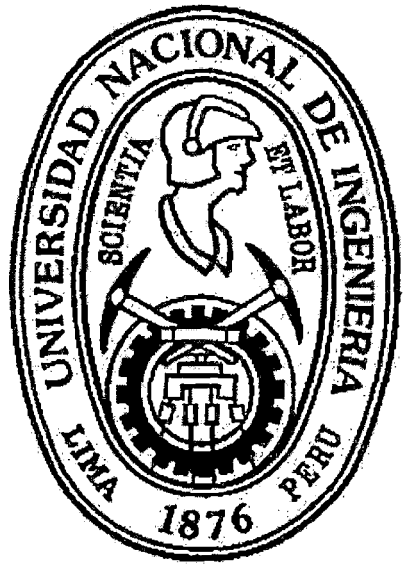
UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS

SECCIÓN DE POSGRADO



PREDICCIÓN DEL RENDIMIENTO ACADÉMICO EN LA EDUCACIÓN SUPERIOR USANDO MINERÍA DE DATOS Y SU COMPARACIÓN CON TÉCNICAS ESTADÍSTICAS

TESIS

Para optar el Grado Académico de

MAESTRO EN CIENCIAS

CON MENCIÓN EN INGENIERÍA DE SISTEMAS

PEDRO RAÚL ACOSTA DE LA CRUZ

PEDRO ARTURO PIZARRO SOLÍS

LIMA – PERÚ

2011

DEDICATORIA

A mi hija Nicole, con infinita ternura
A mi esposa Maribel por su gran apoyo.
A mis queridos padres y hermanos.

DEDICATORIA

A mi amada esposa Beatriz
A mis queridos padres y hermanos.

AGRADECIMIENTO

Agradecemos a nuestro asesor el Msc. Luis Miguel Sierra Flores, al Msc. Josue Ángulo Pérez, al Lic. Yarko Cerna Valdez, al Ing. Alejandro Huapaya Sánchez y a todos aquellos que han contribuido al logro de este trabajo de investigación.

A la Facultad de Ingeniería Industrial y Sistemas en su conjunto por su cálida acogida.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

DESCRIPTORES TEMÁTICOS

- Data Mining
- Regresión Multivariable
- Regresión Logística
- Redes Neuronales Artificiales
- Predicción del Rendimiento Académico
- Aplicación en la Educación Superior

RESUMEN

En la mayoría de las universidades se sigue un sistema de currículo flexible, esto significa que a partir del segundo ciclo de estudios, los estudiantes universitarios pueden escoger los cursos a llevar, siempre y cuando se cumpla con el currículo y los reglamentos académicos correspondientes. Una gran dificultad en el proceso de inscripción es que el estudiante no tiene un sistema de ayuda o recomendación para tomar una buena decisión en la elección de los cursos a llevar, de tal manera que tenga la mayor probabilidad de salir airoso en su rendimiento académico.

En este trabajo se aplica modelos predictivos (redes neuronales, regresión logística y regresión múltiple), que permitan al estudiante universitario predecir su rendimiento académico de cada curso en que desea inscribirse.

El objetivo del estudio es predecir (a) si el alumno aprobará o no un curso ó (b) la nota del curso. Para ello primero se realiza una selección de las variables predictoras, en base a la experiencia de los autores en la cátedra universitaria y luego confrontando estas variables con las usadas en trabajos publicados relacionados al tema. Después usando la base de datos académica de los alumnos de una especialidad (previamente preparados) y el currículo correspondiente, se obtendrán los datos para las variables seleccionadas, mediante un programa en Java.

Se aplica las técnicas de predicción a 7 cursos de la especialidad de Ingeniería Química de la Universidad Nacional de Ingeniería, usando los datos de los períodos académicos del 1993-1 al 2010-2.

La aplicación de las técnicas de redes neuronales de retropropagación y de regresión logística para la predicción de la aprobación o no de un curso, arrojan promedios de porcentajes de aciertos similares, de 70.45 % y 70.39 % para los modelos, y de 72.83 % y 74.04 % para los pronósticos, respectivamente.

La aplicación de las técnicas de redes neuronales de retropropagación y de regresión múltiple para la predicción de la "nota" de un curso, arrojan promedios de raíz de errores medios cuadráticos similares, de 0.1495 y 0.1430 para los modelos, y de 0.1397 y 0.1380 para los pronósticos, respectivamente.

No se requiere de una herramienta sofisticada para la aplicación del modelo de redes neuronales de retropropagación. En este trabajo se ha utilizado el Excel de Microsoft con su complemento Solver para la implementación de la red neuronal con diferentes número de capas y neuronas por capas.

INTRODUCCION

El estudiante universitario gracias al sistema de currículo flexible, tiene la libertad de tomar sus decisiones con respecto a los cursos que llevará por ciclo, respetando el currículo y el reglamento académico correspondiente. Si bien es cierto, esto es una buena forma de tomar en consideración las situaciones particulares de cada estudiante, éste debería tomar estas decisiones teniendo toda la información necesaria. Una de ellas por ejemplo, el estimado de las notas (o la aprobación o no) de los cursos en los cuales desea inscribirse. Esto le permitiría rediseñar el conjunto de cursos que desea llevar y/o realizar un plan de preparación para el nuevo ciclo.

En la Facultad de Ingeniería Química y Textil (FIQT) de la UNI, la inscripción en un nuevo ciclo se realiza a través de Internet, para ello los estudiantes reciben, el currículo de su especialidad, su record académico, que consiste en una relación de los cursos que han llevado cada ciclo con sus notas obtenidas, así como el promedio ponderado por ciclo y el acumulado. Con esta información se inscriben en los cursos a llevar. El sistema de inscripción luego verifica: (a) que el estudiante haya aprobado los pre-requisitos de cada curso y (b) que los cursos inscritos estén dentro de tres ciclos. Si bien es cierto, en el reglamento de inscripción existen restricciones con respecto al número de créditos que pueden llevar y que están en función del promedio ponderado de los dos últimos ciclos, estas restricciones no son tan fuertes ya que con un promedio ponderado de 7 o más, se puede llevar hasta 20 créditos.

Una gran dificultad en el proceso de inscripción, es que el estudiante no tiene un sistema de ayuda o recomendación para tomar una buena decisión. El estudiante actualmente toma su decisión, teniendo muchas veces como meta sus deseos de terminar prontamente la carrera, es decir, inscribiéndose en todos los cursos posibles; y si tiene que escoger entre varias alternativas de cursos, consulta principalmente a sus amigos más cercanos, sin basarse en su record académico; y más aún, una vez inscrito no elabora ninguna estrategia de repasar o revisar cursos llevados para la mejora de su rendimiento académico en el nuevo ciclo.

