada semestre que se desee estimar

- Datos de prueba.- Información del semestre que se desea estimar

3. Preparación de la base de datos, incluye las tareas de completar las tablas de predicción (matriz) y de resultados (semestre), llega hasta completar la información de datos a priori
4. Se ejecutan las tareas de estimación sobre los datos de predicción
5. Se generan los reportes para uso de la información del sistema

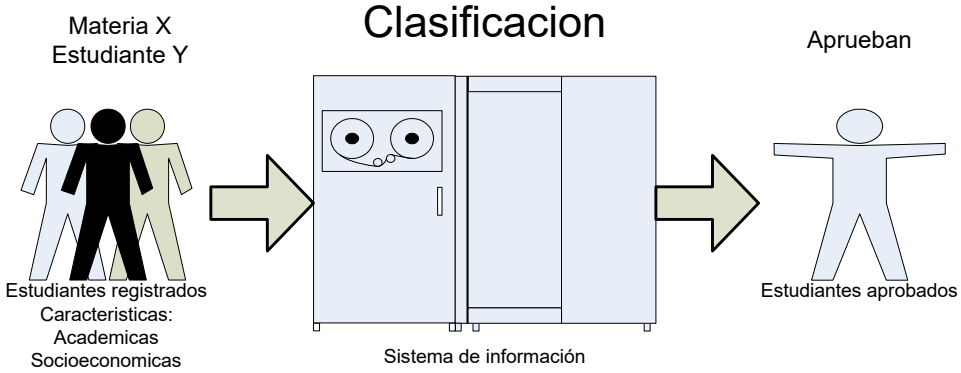
Gráfico 3.1
Diagrama de Flujo de la Solución



Elaborado por: Miguel A. Chang

3.1.3. Definición de Variables de Clasificación

Gráfico 3.2
Esquema General de Clasificación



Elaborado por: Miguel A. Chang

Grupo inicial de variables que se esperaba incluir:

Inicialmente se consideraron los siguientes factores:

1. Factor socioeconómico del estudiante
2. Promedio general de materias aprobadas
3. Genero
4. Nivel de eficiencia
5. Nota del Primer Parcial
6. Horario de la Materia
7. Profesor de la Materia
8. Cantidad de estudiantes por paralelo
9. Carrera del estudiante

Resultado:

Las variables descritas en los literal 6-9 no se consideran relevantes ni de valor a efectos del estudio por tanto sin un mayor análisis serán obviadas y no incluidas al modelos. Las características definitivas a ser incorporadas, las cuales se utilizaron para la construcción del mismo son:

1.- Factor socioeconómico del estudiante

2.- Promedio general de materias aprobadas

3.- Genero

4.- Nivel de eficiencia (materias aprobadas/materias tomadas)

5.- Nota del primer parcial

El resultado se basa principalmente en la *no disponibilidad* de forma consistente para algunos de los factores considerados inicialmente para ser incluidos en el modelo, dado que podrían generar desviaciones o que su peso se considera sería relativamente bajo al discriminar si el estudiante aprueba o reprueba.

3.1.4. Obtención de los datos

Los datos en su totalidad fueron obtenidos de la base de datos del Sistema Académico de la ESPOL. Se considera lo anterior la mejor estrategia ya que evita para el mantenimiento del presente sistema la estimación de la datos requeridos y mas bien el proceso implicaría varias tareas que se descritas en los próximos capítulos que permitirán el uso sostenible del sistema a corto y mediano plazo, teniendo claro que a largo plazo deberán hacerse revisiones al mismo y ajustarlo en base a la realidad.

3.2. Preparando los datos

Una vez definida la base de datos y establecido de forma clara los objetivos que desean cubrirse con el presente trabajo se debe hacer la identificación de los datos requeridos para el análisis, preparar los datos de entrada al modelo de minería de datos pasando por las etapas de selección de la data, y la limpieza de los datos o 'Data Cleaning'.

3.2.1. Selección de los datos

La obtención de los datos para realizar el presente análisis corresponde a la información almacenada en la base de datos del Sistema Académico de la ESPOL para los años 2000 hasta 2004.

Los requerimientos de información señalados corresponden a la totalidad de la información disponible sobre:

Listado de alumnos del ICM e información adicional de los mismos.

Los registros para los alumnos del ICM por término académico desde 2000 hasta 2004.

Información disponible de profesores.

Información disponible de materias.

Se describen en el **Anexo 1** las definiciones de las tablas del modelo listadas a continuación:

Estudiantes.- Estudiantes que se registraron en el ICM en el periodo de estudio 2000 1S – 2004 2s. Para dichos estudiantes implica que en cualquiera de los periodos académicos por lo menos tomaron una materia en cualquiera de las carreras del ICM.

Profesores.- Constan todos los profesores de la ESPOL.

Materias.- Detalla las materias dictadas en la ESPOL identificadas por código.

Semestre.- Toda la información almacenada en la tabla estudiantes y materias para el año 2004, también se definió a este grupo de datos como datos de prueba, con los datos

requeridos para completar el modelo, incluyendo las notas del primer parcial de los estudiantes.

Registros.- Contiene los datos de entrenamiento del sistema, incluye la información de los registros de estudiantes del ICM del año 2000 al 2003.

Matriz.- Tabla base de la cual se obtendrán los valores resultantes del modelo, conocidos como datos a priori, los cuales permitirán completar el mismo.

La matriz siguiente permite visualizar la correlación existente entre las variables que se incluyan al modelo y que demuestra la independencia estadística de las mismas.

Tabla 3.1

"Creación e implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

MATRIZ DE CORRELACIONES DE LAS VARIABLES DEL ANÁLISIS

	GÉNERO	EFICIENCIA	NOTA PARCIAL	FACTOR P
EFICIENCIA	-0,194			
NOTA PARCIAL	-0,07	0,334		
FACTOR P	0,126	-0,002	0,008	
PROMEDIO	-0,162	0,619	0,3	-0,019

Elaborado por: Miguel A. Chang

3.2.2. Limpieza de los datos (Data cleaning)

El data clearing es una actividad muy importante en el proceso de minería de datos, dado que permitirá identificar los datos fuera de

rango y los casos atípicos que potencialmente podrían afectar al modelo. En el **Anexo 2** se incluyen las actividades de limpieza de datos realizadas sobre la data del presente estudio.

3.2.3. Transformación de los datos

Dado que se trabajó de una fuente de datos homogénea no se requirió realizar transformaciones de fondo a los datos obtenidos.

Es necesario señalar en este punto del análisis que para obtener mejores resultados de clasificación utilizando el modelo definido, la mejor estrategia es discretizar los datos. Las estrategias de discretización de datos se realizó sobre los datos de las tablas matriz y semestre, correspondientes a datos de entrenamiento y de prueba previamente definidos, sobre las columnas: Factor p, Promedio, Eficiencia, Nota Parcial; el método de discretización utilizado fue la Discretización simple, la cual extrae los valores máximo y mínimo del conjunto de datos y define las clases dividiendo la diferencia anterior para un k determinado.

$$\delta = \frac{x_{\max} - x_{\min}}{k}$$

La discretización entonces fue realizada para los datos con el fin de obtener mejores resultados utilizándose para todas las variables un k=3.

3.3. Construyendo el modelo

3.3.1. Análisis exploratorio de los datos

Se describe en el **Anexo 3** el análisis univariado de cada una de las variables de clasificación a ser consideradas en el proceso de construcción del clasificador descrito en el próximo paso.

3.3.2. Creación de grupos de datos necesarios

Es importante como se realizará a continuación, redefinir los datos de entrenamiento del clasificador y los datos de prueba sobre los cuales se realizarán, tema que ya fue abordado en el documento y que se formalizará en este capítulo:

3.3.2.1. Datos de entrenamiento

Los datos de entrenamiento del modelo comprenden la información obtenida y ajustada para todos los semestres del 2000 al 2003, sobre los cuales se desarrolla y construye el modelo. La cantidad de registros disponibles como datos de entrenamiento son 37750 que corresponden al 87% de datos como datos de entrenamiento.

3.3.2.2. Datos de prueba

Los datos de prueba comprenden todos los registros obtenidos y ajustados del año 2004. Independientemente de estos datos de prueba el modelo una vez ajustado podrá hacer inferencias sobre cualquier conjunto de datos de entrada que sigan el patrón de estimación y se ajusten al formato establecido y según las indicaciones para usos futuros del software definidas en el **Anexo 4**. La cantidad de registros disponibles como datos de prueba son 5608 que corresponden al 13% de datos como datos de prueba.

3.3.2.3. Construcción del Modelo

La construcción del modelo se define de la siguiente forma:

$$p(x|y_j) = p(x1|y_j) * p(x2|y_j) * \ldots * p(xn|y_j)$$

Por lo que se define de forma específica que:

P(Aprobar **|** Factor P,Género,Promedio,Eficiencia,Nota parcial**)=**

P (Factor P | Aprobó) * P (Género | Aprobó) * P (Promedio | Aprobó) * P (Eficiencia | Aprobó) * P (Nota parcial | Aprobó)

Donde se calculó a partir de los datos de prueba las probabilidades a priori de:

P (Factor P | Aprobó) = Probabilidad a priori de que dado un factor p específico aprobar.

P (Género | Aprobó) = Probabilidad a priori de que dado un género aprobar.

P (Promedio | Aprobó) = Probabilidad a priori de que dado un promedio específico aprobar.

P (Eficiencia | Aprobó) = Probabilidad a priori de que dado un nivel de eficiencia específico aprobar.

P (Nota parcial | Aprobó) = Probabilidad a priori de que dada una nota parcial aprobar.

El modelo contempla, almacenar los valores a priori en una tabla y en base a la discretización ya realizada sobre los campos a estimar obtener las clases y asignar los valores respectivos con fines de lograr eficiencia en la obtención de resultados.

