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

Juan Alvarado Ortega
Miguel Chang Aguilar
Instituto de Ciencias Matemáticas
Escuela Superior Politécnica del Litoral
Kilometro 30.5 Vía Perimetral, Guayaquil, Ecuador
jao_ec@yahoo.com
machang@espol.edu.ec

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.

Palabras Claves: *Minería de datos, clasificación bayesiana, Bayes ingenuo.*

Abstract

The statistic conjunction and computer science obtains knowledge as a result in pure state, it is very difficult to imagine in independent way that advances it existed in each branch in an individual way if the other one didn't exist, although at first sight we could define as the calculation like an independent area, the development sustained in the last times of the same one would not be possible if the numeric sciences didn't exist .The present study contains a great part of both, achieving in certain way to make a natural interaction among the same ones. The classification bayesiana applied to the resolution of the problem of estimate of approval of matters for students of the Institute of Mathematical Sciences of the ESPOL, responds in great measure to that we can analyze the entirety of available registrations so quickly for the semester specific estimate object on the defined pattern of on like we will classify, and to apply it to a same quantity or smaller than additional data or to extend it has the other academic units, having results that compared to the reality, they approach in a reliable enough way as it is explained along the study.

1. 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 período académico.

2. 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.

La presente guía sirve además como plantilla con los márgenes que aquí se describen.

3. 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.

4. 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.

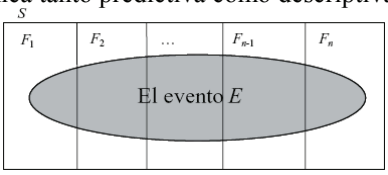

Figura 1. Evento E en Marco Muestral S

5. 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.

6. Planificación académica

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."

7. El proceso de Planificación académica

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

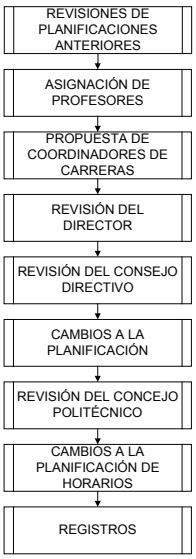

Figura 2. Estructuración de un Sistema Informático Gerencial.

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.

8. 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".

8.1. 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.

8.2. 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.

9. 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.

10. 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.

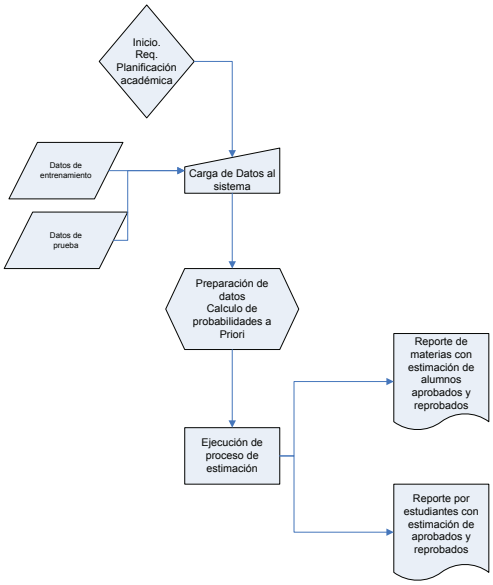

Figura 3. Diagrama de Flujo de la Solución.

10.1. Definición de Variables de Clasificación

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.

10.3. 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.

10.4. 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'.

10.5 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.

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 demuestre la independencia estadística de las mismas.

Tabla 1. 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

10.6. 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.

10.7. 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.

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

10.8. Construcción del Modelo

Es importante como se realizará a continuación, redefinir los datos de entrenamiento del clasificador y los datos de prueba sobre los cuales se realizarán:

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.

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. La cantidad de registros disponibles como datos de prueba son 5608 que corresponden al 13% de datos como datos de prueba.

La construcción del modelo se define de la siguiente forma:

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

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.

10.9. Validación del Modelo

Resultados de la 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.

Prueba de Decisión vs. 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.

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.

11. Referencias

[1] Han, J., & Kamber, M, *Data Mining : Concepts and Techniques*, Morgan Kaufmann, 2000.

[2] Rencher, A., *Methods of Multivariate Analysis – Second Edition*, John Wiley & Sons, 2002.

[3] Alvarado Ortega, Juan., "Algoritmos de la Minería de Datos" Revista Matemática, ICM - ESPOL, 2003.

12. 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. 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.

13. 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.