En este trabajo se pretende elaborar modelos predictivos usando las técnicas de Redes Neuronales de retropropagación, la Regresión Logística y la Regresión múltiple La predicción del rendimiento académico consiste en predecir la nota que obtendrá en el curso o predecir si saldrá aprobado en el curso. Los datos reales que se utilizan son la base histórica académica de los estudiantes de la especialidad de Ingeniería Química de la Universidad Nacional de Ingeniería desde 1993 hasta el 2010.

La información suministrada por estos modelos predictivo, permitirá al estudiante tomar para su inscripción decisiones que le ayudarán a avanzar su carrera en forma satisfactoria. Daremos algunos ejemplos, así en el caso (a) de cursos obligatorios, que necesariamente tiene que llevar, en los cuales se predice una desaprobación o baja nota, tendrá que repasar el o los pre-requisitos, o disponer de mayor tiempo para su estudio, (b) en cursos obligatorios, que no necesariamente tiene que llevar, en los cuales se predice una desaprobación o baja nota, tendrá para escoger una mejor alternativa ó disponer de más tiempo para su estudio, si es que de todas maneras quiere llevarlo, (c) de cursos electivos en los cuales se predice una baja nota o desaprobación, podrá seleccionar una mejor opción.

Actualmente existe un creciente interés en el empleo de las técnicas de Minería de datos en el campo de la Educación. La mayoría de estos trabajos están aplicados a sistemas de educación tradicionales, a cursos particulares a distancia vía Web, a sistemas de manejo de contenidos de aprendizaje y a sistemas educacionales vía Web adaptativos e inteligentes. Estos trabajos se han presentado desde el año 2008 en las diferentes Conferencias anuales sobre Minería de Datos en Educación (MDE). Sin embargo existen muy pocos trabajos relacionados con la predicción del rendimiento académico de un estudiante universitario en un curso en el cual se quiere inscribir. Existen por ejemplo los trabajos de Romero y los de Vialardi.

CAPITULO I

PLANTEAMIENTO DEL PROBLEMA

1.1 Introducción

El estudiante universitario, a diferencia del estudiante de educación secundaria, tiene la libertad de tomar sus decisiones con respecto a los cursos que llevará por ciclo, respetando el currículo y el reglamento académico correspondiente. Esto gracias al sistema de currículo flexible que existe en las universidades. Si bien es cierto, esto es una buena forma de tomar en consideración las situaciones particulares de cada estudiante, éste debería tomar estas decisiones teniendo toda la información necesaria. Una de ellas por ejemplo, el estimado de las notas (o la aprobación o no) de los cursos en los cuales desea inscribirse. Esto le permitiría rediseñar el conjunto de cursos que desea llevar y/o realizar un plan de preparación para el nuevo ciclo.

1.2 Identificación del Problema

En la mayoría de las universidades se sigue un sistema de currículo flexible, esto significa que a partir del segundo ciclo de estudios, los estudiantes universitarios pueden escoger los cursos a llevar, siempre y cuando se cumpla con ciertos requisitos. En la Facultad de Ingeniería Química y Textil (FIQT) de la UNI, la inscripción en un nuevo ciclo se realiza a través de Internet, para ello los estudiantes reciben, el currículo de su especialidad, su record académico, que consiste en una relación de los cursos que han llevado cada ciclo con sus notas obtenidas, así como el promedio ponderado por ciclo y el acumulado. Con esta información se inscriben en los cursos a

llevar. El sistema de inscripción luego verifica: (a) que el estudiante haya aprobado los pre-requisitos de cada curso y (b) que los cursos inscritos estén dentro de tres ciclos. Si bien es cierto, en el reglamento de inscripción existen restricciones con respecto al número de créditos que pueden llevar y que están en función del promedio ponderado de los dos últimos ciclos, estas restricciones no son tan fuertes ya que con un promedio ponderado de 7 o más, se puede llevar hasta 20 créditos. Una gran dificultad en el proceso es que el estudiante no tiene un sistema de ayuda o recomendación para tomar una buena decisión en la elección de los cursos a llevar, de tal manera que tenga la mayor probabilidad de salir airoso en su rendimiento académico.

En este trabajo se pretende elaborar modelos predictivos mediante el uso de minería de datos y técnicas estadísticas, que permitan al estudiante universitario predecir su rendimiento académico en cada curso en que desea inscribirse.

1.3 Definición del Problema

El problema de este trabajo de investigación queda definido con la siguiente interrogante:

¿Es posible aplicar técnicas de minería de datos, como redes neuronales de retropropagación, y técnicas estadísticas como regresión logística y regresión múltiple para predecir el rendimiento académico que obtendrá el estudiante (considerando si aprobará o no, así como la nota que sacará) en cada uno de los cursos en que se inscribe?

1.4 Importancia del Problema

Hace varios años, los estudiantes previos a su inscripción, tenían que pasar por un asesoramiento de profesores, que básicamente consistía en que el profesor (en pocos minutos) en base al reglamento de inscripción y al record académico del estudiante, rellenaba la ficha de inscripción. El profesor asesor no tenía potestad para reducir el número de cursos que el alumno podía llevar por reglamento, ni tampoco imponer alguna alternativa de cursos (en caso las

hubiera) solo recomendaba. El alumno mientras se hallaba dentro del reglamento tenía la decisión final. Actualmente el alumno ya no pasa por esta asesoría, se inscribe por internet y el cumplimiento del reglamento de inscripción se verifica luego de la inscripción. Esto significa que el alumno toma su decisión, tomando muchas veces como meta sus deseos de terminar prontamente la carrera, es decir, inscribiéndose en todos los cursos posibles; y si tiene que escoger entre varias alternativas de cursos, consulta principalmente a sus amigos más cercanos, sin basarse en su record académico; y más aún, una vez inscrito no elabora ninguna estrategia de repasar o revisar cursos llevados para la mejora de su rendimiento académico en el nuevo ciclo.

Este trabajo permitirá al estudiante tener una herramienta que le permitirá realizar un autodiagnóstico y predecir cómo será su rendimiento académico en cada curso que desea inscribirse. Esto le dará oportunidad de tomar decisiones para su inscripción, tal como se expresa en el diagrama de flujo de la figura 1.1. Lo que es cierto, que la decisión final, la tendrá siempre el estudiante.