3.4. Validación del modelo

3.4.1. Resultados de clasificación

La aplicación del modelo de clasificación se aplica a la totalidad de registros del ICM del año 2004, en total 5608 registros son filtrados por nuestro estudio, los cuales presentan de forma

razonable un ajuste a la realidad de los resultados. De los 5608 registros, que implica que de todos los estudiantes del ICM que se registraron para ese año en al menos una materia, 3844 resultaron según el software como aprobados y 1764 como reprobados. Los resultados reales arrojan las siguientes cifras: 4140 aprobados, 1468 reprobados; con lo cual se concluirá sobre la efectividad de clasificación del modelo en la siguiente sección.

La ***tasa de clasificación errada*** es la probabilidad de que la regla de clasificación (o simplemente el clasificador) clasifique mal una observación proveniente de una muestra obtenida posteriormente a la muestra usada para establecer el clasificador siendo que incluya o no registros de la muestra de entrenamiento. Existen varios métodos para estimar esta tasa de error de clasificación, los cuales se mencionan únicamente para referencia puesto que la eficiencia del modelo que se muestra en la siguiente sección se realizó sobre la totalidad de los datos de prueba (registros 2004: semestres 1, 2 y 3); el método más popular se detalla a continuación:

Estimación por resustitusión o Error Aparente.- Este es simplemente la proporción de observaciones de la muestra que son erróneamente clasificadas por el modelo. Por lo general un estimador demasiado optimista puede conducir a falsas

conclusiones si el tamaño de la muestra no es muy grande comparado con el número de variables predictoras.

3.4.2. Prueba de precisión versus la data real

Dado que se tienen todos los resultados del periodo sobre el cual se realizó una comparación dando al modelo una precisión del 92.85% **(3844 (aprobados según sistema) / 4140 (aprobados en realidad))** de confiabilidad de resultados.

El análisis de precisión de clasificación se realizó comparando las estimaciones realizadas del sistema contra el resultado real obtenido para cada estudiante del ICM que tomó una materia en el año 2004, el **Anexo 5** muestra los resultados tabulados a nivel de materias.

Por lo tanto podemos asignar al clasificador una ***tasa de clasificación errada*** del 7.15%, que implica que de cada 100 registros clasificados aproximadamente 7 serán erróneos. Teniendo en cuenta que en promedio se tienen algo menos de 6000 registros en el ICM por semestre, se esperarían tener 429 registros erróneos de alrededor de 94 materias y 1200 alumnos por semestre.

3.5. Despliegue del modelo

3.5.1. Creación del modelo de despliegue

El modelo de despliegue se definirá bajo requerimientos del ICM, en cuanto al interés que se demuestre para la utilización del sistema, el Manual de uso del sistema incluye todas las tareas necesarias para la utilización del mismo en el tiempo. Por tanto será una definición que escapa al presente análisis la implementación del presente sistema en el ICM.

3.5.2. Evaluar el modelo en el ambiente de producción

La implementación implica la creación del programa final mediante el cual se hará uso del sistema, en el manual de usuario **Anexo 6** se muestran las principales pantallas y sus utilidades, así como también las opciones de reporte del mismo.

Se debe recalcar que la eficiencia en diseño implica la fácil adaptabilidad del modelo sobre la incorporación de nuevas variables de clasificación siempre y cuando las mismas cumplan los requisitos de independencia y completitud requeridos. Esto es que al momento de evaluar agregar una variable al modelo, esta sea estadísticamente independiente de las anteriores y además

se obtenga toda la información requerida en los datos de prueba y en los datos a estimar.

CAPITULO IV

CONCLUSIONES Y RECOMENDACIONES

INTRODUCCIÓN

En este capítulo se presentan las conclusiones de mayor valor identificadas en la realización del estudio realizado con el objetivo de ayudar en el desarrollo de trabajos posteriores de esta naturaleza y describir algunas interrogantes del alcance de este trabajo que pueden no estar descritas en el contenido del mismo, las recomendaciones en parte al igual que las conclusiones pretenden dar una guía estudio de futuros trabajos en esta área.

4.1. Conclusiones

1. Es importante definir de forma clara el alcance del trabajo y los objetivos del mismo, para validar el camino seguido se deben revisar las actividades para asegurar que las mismas están contribuyendo de manera positiva al logro del objetivo final trazado.

2. El nivel de complejidad aumenta en gran medida si no se posee una guía o trabajos anteriores realizados en el área para tener como referencia. Realizar un trabajo estructurado es fundamental por lo se debe contar con un modelo metodológico a seguir.

3. Cada técnica de minería de datos a aplicarse debe ser revisada y validada ya que las técnicas multivariadas tienen varios prerrequisitos que los datos deben cumplir y se deben revisar previa la aplicación de la técnica de análisis de datos.

4. Naive Bayes trabaja sorprendentemente bien (aun si la asunción de independencia es violada claramente) - Por qué? Porque la clasificación no requiere estimados de probabilidad exacta mientras que la probabilidad máxima es asignada a la clase correcta.

5. El modelo puede albergar un mayor número de variables, dada la flexibilidad del mismo, se deben siempre tener en cuenta no inluir factores redundantes o factores que no agreguen ganancia al poder clasificatorio del mismo.

6. El modelo no se considera completamente terminado, en realidad es posible que variables 'importantes' no hayan sido incluidas al mismo, es necesaria se haga de forma periódica una revisión al mismo y se determine su suficiencia o los requerimientos de ajuste necesarios. La mejora continua, como en muchos campos prácticos, se considera un factor indispensable también en la minería de datos y es necesario considerarla en ese sentido.

7. Cualquier mejora o cambio mayor al sistema puede ser considerado para futuros trabajos, ya sea como proyectos en materias relacionadas ó como un nuevo trabajo de grado.

8. El proceso de planificación académica de la ESPOL está definido a nivel macro y no detalla de forma precisa la actividad que supone mejoramos con el presente estudio, por lo tanto no es comparable el nivel de mejora que pudiera lograrse al implementar el sistema desarrollado. Al no existir confrontaciones entre la forma en la que se

lleva la tarea actualmente con la sugerida es posible y recomendable acoger al sistema y profundizar y habilitar la cultura de planificación sobre información en la ESPOL.

4.2 Recomendaciones

1. Las tareas de Minería de Datos tienen gran aplicabilidad y amplio uso, se debe crear un departamento interno a nivel de ESPOL donde se defina un equipo multidisciplinario de trabajo al cual se le provean inicialmente los recursos para que luego tenga la posibilidad de brindar servicios y pueda autofinanciarse.

2. Se incluye en el **Anexo 4**. El uso del sistema a futuro para el soporte de la toma de decisiones en el ICM dependerá de la persona que deba realizarlo, la guía descrita pretende ser el soporte para que se mantenga el uso del sistema.

3. Es importante se de continuidad a la investigación y desarrollo de trabajos en este campo, existen instituciones internacionales que promueven el desarrollo del área y que podrían dar un nivel amplio de asesoramiento que ayude a mantener la tendencia y a producir material de calidad en Business Intelligence.

4. La Inteligencia de Negocio es un área que está siendo considerada en varias empresas a nivel local, dar a los profesionales del ICM y de la ESPOL un perfil sólido para ocupar

esas plazas de trabajo podría ser un objetivo de mediano plazo institucional ajustándonos a las exigencias del mercado.

BIBLIOGRAFIA

[1] **Han, J., & Kamber, M.** (2000); "Data Mining : Concepts and Techniques", San Francisco, CA: Morgan Kaufmann

[2] **Rencher, A.** (2002), "Methods of Multivariate Analysis – Second Edition", John Wiley & Sons, New York - USA

[3] **Alvarado Ortega, Juan.** (2003), "Algoritmos de la Minería de Datos", ***Revista Matemática***, Una Publicación del ICM-ESPOL, ESPOL Vol.1 No 2., Guayaquil-Ecuador

[4] **Acuña Fernández, Edgar.** (2000), "Notas de Análisis Discriminante", Departamento de Matemáticas - Universidad de Puerto Rico en Mayagüez

[5] **Palmer, Alfonso.** ET AL. (2001), "Minería de datos en Economía", Universidad de les Illes Balears. 07122 Palma de Mallorca

[6] **Vázquez, Salvador & Martínez, Antonio.** (2003), "Una Arquitectura para el análisis automático de bases de datos", Comunicaciones en Socioeconomía, Estadística e Informática. Vol. 7 No. 2. pg. 89-106

[7] **SQL Data Mining, Microsoft:**
www.microsoft.com/spain/sql/productinfo/datasheet/DataMining.mspx

[8] **Sistemas de Información. Base de Datos**.
http://html.sistemas-de-informacion_bases-de-datos.html

ANEXOS

ANEXO I
DEFINICIÓN DE TABLAS DE BASE DE DATOS

Entre los objetos más importantes podemos encontrar las tablas con sus respectivos campos que conforman nuestra Base de Datos:

Tabla 1

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

ESTUDIANTES

Descripción: Tabla que contiene la información de los estudiantes que tomaron al menos una materia en los años 2000 - 2004		**Autor de creación:** Miguel Angel Chang Aguilar		**Tabla:** 1/1
Nombre del Campo	**Tipo de datos**	**Long.**	**Descripción**	**Null**
est_codigo	Float	9	Número de matrícula del estudiante	Not Null
est_apellidos	varchar	50	Apellidos del estudiante	Not Null
est_nombres	varchar	50	Nombres del estudiantes	Not Null
est_genero	float	8	Género del estudiante	Not Null
est_fecha_nac	smalldatetime	4	Fecha de nacimiento del estudiante	Not Null
est_factor_p	float	8	Factor p del estudiante	Not Null
est_mat_tom	float	8	Materias tomadas por el estudiante	Not Null

Tabla 2

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

MATERIAS

Descripción: Tabla que contiene la información de las materias que en las que se han registrado los estudiantes	**Autor de creación:** Miguel Angel Chang Aguilar	**Tabla:** 1/2