1.5 Justificación del Problema

El disponer de esta herramienta le permitirá al estudiante tomar para su inscripción decisiones que le ayudarán a avanzar su carrera en forma satisfactoria. Daremos algunos ejemplos, así en el caso (a) de cursos obligatorios, que necesariamente tiene que llevar, en los cuales se predice una desaprobación o baja nota, tendrá que repasar el o los pre-requisitos, o disponer de mayor tiempo para su estudio, (b) en cursos obligatorios, que no necesariamente tiene que llevar, en los cuales se predice una desaprobación o baja nota, tendrá para escoger una mejor alternativa ó disponer de más tiempo para su estudio, si es que de todas maneras quiere llevarlo, (c) de cursos electivos en los cuales se predice una baja nota o desaprobación, podrá seleccionar una mejor opción.

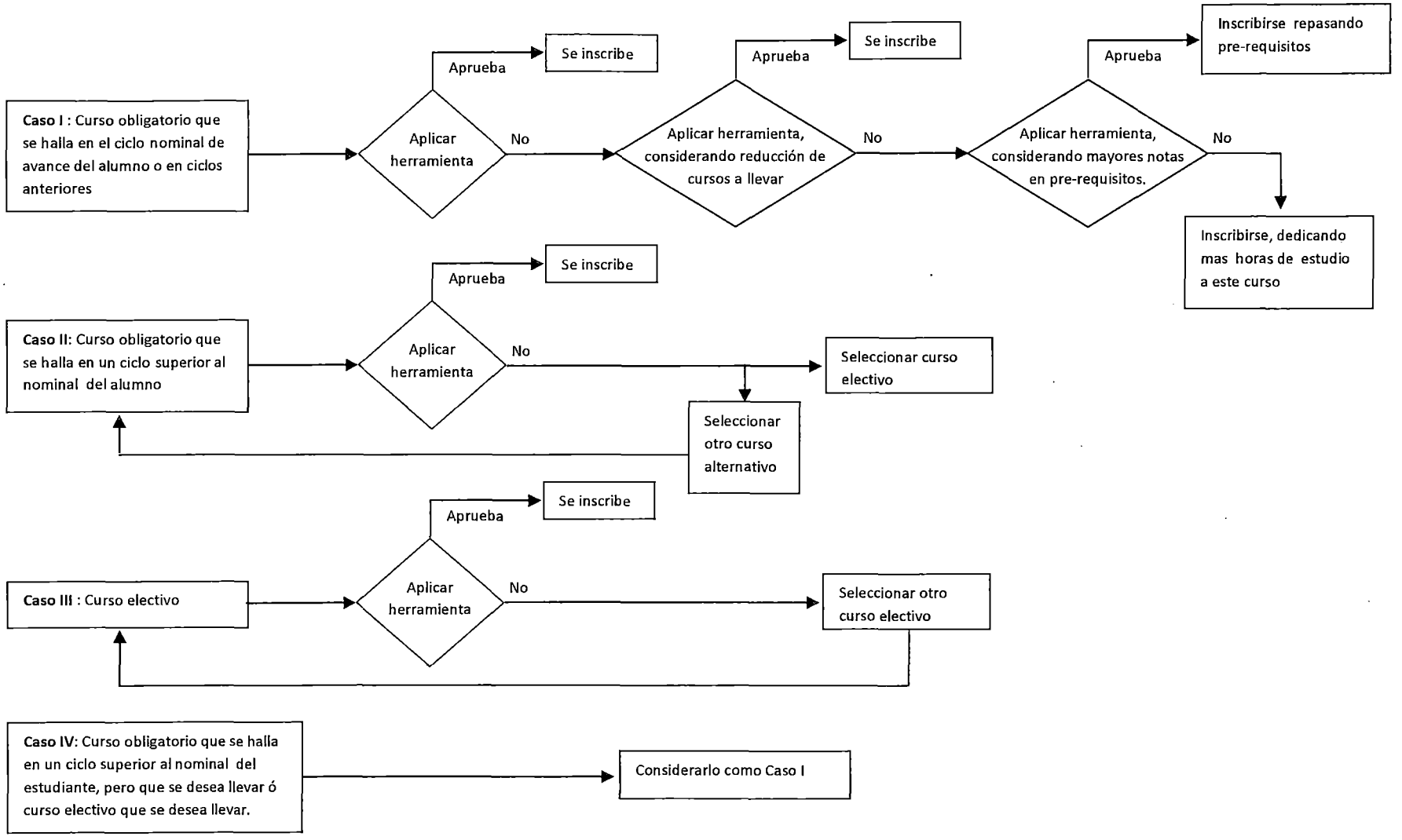


Figura 1.1 Diagrama de flujo para la aplicación de una herramienta predictiva (del rendimiento de un estudiante en un curso) en su inscripción en un nuevo período académico

1.6 Delimitación del Problema

En este trabajo se busca predecir el rendimiento académico que obtendrá un estudiante universitario (considerando si aprobará o no, así como la nota que sacará) en cada uno de los cursos en que se inscribe, utilizando tres técnicas de predicción, a saber, redes neuronales de retropropagación, regresión logística y regresión múltiple. La predicción se hará en base al rendimiento académico previo del estudiante y al rendimiento académico de otros estudiantes con perfiles similares que hayan llevado el curso. Se realizará la aplicación para la especialidad de Ingeniería Química de la Facultad de Ingeniería Química y Textil de la UNI.

1.7 Objetivos de la Investigación

1.7.1 Objetivo General

Utilizar técnicas de minería de datos y técnicas estadísticas para predecir el rendimiento académico de un estudiante universitario en un curso en que se desea inscribir.

1.7.2 Objetivos Específicos

1) Utilizar técnicas de redes neuronales de retropropagación para predecir si un estudiante universitario aprobará o no un curso en el que se desea inscribir.
2) Utilizar técnicas de regresión logística para predecir si un estudiante universitario aprobará o no un curso en el que se desea inscribir
3) Utilizar técnicas de redes neuronales de retropropagación para predecir la nota que obtendrá un estudiante universitario en un curso en el que se desea inscribir
4) Utilizar técnicas de regresión múltiple para predecir la nota que obtendrá un estudiante universitario en un curso en el que se desea inscribir

1.8 Formulación de la Hipótesis

1.8.1 Hipótesis General

La utilización de técnicas de minerías de datos y técnicas estadísticas aplicadas a la universidad peruana, permitiría al estudiante predecir su rendimiento académico en un curso en que se desea inscribir en un nuevo ciclo.

1.8.2 Hipótesis Específicas

1.8.2.1 Hipótesis Específica Nº 1

Las redes neuronales de retropropagación aplicadas a la universidad peruana, permitirían al estudiante predecir si aprobará o no un curso en que se desea inscribir en un nuevo ciclo.

1.8.2.2 Hipótesis Específica Nº 2

La técnica de regresión logística aplicada a la universidad peruana, permitiría al estudiante predecir si aprobará o no un curso en que se desea inscribir en un nuevo ciclo.

1.8.2.3 Hipótesis Específica Nº 3

Las redes neuronales de retropropagación aplicadas a la universidad peruana, permitirían al estudiante predecir la nota que obtendrá en un curso en que se desea inscribir en un nuevo ciclo.

1.8.2.4 Hipótesis Específica Nº 4

La técnica de regresión múltiple aplicada a la universidad peruana, permitiría al estudiante predecir la nota que obtendrá en un curso en que se desea inscribir en un nuevo ciclo.

1.8.3 Definición conceptual de las variables

Tanto para la Hipótesis General como para las 4 Hipótesis específicas las variables son las siguientes:

1.8.3.1 Variables independientes:

a) Base histórica académica de notas de la Facultad de Ingeniería Química y Textil, correspondientes a todos los alumnos de la especialidad de Ingeniería Química.
b) Currículo de la Especialidad de Ingeniería Química
c) Relación de cursos en que el estudiante desea inscribirse.

1.8.3.2 Variable dependiente:

Rendimiento del alumno en un curso, expresado como nota del alumno en el curso. De aquí se obtendrá la variable discreta aprobado/desaprobado y la variable continua la nota misma.

1.8.4 Definición operacional de las variables (ver sección 3.2.2)

Tanto para la Hipótesis General como para las 4 Hipótesis específicas las variables son las siguientes:

1.8.4.1 Variables independientes:

a) **Variables relacionadas con el rendimiento global del estudiante** En estas variables se toman en cuenta el desempeño académico del alumno en base a todos los cursos llevados y al tiempo que lleva en la universidad y son:

1. Promedio ponderado acumulado al semestre previo
2. Antigüedad en años del alumno desde el año de ingreso a la UNI hasta el año del semestre previo inclusive.

b) Variables que dependen del rendimiento del estudiante relacionado al curso

Estas variables se refieren a los rendimientos de los estudiantes, pero más orientados al curso. Para ello se consideraran al o a los cursos pre-requisitos del curso en estudio. Se tienen a:

3. Nota del curso pre-requisito 1
4. Nota del curso pre-requisito 2 (en caso exista)

c) Variables relacionadas con el grado de dificultad de aprobación de un curso

En estas variables se toman en cuenta en forma global, la influencia de los diferentes factores que hacen que un curso sea más difícil de aprobar que otros y es:

5. Promedio del Grado de Dificultad de aprobación de un curso por semestre hasta el semestre previo.

d) Variables relacionadas con la influencia de los otros cursos que se quiere llevar en el nuevo semestre.

Estas variables toman en cuenta la carga académica del estudiante, es decir, el total de cursos en los que se quiere matricular.

6. Número de créditos totales que se quiere llevar en el semestre actual
7. Sumatoria del promedio del Grado de Dificultad de aprobación de un curso (por semestre hasta el semestre previo), evaluado para todos los cursos que se llevaran en el semestre actual.

1.8.4.2 Variable dependiente:

El rendimiento académico del estudiante en un curso está expresado por la calificación final que obtiene después de haberlo llevado. Esta calificación llamada Nota nos dará dos variables dependientes, que serán estudiadas por separado, a saber:

i. Variable continua: Nota del alumno

ii. Variable discreta: Aprobado o Desaprobado.

CAPITULO II

FUNDAMENTO TEÓRICO Y METODOLÓGICO

2.1 Marco Referencial de la Investigación

2.1.1 Antecedentes

Actualmente existe un creciente interés en el empleo de las técnicas de Minería de datos en el campo de la Educación. La mayoría de estos trabajos están aplicados a sistemas de educación tradicionales, a cursos particulares a distancia vía Web, a sistemas de manejo de contenidos de aprendizaje y a sistemas educacionales vía Web adaptativos e inteligentes [16]. En el año 2008, en Quebec (Canadá) se llevó a cabo la Primera Conferencia Internacional sobre Minería de Datos en Educación (MDE), la segunda se realizó el 2009 en Córdoba (España), la tercera en Pittsburgh (EEUU) en el 2010 y la cuarta en Eindhoven (Holanda) en Julio del 2011. Sin embargo existen muy pocos trabajos relacionados con la predicción del rendimiento académico de un estudiante universitario en un curso en el cual se quiere inscribir. A continuación comentaremos algunos trabajos realizados relacionados al tema de nuestro estudio:

i. Trabajo de Romero, Ventura, Espejo y Hervás [17]

En este estudio se compara diferentes técnicas de Minería de Datos para Clasificación (Árbol de decisión, clasificador estadístico, reglas de inducción, lógica difusa y redes neuronales) de estudiantes. La habilidad de predecir el rendimiento de un estudiante es muy importante en la educación a distancia vía Web, de hecho una de las principales tareas en e-learning es

la clasificación. Los datos usados son, los atributos que se recogen para cada estudiante en el Curso vía Web, para predecir la nota final del Curso en forma categórica. Se usaron datos reales de siete cursos con estudiantes de la Universidad de Córdoba. Se aplicó la discretización y el rebalanceo dentro de las técnicas de pre-procesamiento para los datos originales. El porcentaje global de los resultados correctamente clasificados se encuentra entre el 50 y el 67 %. Este valor del 67 % se obtuvo con el algoritmo CART de la técnica Árbol de Decisión.

ii. Trabajo de Vialardi, Bravo, Shafti y Ortigosa [21]

Propone un sistema de recomendación basado en técnicas de Minería de Datos para ayudar al estudiante para escoger mejor en cuantos y en qué cursos se debe inscribir, teniendo como base la experiencia de estudiantes anteriores con similares perfiles académicos.