Nombre del Campo	**Tipo de datos**	**Long.**	**Descripción**	**Null**
mat_codigo	nvarchar	255	Código de materia	Not Null
mat_nombre	nvarchar	255	Nombre de materia	Not Null
mat_cred_teor	float	8	Créditos de horas teóricas	Not Null
mat_cred_prac	float	8	Créditos de horas practicas	Not Null

Tabla 3

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

PROFESORES

Descripción: Tabla de que contiene la información de los profesores	**Autor de creación:** Miguel Angel Chang Aguilar	**Tabla:** 1/3

Nombre del Campo	**Tipo de datos**	**Long.**	**Descripción**	**Null**
prof_cedula	nvarchar	255	Cédula del profesor	Not Null
prof_apellidos	nvarchar	255	Apellidos del profesor	Not Null
prof_nombres	nvarchar	255	Nombres del profesor	Not Null

Tabla 4

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

REGISTROS

Descripción: Tabla que describe registros por materia de un estudiante	**Autor de creación:** Miguel Angel Chang Aguilar	**Tabla:** 1/4

Nombre del Campo	Tipo de datos	Long.	Descripción	Null
reg_anio	Numeric	9(18,0)	Año de registro	Not Null
reg_term	nvarchar	255	Ternito de registro	Not Null
reg_cod_carrera	nvarchar	255	Código de carrera del estudiante	Not Null
reg_nombre_carrera	nvarchar	255	Nombre de la carrera del estudiante	Not Null
reg_matricula	float	8	Matrícula del estudante	Not Null
reg_cod_materia	nvarchar	255	Código de la materia en que se registra el estudiante	Not Null
reg_nota_parcial	float	8	Nota parcial de la materia registrada	Not Null
reg_nota_final	float	8	Nota final de la materia registrada	Not Null
reg_nota_mej	float	8	Nota mejoramiento de la materia registrada	Not Null
reg_paralelo	int	4	Paralelo de la materia registrada	Not Null
reg_profesor_id	nvarchar	255	Identificación de profesor que dicta la materia	Not Null

Tabla 5

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

SEMESTRE

Descripción:			**Autor de creación:** Miguel Angel Chang Aguilar	**Tabla:** 1/5
Nombre del Campo	**Tipo de datos**	**Long.**	**Descripción**	**Null**
sem_id	numeric	9(18,0)	Identificación de vehículos	Not Null
Sem_anio	float	8	Describe las placas del vehículo	Not Null
Sem_termino	nvarchar	50	Describe el nombre del responsable del vehículo	Not Null
Sem_cod_mater	nvarchar	50	Describe alguna particularidad del vehículo	Not Null
Sem_cod_prof	nvarchar	50	Código del profesor	Not Null
Sem_est_cod	float	8	Código del estudiante	Not Null
Sem_est_fp	float	8	Factor P del estudiante	Not Null
Sem_est_prom	float	8	Promedio del estudiante	Not Null
Sem_est_genero	float	8	Género del estudiante	Not Null
Sem_est_eficiencia	float	8	Eficiencia del estudiante	Not Null
Sem_est_nota_p	numeric	9(18,0)	Nota parcial del estudiante	Not Null
Sem_clase_fp	numeric	9(18,0)	Clase de Factor P	Not Null
Sem_clase_prom	numeric	9(18,0)	Clase de Promedio	Not Null
Sem_clase_np	numeric	9(18,0)	Clase de Nota parcial	Not Null
Sem_prob_aprob_nb	nvarchar	53	Probabilidad de aprobar	Not Null

Tabla 6

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

MATRIZ

Descripción:		**Autor de creación:** Miguel Angel Chang Aguilar		**Tabla:** 1/6
Nombre del Campo	**Tipo de datos**	**Long.**	**Descripción**	**Null**
mat_id	numeric	9(18,0)	Id de la tabla	Not Null
mat_fp	numeric	9(18,0)	Factor P del estudiante	Not Null
mat_prom	float	8	Promedio del estudiante	Not Null
mat_cod_mat	float	8	Código de materia	Not Null
mat_cod_prof	nvarchar	255	Código de Profesor	Not Null
mat_cant_mat_tom	numeric	9(18,0)	Cantidad de materias tomadas	Not Null
mat_genero	int	4	Género del estudiante	Not Null
mat_eficiencia	float	8	Eficiencia del estudiante	Not Null
Mat_nota_parcial	float	8	Nota parcial	Not Null
Mat_est_cod	numeric	9(18,0)	Código del estudiante	Not Null
Mat_decision	int	4	1= si estudiante aprobó la materia, 0= si estudiante reprobó la materia	Not Null
Mat_ter	nvarchar	50	Termino academizo	Not Null
Mat_anio	numeric	9(18,0)	Año académico	Not Null
Mat_clase_fp	int	4	Clase de Factor P	Not Null
Mat_clase_prom	int	4	Clase de Promedio	Not Null

Mat_clase_eficiencia	int	4	Clase de Eficiencia	Not Null
Mat_clase_nota_parcial	int	4	Clase de Nota Parcial	Not Null

Tabla 7

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

PRIORI

Descripción:		**Autor de creación:** Miguel Angel Chang Aguilar		**Tabla:** 1/7
Nombre del Campo	**Tipo de datos**	**Long.**	**Descripción**	**Null**
id	Numeric	9(18,0)	Identificador de la tabla	Not Null
priori_caract	Numeric	9(18,0)	Característica	Not Null
priori_cat	Numeric	9(18,0)	Categoría de la característica	Not Null
priori_val	nvarchar	53	Valor por característica y categoría	Null

Tabla 8

"Creación e Implementación de un clasificador suave para estimar la aprobación de materias de los estudiantes del Instituto de Ciencias Matemáticas de la ESPOL"

REPORTE_MATERIAS

Descripción: Tabla que describe los clientes de la empresa		**Autor de creación:** Miguel Angel Chang Aguilar		**Tabla:** 1/8
Nombre del Campo	**Tipo de datos**	**Long.**	**Descripción**	Not Null
Id	int	4	Identificador	Not Null
rep_semestre	char	10	Semestre de estudio	Not Null
rep_termino	char	10	Termino de estudio	Not Null
rep_cod_mat	nvarchar	50	Código de materia	Not Null

rep_nom_mat	nvarchar	50	Nombre de materia	Not Null
rep_est_apr	numeric	9(18,0)	Cantidad de estudiantes aprobados	Not Null
rep_est_rep	numeric	9	Cantidad de estudiantes reprobados	Not Null

ANEXO II

ACTIVIDADES DE LIMPIEZA DE DATOS REALIZADAS

Fue necesario realizar varias actividades de limpieza de datos: primero antes de cargar los mismos a la base de datos y segundo, actividades desarrolladas internamente en la base de datos por medio de procedimiento almacenados propios de SQL – Server, para completar los pasos requeridos para finalizar el sistema, dichas actividades serán descritas por grupo de datos de forma general con una breve descripción de la tarea realizada.

Se presenta el listado de los datos antes de ingresar al sistema, para cada ítem de la lista se comentará las actividades de limpieza realizadas, para el grupo de datos que no se comente ninguna actividad de limpieza, pues debe asumirse que los mismos fueron cargados al sistema sin cambio alguno.

ACTIVIDADES DE LIMPIEZA SOBRE LOS DATOS ANTES DE CARGARLOS A LA BASE DE DATOS:

Datos obtenidos:

Listado de estudiantes del ICM - ESPOL.

Listado de materias de la ESPOL.

Listado de profesores de la ESPOL.

Listado de los registros del ICM 2000 – 2003.

Listado de los registros del ICM 2004.

Actividades de limpieza:

Sobre Listado de estudiantes del ICM – ESPOL se realizaron las siguientes actividades de limpieza:

- Se incorporó una columna que afectó a todos los registros y se incluyó el género del estudiante, pues esta información no la contiene el Sistema Académico de la ESPOL
- No se realizó ninguna actividad adicional sobre esta tabla

Sobre Listado de Materias se realizaron las siguientes actividades de limpieza:

- Se procedió únicamente a importar los datos al sistema

Sobre Listado de Profesores se realizaron las siguientes actividades de limpieza:

- Se procedió únicamente a importar los datos al sistema

Sobre Listado de registros del ICM 2000 – 2003 se realizaron las siguientes actividades de limpieza:

- Dado que se nos proporcionó la información por año, se procedió a unificar los registros de los años 2000, 2001, 2002 y 2003; utilizando como control de esta unificación los totales por año contra el total final del archivo

- No se realizó ninguna actividad adicional sobre esta tabla

Sobre Listado de registros del ICM 2004 se realizaron las siguientes actividades de limpieza:

- Se procedió únicamente a importar los datos al sistema

ACTIVIDADES DE LIMPIEZA SOBRE LOS DATOS YA INGRESADOS A LA BASE DE DATOS:

Se incluye en este grupo, el llenado de tablas base, describiendo los procedimientos almacenados que permiten completar esta actividad.

Llenado de la tabla matriz:

El llenado de la tabla matriz, o tabla principal del sistema, la cual contiene los datos de entrenamiento del sistema incluye tomar toda la información de registros 2000 – 2003 y cruzar esta información filtrando únicamente los registros pertenecientes a estudiantes del ICM y de los cuales se podrá extraer información útil para entrenar el clasificador.

Adicionalmente a la actividad descrita, se discretizan las variables del modelo a excepción de género, la cual no requiere discretización, a fines de lograr mayor eficiencia en el modelo.

Se utilizan para las dos tareas descritas, los procedimientos almacenados: llena_matriz y discetizar1 respectivamente.

Llenado de la tabla semestre:

El llenado de la tabla semestre, o tabla de prueba del sistema, la cual contiene los datos de prueba del sistema incluye tomar toda la información de registros 2004 y cruzar esta información filtrando únicamente los registros pertenecientes a estudiantes del ICM y de los cuales se podrá extraer información útil para probar el clasificador.

Adicionalmente a la actividad descrita, se discretizan las variables del modelo a excepción de género, la cual no requiere discretización, a fines de lograr mayor eficiencia en el modelo, tal como se realizó para tabla matriz.