Se han analizado datos reales de estudiantes correspondientes a siete años en la Escuela de Ingeniería de la Universidad de Lima. Las variables usadas son: el número de cursos tomados simultáneamente, nombre del curso, el promedio ponderado acumulado, que servirán para predecir el Grado, que discretizado sería: desaprobado (nota de 0 a 10.99) y aprobado (de 11.00 a 20). La técnica de Minería de Datos usada es la de Árbol de Decisión, específicamente el algoritmo C4.5. El porcentaje de instancias correctamente clasificadas asciende a 73.9 %.

iii. Trabajo de Vialardi [22]

En este estudio, al igual que en caso anterior, se implementa una herramienta para predecir la conveniencia de cursar una asignatura específica para un estudiante determinado, sobre la base de los resultados obtenidos por estudiantes de rendimientos académicos similares que hayan

cursado antes esa asignatura. Se ha utilizado datos reales de la Facultad de Ingeniería de Sistemas de la Universidad de Lima desde 1991 hasta 2009. Utiliza los sistemas de filtrado colaborativos clásicos basados en memoria, el clasificador Árbol de Decisión C4.5 y los clasificadores Bagging y Boosting. Se emplea como variables predictoras al "potencial", que es la habilidad que tiene el alumno para el curso en estudio y a la variable "dificultad", que indica la dificultad del curso en base al promedio de los alumnos que la han llevado.

Nuestro trabajo, tiene los mismos objetivos que las dos referencias anteriores, pero utilizamos siete variables predictoras, y empleamos como técnicas de predicción a las Redes Neuronales de retropropagación, a la Regresión múltiple y a la Regresión Logística. Los datos reales que se utilizan son la base histórica académica de los estudiantes de la especialidad de Ingeniería Química de la Universidad Nacional de Ingeniería desde 1993 hasta el 2010.

2.1.2 Marco Teórico

Las técnicas predictivas a emplear son: Redes Neuronales Artificiales (RNA), Regresión Múltiple (RM) y Regresión Logística (RL).

2.1.2.1 Minería de Datos

La Minería de datos de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.

La disponibilidad de grandes volúmenes de información y el uso generalizado de las herramientas informáticas han transformando el análisis de datos orientándolo hacia determinadas técnicas

especializadas englobadas bajo el nombre de Minería de datos o *Data Mining*.

Las técnicas de Minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir, perfiles y tendencias a través del análisis de datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos.

Las técnicas predictivas especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido. Formalmente, la aplicación de todo modelo debe superar las fases de *identificación objetiva* (a partir de los datos se aplican reglas que permiten identificar el mejor modelo posible que ajuste los datos), *estimación* (proceso de cálculo de los parámetros del modelo elegido para los datos en la fase de identificación), diagnosis (proceso de contraste de la validez del modelo estimado) y *predicción* (proceso de utilización del modelo identificado, estimado y validado para predecir valores futuros de las variables dependientes). En algunos casos, el modelo se obtiene como mezcla del conocimiento obtenido antes y después del *Data Mining* y también debe contrastarse antes de aceptarse como válido. Por ejemplo, *las redes neuronales* permiten descubrir modelos complejos y afinarlos a medida que progresa la exploración de los datos. Gracias a su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa. Podemos incluir

entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. Tanto los árboles de decisión, como las redes neuronales y el análisis discriminante son a su vez *técnicas de clasificación* que pueden extraer perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato. Los árboles de decisión permiten clasificar los datos en grupos basados en los valores de las variables. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas.

2.1.2.2 Redes Neuronales

a) Definición

La Red neuronal se puede definir como un conjunto de elementos de procesamiento de la información altamente interconectados, que son capaces de aprender con la información que se les alimenta. La principal característica de esta nueva técnica es que puede aplicarse a un gran número de problemas que pueden ir desde problemas complejos reales a modelos teóricos sofisticados como por ejemplo reconocimiento de imágenes, análisis y filtrado de señales, clasificación, predicción dinámica, etc.

Las Redes neuronales tratan de emular al sistema nervioso, de forma que son capaces de reproducir algunas de las principales tareas que desarrolla el cerebro humano, al reflejar las características fundamentales del comportamiento del mismo. Lo que realmente intenta modelar las redes neuronales es una de las estructuras

fisiológicas del soporte del cerebro, las neuronas y los grupos estructurados e interconectados de varias de ellas, conocidos como redes de neuronas. De este modo construyen sistemas que presentan un cierto grado de inteligencia. No obstante, debemos insistir en el hecho de que los Sistemas Neuronales Artificiales, como cualquier otra herramienta construida por el hombre, tienen limitaciones y solo poseen un parecido superficial con sus contrapartidas biológicas. Las redes neuronales, en relación con el procesamiento de la información, heredan tres características básicas de las redes de neuronas biológicas: paralelismo masivo, respuesta no lineal de las neuronas frente a las entradas recibidas y procesamiento de la información a través de múltiples capas de neuronas.

Una de las principales propiedades de estos modelos es su capacidad de aprender a partir de ejemplos reales. Es decir, la red aprende a reconocer la relación (que no deja de ser equivalente a estimar una dependencia funcional) que existe entre el conjunto de entradas proporcionadas como ejemplos y sus correspondientes salidas, de modo que, finalizado el aprendizaje, cuando a la red se le presenta una nueva entrada, en base a la relación funcional establecida en el mismo, es capaz de generalizarla ofreciendo una salida. En consecuencia, podemos definir una red neuronal artificial como un sistema inteligente capaz, no solo de aprender, sino también de generalizar.

Una red neuronal está formada por unidades de procesamiento que reciben el nombre de neuronas o nodos. Estos nodos están organizados en grupos que se llaman "capas". Generalmente existen

tres tipos de capas: una capa de entrada, una o varias capas ocultas y una capa de salida. Las conexiones se establecen entre los nodos de capas adyacentes. La capa de entrada, mediante la cual se presentan los datos a la red, está formada por nodos de entrada que reciben la información directamente del exterior. La capa de salida representa la respuesta de la red a una entrada dada, siendo esta información transferida al exterior. Las capas ocultas o intermedias se encargan de procesar la información y se interponen entre las capas de entrada y salida y son las únicas que no tienen conexión con el exterior.

La estructura de red más habitual es la denominada red de alimentación hacia delante o *feedforward*, ya que las conexiones entre neuronas se establecen en un único sentido, por el consiguiente orden: capa de entrada, capa(s) ocultas(s) y capa de salida. Por ejemplo en la figura 2.1 se muestra una red con dos capas ocultas. No obstante, existen también redes retroalimentadas o *feedback*, que pueden tener conexiones hacia atrás, es decir, de nodos de una capa a elementos de proceso de capas anteriores, así como redes recurrentes, que pueden poseer conexiones, tanto entre neuronas de una misma capa, como de nodo a sí mismo. La figura 2.2 ilustra un modelo de red en que coexisten los distintos tipos de conexiones que hemos señalado, es decir, hacia delante, hacia atrás y recurrentes, mostrando una interconexión total.

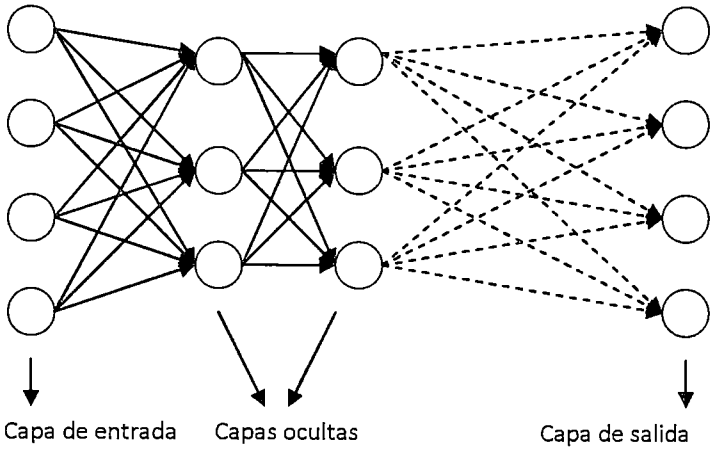


Figura 2.1 Red neuronal feedforward de dos capas

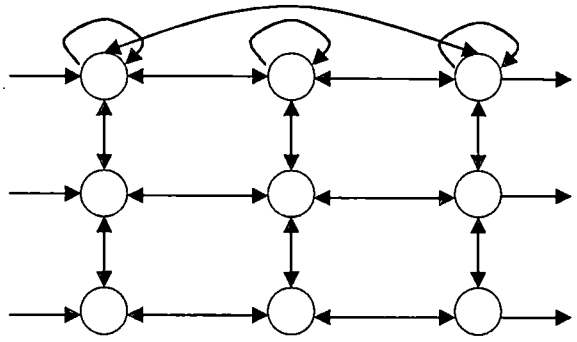


Figura 2.2 Red neuronal con interconexión total

b) Función de salida y funciones de transferencia o activación

La red neuronal totalmente interconectada es aquella en la que los nodos de cada capa están conectados con todos los nodos de la capa siguiente. La capa de entrada tiene por única misión la de distribuir la información que se le presenta a la red neuronal para el procesamiento de la capa siguiente. Los nodos de las capas ocultas y de la capa de salida procesan las señales aplicando factores de

procesamiento, llamados pesos. Cada capa tiene un nodo adicional llamado sesgo (bias), que añaden un término adicional a la salida de todos los nodos de la capa. Todas las entradas de un nodo son ponderadas, combinadas y procesadas a través de una función, llamada "función de transferencia" o "función de activación" que controla el flujo de salida de ese nodo para conectar con todos los nodos de la capa siguiente. Esta función de transferencia sirve para normalizar la salida.

Una red neuronal artificial no es más que la conexión de varias neuronas. Así, las neuronas artificiales, denominadas también unidades, nodos o elementos de proceso, constituyen la unidad básica de una red neuronal (análoga a la neurona biológica). Dichas neuronas artificiales operan a modo de microprocesadores simples, cuya función consiste en dar respuesta a un determinado patrón de entrada. Cada elemento de proceso, al igual que ocurre en una neurona biológica, recibe entradas procedentes de otros nodos vecinos, o del exterior, en el caso de la capa de entrada y su función consiste en transformar, mediante sencillos cálculos internos, dichas entradas en un solo valor de salida que envía al resto de nodos (constituyéndose la entradas de éstos) o bien, al exterior, si la neurona en cuestión pertenece a la capa de salida. Las conexiones entre elementos de proceso llevan asociadas un peso o una fuerza de conexión *W* que determina cuantitativamente el efecto que producen unos elementos sobre otros. Es decir, en los pesos se almacena la información de la red, al igual que sucede en las redes de neuronas biológicas.

El que una entrada tenga un efecto excitatorio o inhibitorio, depende de que el signo de peso correspondiente sea, respectivamente,

positivo o negativo. La efectividad de las entradas está determinada por la fuerza de la conexión, representada por el valor absoluto de los pesos. Así, cada uno de los elementos W_{ij} de la matriz de pesos *W*, conocida como patrón de conexiones, representa la intensidad y sentido de la relación del elemento del proceso *j*, con respecto al elemento del proceso *i*.

El proceso de transformación de las entradas en salidas, en una red neuronal artificial alimentada hacia adelante, con r entradas, una única capa oculta, compuesta de q elementos de proceso, y una unidad de salida puede resumirse en la siguiente formulación de la "función de salida de la red":

$$\hat{f}(x, W) = F(\beta_0 + \sum_{j=1}^{q} \beta_j G(x'\gamma_j)$$

Donde, $\hat{f}(x, W)$ es la salida de la red, el vector $x = (1, x_1, x_2, \ldots, x_r)'$ representa las entradas de la red (el uno se corresponde con el sesgo de un modelo tradicional), $\gamma_j = (\gamma_{j0}, \gamma_{j1}, \ldots, \gamma_{ji}, \ldots, \gamma_{jr})' \in \Re^{r+1}$ son los pesos de las neuronas de la capa de entrada a las de la intermedia u ocultas, β_j, j = 0, ... , q, representa la fuerza de conexión de las unidades ocultas a las de salida (j = 0 indexa la unidad sesgo), q es el número de unidades intermedias, es decir, el número de nodos de la capa oculta, $F: \Re \to \Re$ es la función de activación de la unidad de salida y $G: \Re \to \Re$ se corresponde con la función de activación de las neuronas intermedias, *W* es un vector que incluye todos los pesos de la red, es decir, γ_j y β_j. La Figura 2.3 representa la función $\hat{f}(x, W)$