Se utilizan para las dos tareas descritas, los procedimientos almacenados: llena_semestre y discetizar2 respectivamente.

Llenado de la tabla priori:

El llenado de la tabla priori, o tabla de probabilidades de clases, la cual contiene las probabilidades por clases para cada variable del modelo, son los valores que para cada corrida se toman y se define la decisión de aprobar o reprobar.

El cálculo de los valores a priori se realiza mediante el procedimiento almacenado datos_a_priori.

ANEXO III

ANÁLISIS UNIVARIADO DE LAS VARIABLES DEL MODELO

Análisis Univariado de los datos de los registros desde el año 2000 – 2003 de los cuales se calcularon las probabilidades a priori del modelo

Factor P	
Media	9.035444
Error Estándar	0.022974
Mediana	8
Moda	7
Desviación Estándar	4.463683
Varianza de la Muestra	19.92446
Rango	33
Mínimo	3
Máximo	36

Eficiencia	
Media	0.752212
Error Estándar	0.000901
Mediana	0.770492
Moda	1
Desviación Estándar	0.175141
Varianza de la Muestra	0.030675
Rango	0.933333
Mínimo	0.066667
Máximo	1

Nota Parcial	
Media	62.02726
Error Estándar	0.102742
Mediana	64
Moda	60
Desviación Estándar	19.96219
Varianza de la Muestra	398.4892
Rango	100
Mínimo	0
Máximo	100

Promedio	
Media	7.104903
Error Estándar	0.002063
Mediana	7.04
Moda	7.07
Desviación Estándar	0.400881
Varianza de la Muestra	0.160705
Rango	3.2
Mínimo	6
Máximo	9.2

Genero	
Media	0.387709
Error Estándar	0.002508
Mediana	0
Moda	0
Desviación Estándar	0.487234
Varianza de la Muestra	0.237397
Rango	1
Mínimo	0
Máximo	1

Análisis Univariado de la información del período de estudio que se desea estimar el cuál fue al año lectivo 2004 - 2005

Genero	
Media	0.347896
Error Estándar	0.006361
Mediana	0
Moda	0
Desviación Estándar	0.476345
Varianza de la Muestra	0.226905
Rango	1
Mínimo	0
Máximo	1

Factor P	
Media	8.44918
Error Estándar	0.055069
Mediana	7
Moda	7
Desviación Estándar	4.123939
Varianza de la Muestra	17.00687
Rango	33
Mínimo	3
Máximo	36

Eficiencia	
Media	0.736057
Error Estándar	0.002572
Mediana	0.764706
Moda	1
Desviación Estándar	0.192633
Varianza de la Muestra	0.037107
Rango	0.933333
Mínimo	0.066667
Máximo	1

Nota Parcial	
Media	61.14961
Error Estándar	0.310221
Mediana	65
Moda	0
Desviación Estándar	23.23137
Varianza de la Muestra	539.6964
Rango	100
Mínimo	0
Máximo	100

Promedio	
Media	7.198878
Error Estándar	0.00634
Mediana	7.14
Moda	7.07
Desviación Estándar	0.474746
Varianza de la Muestra	0.225384
Rango	3.43
Mínimo	6
Máximo	9.43

ANEXO IV
MANUAL DE USO DEL SISTEMA

En el presente manual de uso del sistema, se procederá a explicar como debe instalarse el sistema, y a detallar los pasos para hacerlo funcionar.

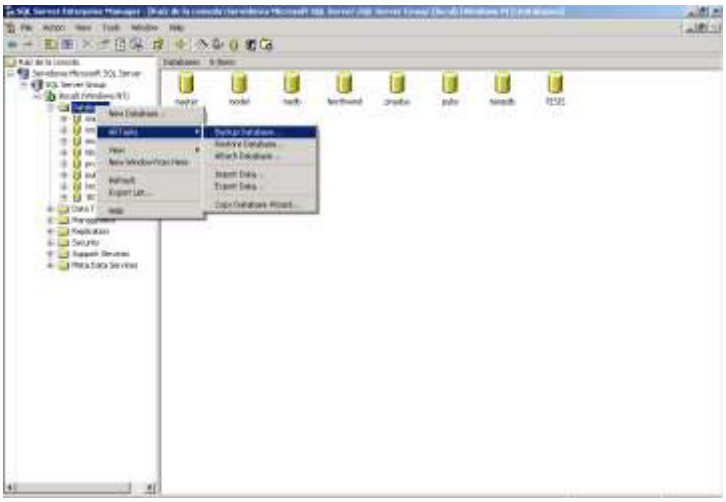
El manual consta de 5 secciones principales, las cuales se van a ir detallando a continuación:

Instalación de la Base de Datos

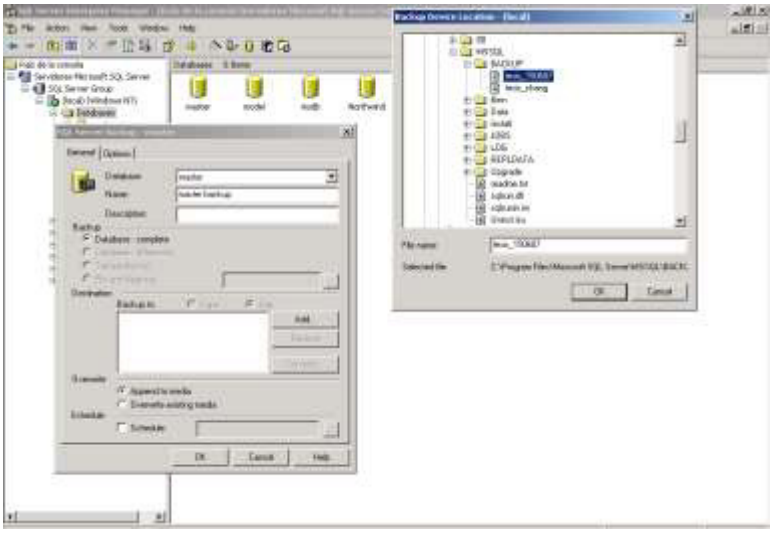
Para la instalación de la base de datos se deben seguir los siguientes pasos:

1. Copiar archivo Base.bck al directorio: C:\Program Files\Microsoft SQL Server\MSSQL\BACKUP
2. Abrir el Administrador Corporativo de MS SQL Server 2000
3. Restaurar la base de datos siguiendo los siguientes pasos:

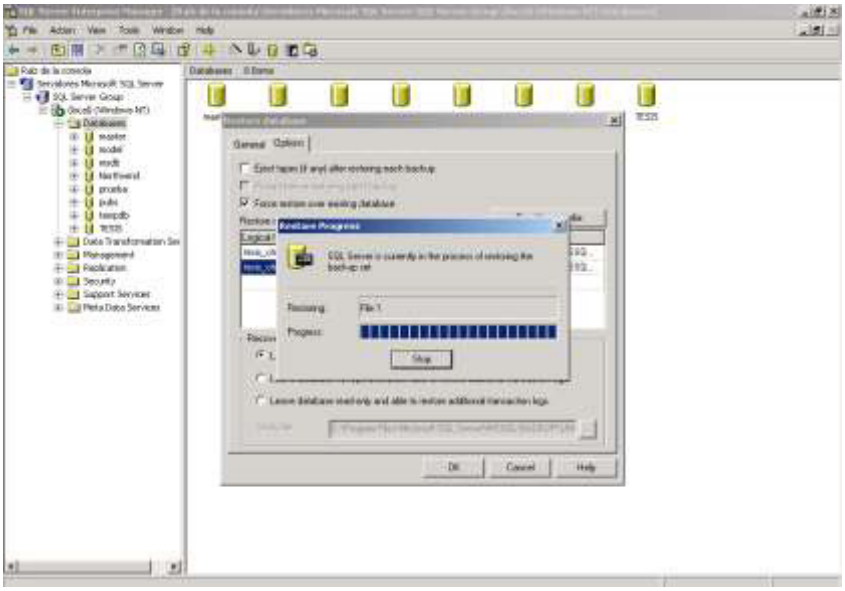
a.



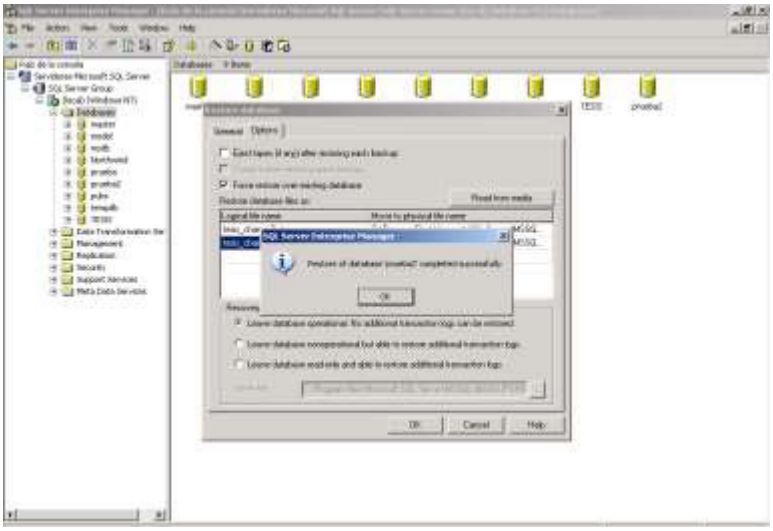
b.



c.



d.



Con esas actividades damos por concluida la instalación de la base de datos del sistema.