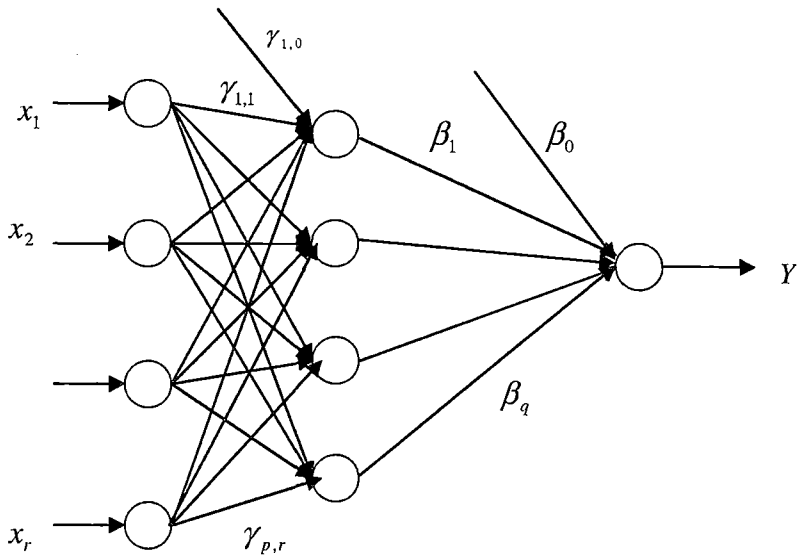


Figura 2.3 La función salida de la red $\widehat{f}(x, W)$

Históricamente se emplearon como funciones de activación "funciones umbral", cuyo efecto es que las unidades se activan bruscamente, esto es, o no se activan, o se activan de golpe. La respuesta puede ser solo blanco o negro, por ello, estas funciones son adecuadas para tareas de clasificación o reconocimiento. Con el tiempo, se introdujeron funciones de activación que permiten que las neuronas se activen gradualmente a medida que el nivel de actividad de sus entradas aumenta, en lugar que su estado pueda ser, únicamente, activación-desactivación. En concreto, la función que se propone es la tangente hiperbólica:

$$G(a) = \frac{2}{1+e^{(-2a)}} - 1$$

que produce una respuesta alisada.

En general las funciones *F* y *G* pueden adoptar cualquier forma en la expresión de $\hat{f}(x,W)$. Ahora bien, es práctica habitual considerar, bien que la función de activación de las neuronas de salida y de las intermedias es idéntica, *F*(*a*) = *G*(*a*), y que se corresponde con la función tangente hiperbólica, o bien, que *F*(*a*) = *a*, es decir, que es la función identidad y que *G*(*a*) se corresponde con la función tangente hiperbólica, lo que es equivalente a considerar que sólo existe función de activación en las unidades ocultas.

Suponiendo, que sólo existe función de activación en las neuronas intermedias, tenemos:

$$\hat{f}(x,W) = \beta_0 + \sum_{j=1}^{q} \beta_j G(x'\gamma_j)$$

Otra posibilidad, de gran utilidad en aplicaciones econométricas, es considerar que en la red que presentamos, una red neuronal artificial alimentada hacia delante, con *r* entradas, una única capa oculta, compuesta de *q* elementos de proceso, y una unidad de salida, también existen conexiones directas entre la capa de entrada y la de salida. En este caso, la salida de la red se obtiene mediante la siguiente expresión:

$$\hat{f}(x,W) = x'\alpha + \beta_0 + \sum_{j=1}^{q} \beta_j G(x'\gamma_j)$$

Donde α es un vector de dimensión $r \times 1$ que representa los pesos de las conexiones directas entre las capas de entrada y salida. Como es lógico, ahora *W*, que recoge la totalidad de pesos de la red, se compone de α, γ_j y β_j.

c) Redes Neuronales y Ajuste de Modelos de Regresión

La función $\hat{f}(x, W)$ podemos descomponerla en dos partes. La primera de ellas, que se corresponde con los dos primeros términos, representa un modelo lineal, de manera que, si tomamos como variables de entrada *r* retardos de la variable *x*, se convierte en una regresión lineal sobre variables de entrada retardadas, que actúan como variables explicativas, y una constante (β_0).

Esta primera parte, como es lógico, capta las dependencias lineales entre los patrones de entrada y las de salida de la red. La segunda parte, que es el tercer término de la formulación anterior, recoge, en caso de que existan, las dependencias no lineales entre las variables de entrada y la salida de la red, dado que la función empleada es no lineal. Concretando, este tercer término es una composición, ponderada con los pesos sinápticos de las neuronas intermedias a las de salida (β_j), de funciones tangentes hiperbólicas de las entradas de la red, estas últimas, por la fuerza de conexión de las unidades de entrada a las intermedias. Este modelo puede considerarse una extensión de los conocidos y, tan frecuentemente utilizados, modelos lineales, ya que se compone de un modelo lineal, aumentado con términos no lineales.

Como podemos apreciar, la red que describimos mediante la expresión de $\hat{f}(x, W)$, goza de tal grado de flexibilidad que permite ajustar todo tipo de funciones, por ello se caracteriza a las redes neuronales artificiales como "aproximadores universales". Es decir, *una red neuronal es capaz de aprender cualquier función*. El modelo de redes neuronales artificiales debe considerarse como uno más

dentro del conjunto de los no paramétricos, al que se pueden aplicar los resultados de la inferencia estadística.

d) Aprendizaje en las Redes Neuronales

Después de diseñar una red neuronal artificial, lo que pretendemos conseguir con la misma es que, para ciertas entradas, o patrones ejemplo que suministramos a la red, ésta sea capaz de generar una salida deseada. Para ello, además de que la topología de la red (entendida como la estructura de la red) sea adecuada, se requiere que la misma aprenda a proporcionar soluciones correctas, es decir, es necesario someter a la red a un proceso de aprendizaje o entrenamiento. El aprendizaje puede entenderse como un procedimiento de prueba y error que permite la estimación estadística de los parámetros del modelo de red neuronal empleado.

Suelen considerarse tres tipos básicos de aprendizaje que dan lugar a diferentes tipos de redes neuronales. Cuando el entrenador proporciona a la red la salida deseada, se dice que el *aprendizaje* es *supervisado*. En caso contrario, nos encontramos ante un *aprendizaje no supervisado*. Por último, un tipo intermedio de *aprendizaje* es el *reforzado o híbrido*, en el cuál el entrenador sólo proporciona a la red una indicación de si la respuesta a una entrada es buena o mala.

Las redes neuronales con *aprendizaje supervisado*, que suelen venir asociadas al *perceptrón multicapa* (*Mutilayer Perceptron* MLP), presentan un patrón de salida o variable dependiente que les permite contrastar y corregir los datos. Las redes neuronales con patrón de salida suelen ser técnicas utilizadas tanto para la clasificación como

para la predicción. Con ello se puede segmentar mercados, posicionar productos, realizar previsiones de demanda, evaluaciones de expedientes de crédito o de análisis del valor de la bolsa y un sinfín de aplicaciones más. El modelo de red neuronal *perceptrón multicapa* se fundamenta en el *aprendizaje por retropropagación del error* (*Back-Propagation*). Y utiliza habitualmente el *algoritmo por retropropagación*, el *algoritmo del gradiente descendente* (*conjugate gradient descent*) y el algoritmo de *Levenberg-Marquardt*.

Formalmente, el proceso del aprendizaje consiste en resolver un problema de mínimos cuadrados no lineales. Para ello, hay que emplear métodos numéricos de optimización como el de *retropropagación de errores* (*Back-Propagation*), que se fundamenta en el algoritmo de aproximación estocástica de Robbins y Moro (1951) aplicados a mínimos cuadrados no lineales.