Instalación del Aplicativo

Para la instalación del aplicativo se deben seguir los siguientes pasos:

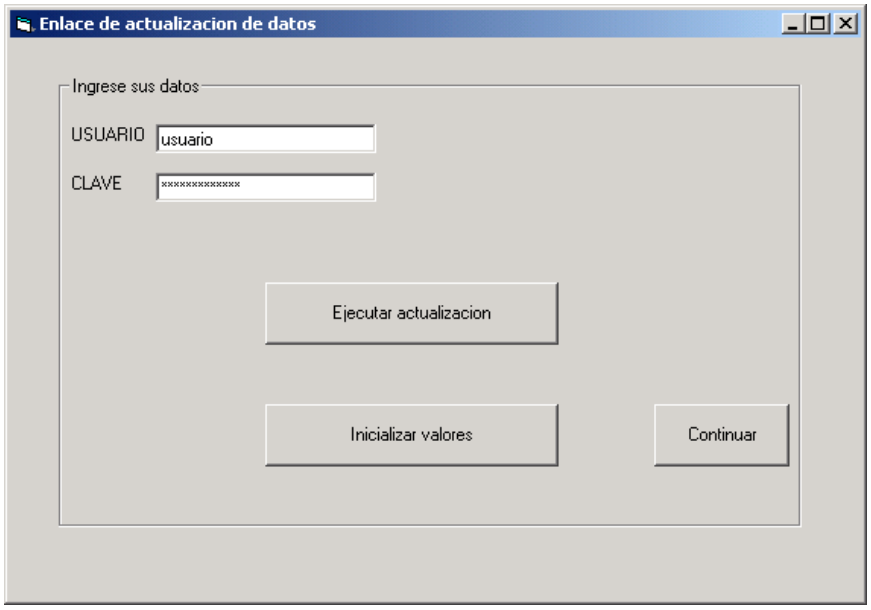
4. Copiar la carpeta Sistema_tesis al directorio: C:\DOCS\Tesis_Estimación
5. Ejecutar el archivo Tesis.vbp
6. Ingresar el usuario y clave que indica el sistema
7. Seguir las instrucciones del Anexo 6.- Manual de usuario del sistema

Carga de datos

La carga de datos debe ser realizada de forma sistemática, una vez se hayan preparado los datos acorde se detalla en el Anexo 2, se deben copiar los

archivos en: C:\DOCS\Tesis_Estimación, directorio de donde los leerá el sistema.

Una vez ingresado el usuario y la clave del sistema, se debe ejecutar el comando ejecutar actualización y la carga de datos se realizará de forma automática.



Preparación de datos

Al tener los datos cargados, únicamente ejecutamos el comando inicializar valores, la consulta se ejecutará el aproximadamente 20 minutos por lo tanto se debe esperar que la misma se complete y concluya.

Pantallas y reportes

Ver Anexo vi.

ANEXO V
RESULTADOS DE LA CLASIFICACIÓN
vs.
DATOS REALES

			APROBADOS		REPROBADOS	
CODIGO DATOS REALES	**CODIGO ESTIMACION**	**NOMBRE MATERIA**	**REAL**	**CLASIFICADOR**	**REAL**	**CLASIFICADOR**
ICHE00877	ICHE00877	TEC.EXP.ORAL ESC. E INV.(B)	126	106	4	24
ICM01354	ICM01354	FORM.EVAL. PROYECTOS	121	118	17	20
ICM00604	ICM00604	ALGEBRA LINEAL (B)	109	110	95	94
ICHE02998	ICHE02998	CIENCIA E INVESTIGACION	107	93	23	37
ICM01412	ICM01412	UTILITARIOS INFORMATICOS	96	93	13	16
ICHE02519	ICHE02519	ORGANIZACION Y METODOS	90	72	3	21
ICM00794	ICM00794	FUND.COMPUTACION(B)	89	88	75	76
ICHE02857	ICHE02857	FUNDAMENTOS DE ADMINISTRACI ŧ N	88	49	8	47
ICHE00604	ICHE00604	MACROECONOMIA	88	78	8	18
ICHE00927	ICHE00927	MARCO LEGAL EMP.	85	66	1	20
ICM01453	ICM01453	ETICA Y FUNDAM. DE AUDITOR ^ A	84	79	29	34
ICM01495	ICM01495	MATEM ŧ TICAS III (AUDIT.)	81	80	7	8
ICM01362	ICM01362	INGENIERIA DE LA CALIDAD	79	79	2	2
ICHE02576	ICHE02576	MONEDA Y BANCA	78	68	1	11
ICHE02477	ICHE02477	CONTAB.COSTOS	78	72	17	23
ICHE00885	ICHE00885	ECOL.EDUC.AMB.(B)	78	77	13	14
ICM01545	ICM01545	AUDITOR ^ A OPERACIONAL	75	72	1	4
ICM01263	ICM01263	MATEMATICAS FINANCIERAS	75	92	61	44
ICHE02865	ICHE02865	FUNDAMENTOS DE CONTABILIDAD	74	50	33	57
ICHE03038	ICHE03038	CONTABILIDAD GRAL. II (AUDIT.)	73	72	20	21
ICM01479	ICM01479	MATEM ŧ TICAS I (AUDIT.)	73	83	98	88
ICHE02592	ICHE02592	INV. DE MERCADO	71	68	0	3
ICHE00893	ICHE00893	MICROECONOMIA	69	74	14	9
ICHE03046	ICHE03046	CONTAB.SIST.BANCARIO (AUDIT.)	68	69	10	9
ICM01537	ICM01537	AUDITOR ^ A ADMINISTRATIVA	68	69	9	8
ICM01560	ICM01560	AUDIT. SIST. INFORMATICO I	65	64	1	2
ICHE02980	ICHE02980	CONTABILIDAD GRAL. I (AUDIT.)	63	53	21	31

			APROBADOS		REPROBADOS	
CODIGO DATOS REALES	**CODIGO ESTIMACION**	**NOMBRE MATERIA**	**REAL**	**CLASIFICADOR**	**REAL**	**CLASIFICADOR**
ICM01552	ICM01552	ESTADISTICA II (AUDIT)	61	63	26	24
ICM01511	ICM01511	ESTAD ^ STICA I	61	82	80	59
FIEC05272	FIEC05272	INTRODUC. A BASES DE DATOS(AUDIT)	59	59	9	9
ICM00901	ICM00901	MATEMATICAS DISCRETAS(IEC)	56	58	97	95
ICM01487	ICM01487	MATEM ƒ TICAS II (AUDIT.)	54	65	75	64
ICM00216	ICM00216	CALCULO I (B)	53	37	55	71
ICHE02600	ICHE02600	POLITICA EMPRESARIAL	52	43	0	9
ICM01834	ICM01834	GEOESTAD ^ STICA APLICADA (electiva prof.)	52	47	2	7
ICM01578	ICM01578	MUESTREO Y SIMULACI ₹ N (AUDIT.)	51	52	5	4
ICM01933	ICM01933	CONSULTOR ^ A PROFESIONAL	48	46	1	3
ICHE03079	ICHE03079	CONTABILIDAD GUBERNAMENTAL (AUDIT.)	47	46	0	1
ICM01503	ICM01503	MATEM ƒ TICAS IV (AUDIT.)	47	72	58	33
ICM01420	ICM01420	PROCESOS ESTOC ƒ STICOS	44	29	4	19
ICM01602	ICM01602	AUDIT. SIST. INFORM ƒ T. II (AUDIT.)	42	43	7	6
ICM01321	ICM01321	SIMULACION MATEMATICAS	41	34	0	7
ICM01677	ICM01677	AUDITOR ^ A GUBERNAMENTAL (AUDIT.)	41	40	0	1
ICM01818	ICM01818	ESTRATEGIAS MEJORAM.CALIDAD (ELECTV.PROF	40	39	7	8
ICHE03087	ICHE03087	ADM. RECURSOS HUMANOS (AUDIT.)	40	40	0	0
ICM01636	ICM01636	DERECHO MERCANTIL Y LABORAL (AUDIT.)	40	40	0	0
ICHE02469	ICHE02469	CONTABILIDAD GRAL.	39	31	30	38
ICM01313	ICM01313	ESTADISTICA COMPUTAC.	39	32	0	7
ICM01842	ICM01842	AUDITOR ^ A Y CONTABILIDAD FORENSE	39	39	0	0
ICM01610	ICM01610	AUDITOR ^ A FINANCIERA (AUDIT.)	38	39	3	2
ICM01651	ICM01651	ADMINISTRACI ₹ N P ν BLICA (AUDIT.)	37	37	0	0
ICHE03053	ICHE03053	ADMINISTRACI ₹ N FINANCIERA (AUDIT.)	36	36	0	0
ICM01669	ICM01669	LEGISLACI ₹ N Y PR ƒ CT. TRIBUTARIA (AUDIT.)	36	36	0	0

			APROBADOS		REPROBADOS	
CODIGO DATOS REALES	**CODIGO ESTIMACION**	**NOMBRE MATERIA**	**REAL**	**CLASIFICADOR**	**REAL**	**CLASIFICADOR**
ICM01289	ICM01289	ANLS.MULTIVARIADO Y DISENO EXPERIMENTO	34	29	7	12
ICHE03061	ICHE03061	ADMINISTRACI ŧ N PRESUPUESTARIA (AUDIT.)	34	32	0	2
ICM01271	ICM01271	MATEMATICAS ACTUARIALES	33	29	26	30
ICM01057	ICM01057	TRATAM.ESTAD.DATOS	31	22	32	41
FIEC05330	FIEC05330	ADMINISTRAC.CENTRO C ŧ MPUT. (AUDIT.)	31	27	1	5
ICM00802	ICM00802	MATEMATICAS SUPERIORES	30	21	7	16
ICM00646	ICM00646	CALCULO II (B)	29	27	16	18
FIEC04820	FIEC04820	REDES COMPUT.(IIT95)	27	22	7	12
ICM01347	ICM01347	ANLS.SERIES TIEMPO	26	22	3	7
ICM01297	ICM01297	DESARROLLO APLICAC.COMPUTACIONALES	24	14	1	11
ICM01248	ICM01248	MUESTREO	24	15	9	18
ICM00653	ICM00653	CALCULO III (B)	24	16	14	22
ICM01214	ICM01214	ANALIS.VARIAB.REAL	23	16	12	19
FIEC04630	FIEC04630	SIST.BASES DE DATOS	22	18	3	7
CELEX00067	CELEX00067	INGL / S B ŦSICO A	22	21	11	12
ICHE02550	ICHE02550	MARKETING	20	18	2	4
ICM01123	ICM01123	ESTADIST.MATEMAT.I	19	14	9	14
FIEC04622	FIEC04622	PROGRAMAC.ORIENTADA OBJETOS	18	12	5	11
ICM01586	ICM01586	ADMINISTRACI ŧ N OPERATIVA (AUDIT.)	17	15	1	3
ICM01172	ICM01172	INVESTIGAC.OPERAC.I	15	12	14	17
FIMP06072	FIMP06072	SIST.GESTION AMBIENTAL(IM)	15	14	0	1
ICM00158	ICM00158	ANALISIS NUMERICO (F)	15	15	26	26
ICM01743	ICM01743	DEMOGRAF ^ A (ELECTV.PROFES.I) ING.ESTADIS	14	11	1	4
ICHE00448	ICHE00448	ADM. DE EMPRESAS	12	12	1	1
ICM01164	ICM01164	ESTADIST.MATEMAT.II	11	7	13	17
ICM01255	ICM01255	INVESTIG.OPERAC.II	11	9	11	13
CELEX00075	CELEX00075	INGL / S B ŦSICO B	5	5	0	0

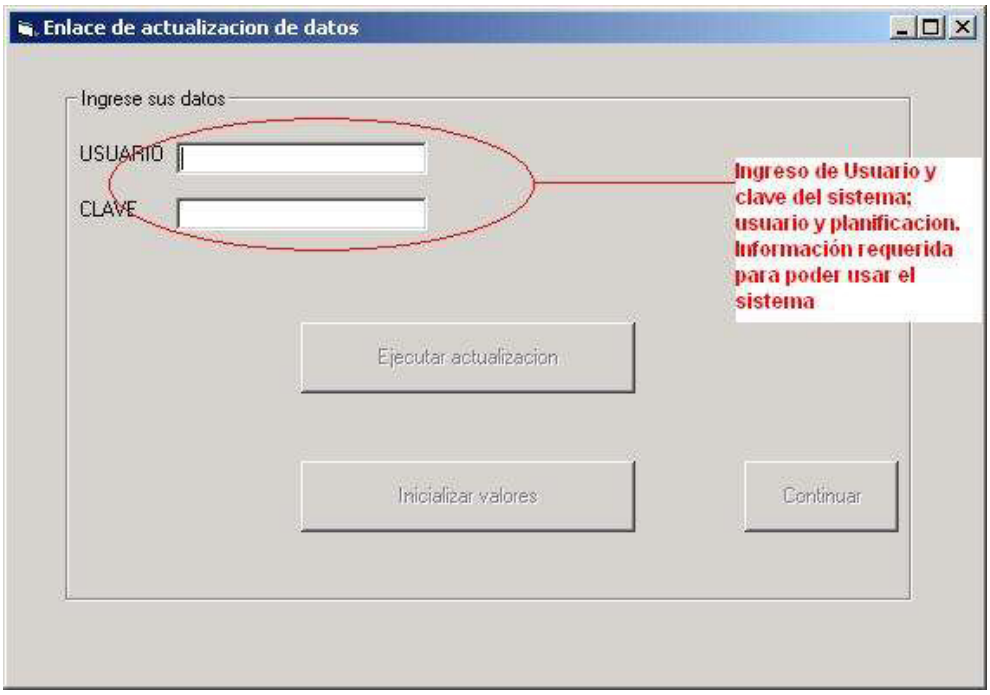
			APROBADOS		REPROBADOS	
CODIGO DATOS REALES	**CODIGO ESTIMACION**	**NOMBRE MATERIA**	**REAL**	**CLASIFICADOR**	**REAL**	**CLASIFICADOR**
ICHE01602	ICHE01602	FINANZ.INTERNAC.	4	4	0	0
ICHE02667	ICHE02667	INTRODUCC. A LA MICROECONOM ^ A	2	2	0	0
ICHE00612	ICHE00612	INGENIERIA ECONOMICA	1	0	0	1
ICHE02105	ICHE02105	SOCIOECON.ANLS.ECONOMICO	1	0	0	1
FIMP07401	FIMP07401	INVESTIG. OPERAC. SIST. MANUFACTURA	1	1	0	0
ICHE01438	ICHE01438	MICROECONOMIA I	1	1	0	0
ICHE02758	ICHE02758	ANALISIS E INVEST.MERCADO	1	1	0	0
CELEX03467	CELEX03467	SUFICIENCIA DE INGL / S 8	0	0	89	89
FIEC05231	FIEC05231	PROGRAMAS UTILITARIOS I	0	0	5	5
FIEC05264	FIEC05264	MODULOS UTILITARIOS INFORMAT.BASICOS	0	0	31	31
ICM01701	ICM01701	MODULO FUNDAM. DE DERECHO (AUDIT.)	0	0	62	62
			4140	3844	1468	1764

presición del clasificador	
REAL EFECT	0,928502415

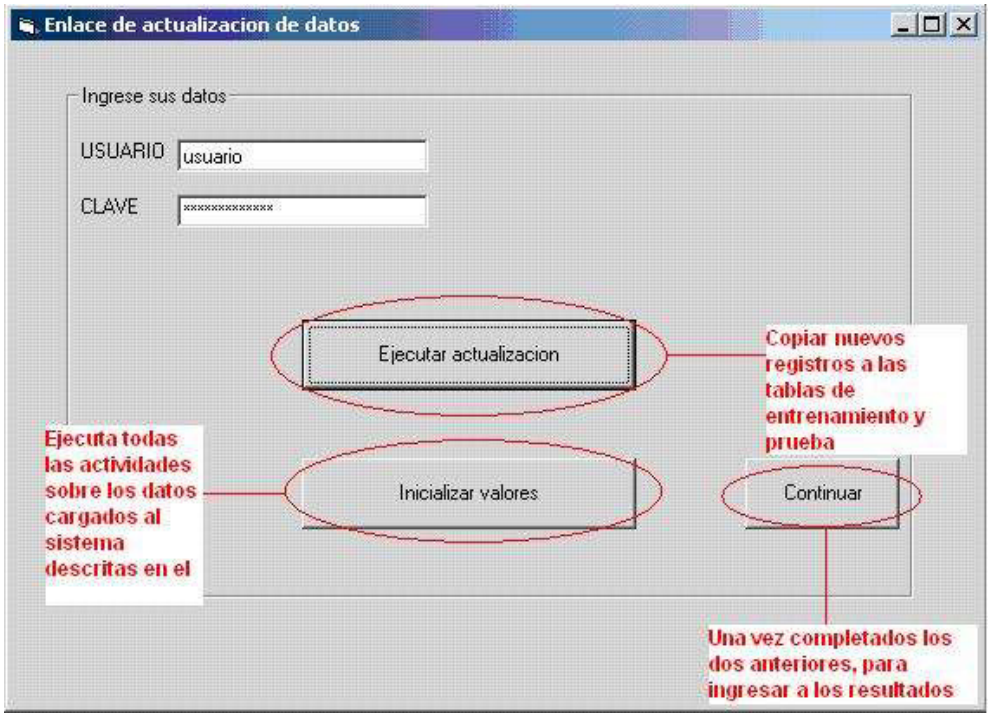
92,85%

ANEXO VI
MANUAL DE USUARIO

En el presente manual se procederá a explicar cada una de las pantallas que integran la aplicación **ESTIMACIÓN DE ESTUDIANTES APROBADOS**, diseñada para estimar la aprobación de estudiantes en una determinada materia.

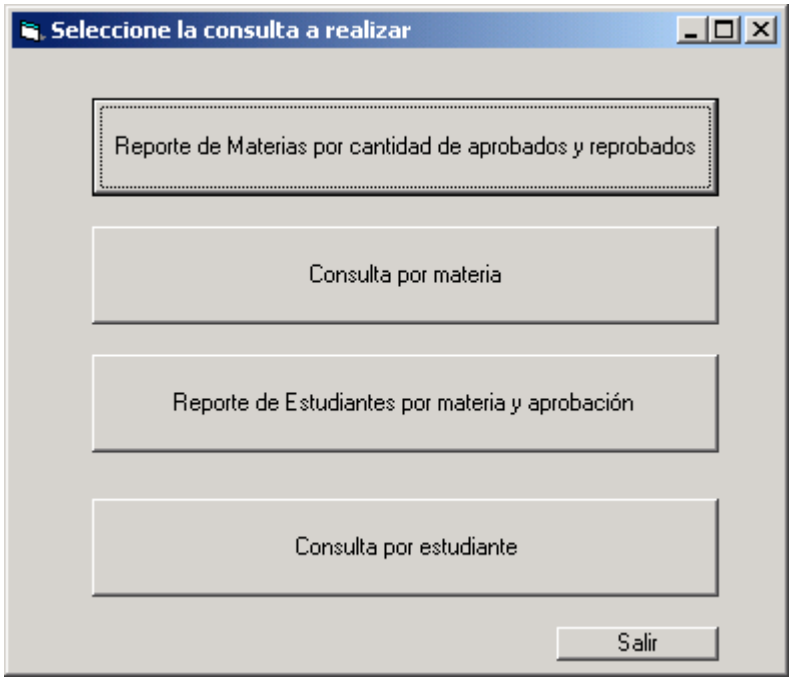


Al ingresar los valores de usuario y clave, se activaran los botones adicionales de la ventana.



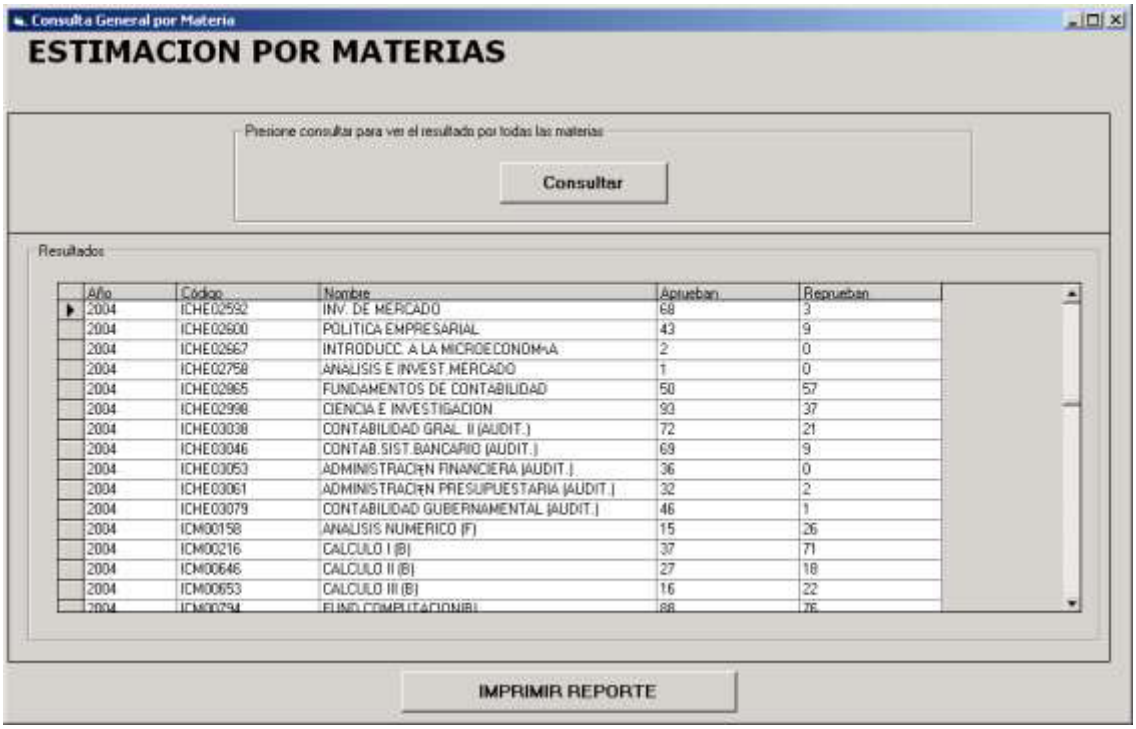
Ejecutar actualización.- Permite ingresar o actualizar los datos de entrenamiento y prueba del sistema, una vez dispuestos los datos en dos archivos ubicados en la carpeta C:\Mis Documentos\Sistema_Estimacion; permite realizar la carga descrita y actualizar la información del sistema antes de comenzar a realizar cálculos de preparación y estimación en ese orden.

Inicializar valores.- Al ejecutar este comando, damos paso a la ejecución de todos los procedimientos almacenados de preparación de datos definidos a nivel de base de datos, los cuales permitirán vizualizar los reportes que se disponen al ejecutar **Continuar**.

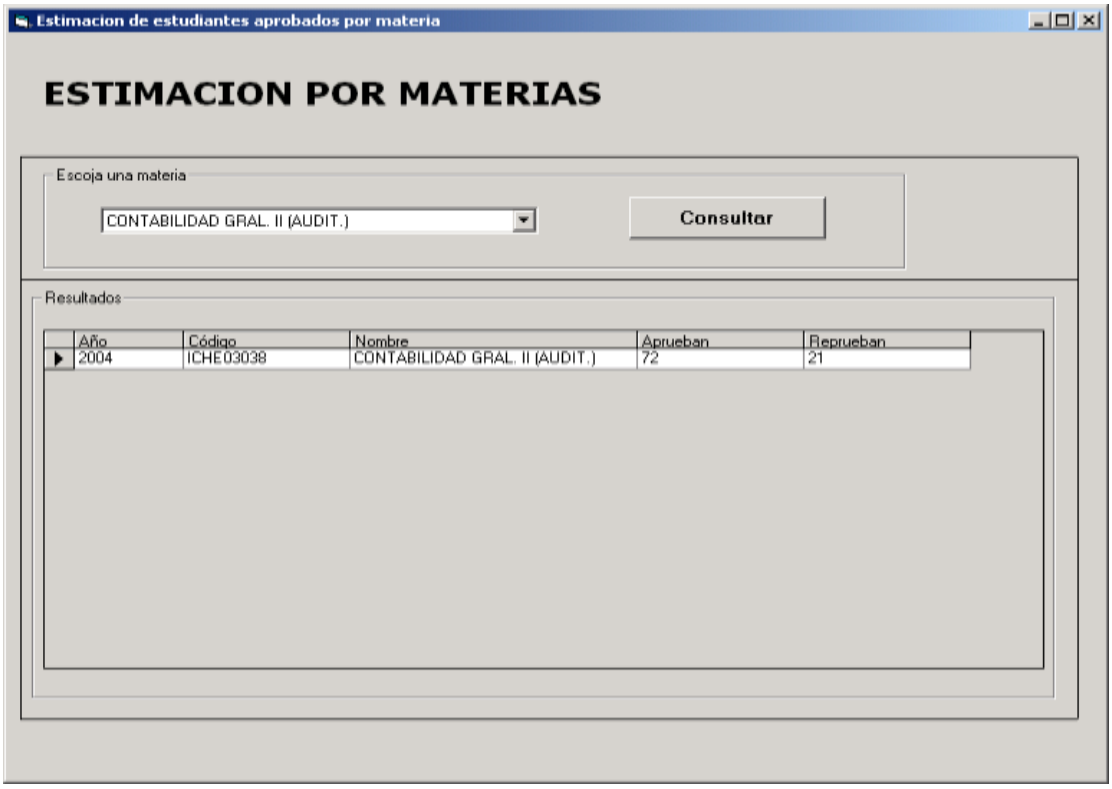


Dependiendo de la consulta a realizar tenemos las siguientes opciones:

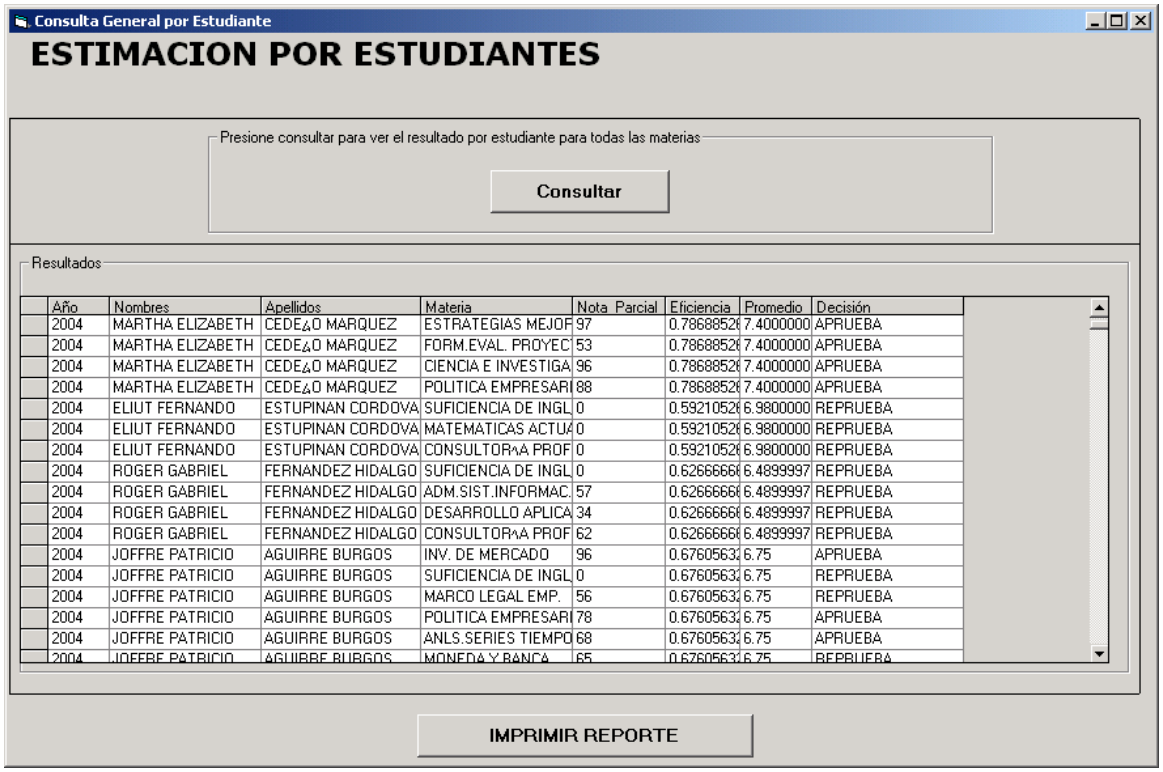
- *Reporte de Materias por cantidad de aprobados y reprobados.*- Donde se muestra para cada materia de la ESPOL donde se encuentre registrado un estudiante del ICM, la cantidad de estudiantes que aprueban y reprueban la misma.



- *Consulta por Materia.-* Muestra la misma información del reporte anterior, pero para materias individuales seleccionadas de un combo.

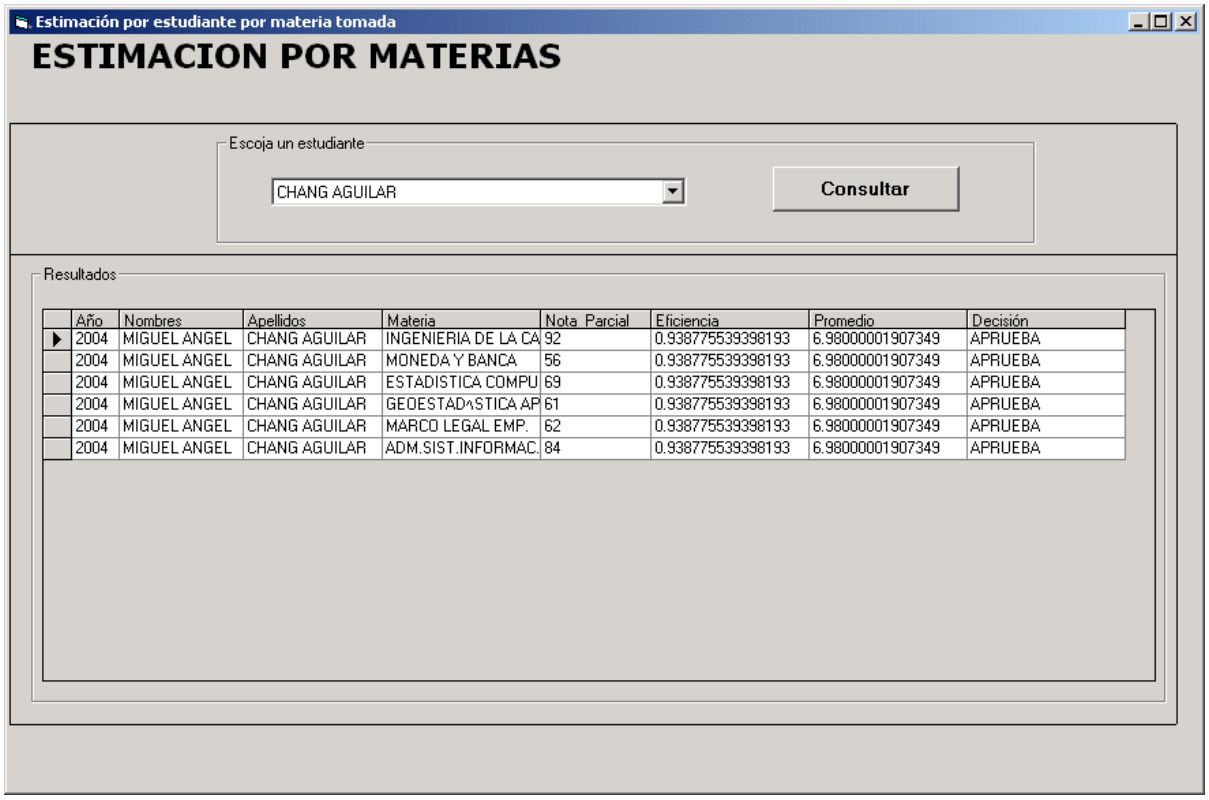


Reporte de Estudiantes por materia y aprobación.- Muestra la información a nivel de estudiante, detallando los datos del mismo y si aprueba o reprueba según el sistema.



Año	Nombres	Apellidos	Materia	Nota Parcial	Eficiencia	Promedio	Decisión
2004	MARTHA ELIZABETH	CEDE¿O MARQUEZ	ESTRATEGIAS MEJOR	97	0.7868852(	7.4000000	APRUEBA
2004	MARTHA ELIZABETH	CEDE¿O MARQUEZ	FORM.EVAL. PROYEC	53	0.7868852(	7.4000000	APRUEBA
2004	MARTHA ELIZABETH	CEDE¿O MARQUEZ	CIENCIA E INVESTIGA	96	0.7868852(	7.4000000	APRUEBA
2004	MARTHA ELIZABETH	CEDE¿O MARQUEZ	POLITICA EMPRESARI	88	0.7868852(	7.4000000	APRUEBA
2004	ELIUT FERNANDO	ESTUPINAN CORDOVA	SUFICIENCIA DE INGL	0	0.5921052(	6.9800000	REPRUEBA
2004	ELIUT FERNANDO	ESTUPINAN CORDOVA	MATEMATICAS ACTUA	0	0.5921052(	6.9800000	REPRUEBA
2004	ELIUT FERNANDO	ESTUPINAN CORDOVA	CONSULTORíA PROF	0	0.5921052(	6.9800000	REPRUEBA
2004	ROGER GABRIEL	FERNANDEZ HIDALGO	SUFICIENCIA DE INGL	0	0.6266666(	6.4899997	REPRUEBA
2004	ROGER GABRIEL	FERNANDEZ HIDALGO	ADM.SIST.INFORMAC.	57	0.6266666(	6.4899997	REPRUEBA
2004	ROGER GABRIEL	FERNANDEZ HIDALGO	DESARROLLO APLICA	34	0.6266666(	6.4899997	REPRUEBA
2004	ROGER GABRIEL	FERNANDEZ HIDALGO	CONSULTORíA PROF	62	0.6266666(	6.4899997	REPRUEBA
2004	JOFFRE PATRICIO	AGUIRRE BURGOS	INV. DE MERCADO	96	0.6760563(	6.75	APRUEBA
2004	JOFFRE PATRICIO	AGUIRRE BURGOS	SUFICIENCIA DE INGL	0	0.6760563(	6.75	REPRUEBA
2004	JOFFRE PATRICIO	AGUIRRE BURGOS	MARCO LEGAL EMP.	56	0.6760563(	6.75	REPRUEBA
2004	JOFFRE PATRICIO	AGUIRRE BURGOS	POLITICA EMPRESARI	78	0.6760563(	6.75	APRUEBA
2004	JOFFRE PATRICIO	AGUIRRE BURGOS	ANLS.SERIES TIEMPO	68	0.6760563(	6.75	APRUEBA
2004	JOFFRE PATRICIO	AGUIRRE BURGOS	MONEDA Y BANCA	65	0.6760563(	6.75	REPRUEBA

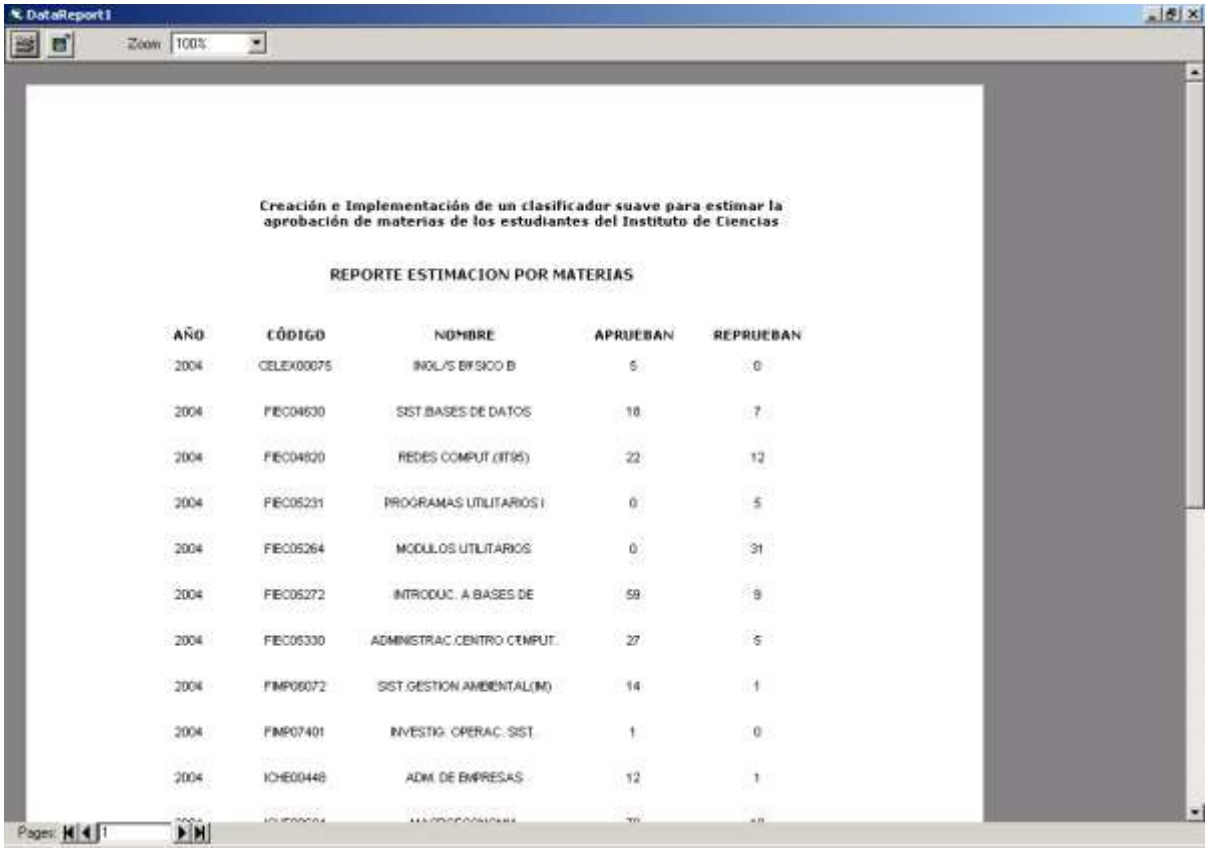
- *Consulta por Estudiante.-* Muestra la misma información del reporte anterior, pero para estudiantes individuales seleccionadas de un combo.



Año	Nombres	Apellidos	Materia	Nota Parcial	Eficiencia	Promedio	Decisión
2004	MIGUEL ANGEL	CHANG AGUILAR	INGENIERIA DE LA CA	92	0.938775539398193	6.98000001907349	APRUEBA
2004	MIGUEL ANGEL	CHANG AGUILAR	MONEDA Y BANCA	56	0.938775539398193	6.98000001907349	APRUEBA
2004	MIGUEL ANGEL	CHANG AGUILAR	ESTADISTICA COMPU	69	0.938775539398193	6.98000001907349	APRUEBA
2004	MIGUEL ANGEL	CHANG AGUILAR	GEOESTADÝSTICA AP	61	0.938775539398193	6.98000001907349	APRUEBA
2004	MIGUEL ANGEL	CHANG AGUILAR	MARCO LEGAL EMP.	62	0.938775539398193	6.98000001907349	APRUEBA
2004	MIGUEL ANGEL	CHANG AGUILAR	ADM.SIST.INFORMAC.	84	0.938775539398193	6.98000001907349	APRUEBA

REPORTES:

REPORTE GERENCIAL.- Contiene información útil para elaborar la planificación académica.



REPORTE DE DETALLE.- Con información por estudiante, válido para niveles de consulta detallados.