Una vez finalizado el aprendizaje se debe proceder a testear la red. La fase de test consiste en introducir nuevos patrones de entrada y comprobar la eficacia del sistema generado. Si no resulta aceptable se repite la fase de entrenamiento utilizando nuevos patrones, e incluso puede ser necesario modificar la estructura de la red.

e) Funcionamiento de una Red Neuronal

Para la creación y aplicación de una red neuronal a un problema concreto, hemos de distinguir los siguientes pasos:

Conceptualización del modelo para el estudio del problema concreto. En este Modelo debemos señalar las entradas, las salidas y la información de que se dispone.

Adecuación de la información de que se dispone a la estructura de la red a crear. Es decir se construirán los patrones de aprendizaje, que es parte de la información que va a ser utilizada para el entrenamiento o aprendizaje de la red, y los patrones de validación, que es parte de la información que va a ser utilizada como validación de la red.

Fase de aprendizaje. Se le va presentando a la red los patrones adecuados y la red va proporcionando una salida, este proceso se repite un cierto número de etapas, estas salidas se comparan con las salidas esperadas y los diversos algoritmos de aprendizaje intentan minimizar el error que hay entre la salida proporcionada por la red y la salida esperada.

Fase de validación. Se presentan a la red entrenada el conjunto de patrones de validación, y se ve el error cometido por la red en este conjunto, este error es una medida de la bondad de la red.

Fase de generalización. Si hemos conseguido una red adecuada se procede a utilizar la red como modelo predictor, aportándole una nueva entrada, la red la procesará y dará una salida.

f) El Algoritmo de Aprendizaje de Retropropagación

El proceso de aprendizaje o entrenamiento de la red consiste en ir presentando a la red el conjunto de patrones, un determinado número de etapas prefijadas de antemano, de forma a minimizar el error de aprendizaje, entendiendo éste como la diferencia cuadrática entre la salida esperada y la salida que aporta la red. En la primera etapa, la red tienen unos pesos de interconexiones elegidos de forma

aleatoria, a la red se le presenta un vector de entrada en la primera etapa, constituido por el primer patrón, éste se va propagando a través de todas las capas hasta proporcionar una salida, la señal de salida se compara con la salida deseada en todos los nodos de la capa de salida. Este proceso se realiza para todos los patrones del conjunto de aprendizaje, y la suma de errores cuadráticos de todos los patrones será el error cometido por la red en esa primera etapa.

El objetivo es ir cambiando o actualizando para la segunda etapa los pesos de interconexiones de forma a disminuir el error total. La idea del *algoritmo back-propagation* consiste en actualizar los pesos de interconexión de forma que la señal de error se transmita hacia atrás partiendo de la capa de salida; sin embargo estas unidades intermedias sólo reciben una fracción de error proporcional a la contribución relativa que haya aportado a la salida. Este proceso se repite capa por capa hasta que todos los nodos hayan recibido una señal de error que describa su contribución al mismo. Una vez que hemos actualizado los pesos, se repite el proceso de presentar de nuevo los patrones de aprendizaje y el cálculo del error, este proceso acaba bien porque el error total es menor que uno prefijado, bien porque hemos concluido con el número de etapas prefijado. La importancia de este proceso radica en que a medida que se entrena la red, los nodos de las capas ocultas aprenden a reconocer distintas características del problema.

Para realizar la descripción matemática del algoritmo en una red con tres capas utilizaremos la siguiente notación:

o_i = Salida del nodo *i* de la primera capa.

w_{ij} = Peso de conexión entre el nodo *i* de la primera capa y el nodo *j* de la capa oculta.

net_j = Entrada neta del nodo *j* de la capa oculta. $net_j = \sum_i w_{ij} o_i$

o_j = Salida del nodo *j* de la capa oculta. $o_j = \frac{1}{1+\exp(-net_j)}$

w_{jk} = Peso de la conexión entre el nodo *j* de la capa oculta y el nodo *k* de la capa final.

net_k = Entrada neta del nodo *k* de la capa oculta. $net_k = \sum_j w_{jk} o_j$

o_k = Salida del nodo *k* de la capa oculta. $o_k = \frac{1}{1+\exp(-net_k)}$

t_k = Salida esperada en el nodo *k* de la capa final.

Para un patrón determinado *p* la salida vendrá dada por o_{pk} y la salida esperada por t_{pk}. El error de toda la red vendrá dado por:

$$E = \frac{1}{2}\sum_p \sum_k (t_{pk} - o_{pk})^2$$

El objetivo de la *back-propagation* es determinar el conjunto de pesos (w_{ij} , w_{jk}), que hagan mínimo el error cuadrático de la red. El algoritmo comienza por un conjunto de pesos arbitrarios y se va actualizando en cada etapa de acuerdo con la siguiente regla:

1. En primer lugar los pesos de la capa final, w_{jk} mediante la técnica del gradiente descendente

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial o_k} \cdot \frac{\partial o_k}{\partial net_k} \cdot \frac{\partial net_k}{\partial w_{jk}} = -(t_k - o_k) o_k (1 - o_k) o_j ,$$

de forma que w_{jk} se actualiza con una tasa de aprendizaje negativa $(-\eta)$, con lo cual w_{jk} actualizado es

$$w_{jk}^{*} = w_{jk} + (-\eta)[-(t_k - o_k)o_k(1-o_k)o_j]$$

2. La actualización de los pesos correspondientes a la capa oculta son:

$$\frac{\partial E}{\partial w_{ij}} = \sum_k \frac{\partial E}{\partial o_k} \cdot \frac{\partial o_k}{\partial net_k} \cdot \frac{\partial net_k}{\partial o_j} \cdot \frac{\partial o_j}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}}$$

$$= \sum_k -(t_k - o_k)o_k(1-o_k)w_{jk}o_j(1-o_j)o_i$$

A veces se añade a la actualización un término momento, con lo cual se acelera el proceso de actualización de los pesos.

2.1.2.3 Regresión Logística

La regresión logística es un instrumento estadístico de análisis bivariado o multivariado, de uso tanto explicativo como predictivo. Se emplea cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de *p* variables predictoras o independientes, que pueden ser cuantitativas (denominadas covariables) o categóricas.

El propósito del análisis de regresión logística es:

- Predecir la probabilidad de que a alguien le ocurra cierto evento, como por ejemplo en nuestro caso de estudio: "aprobar un curso" =1 o "desaprobarlo"= 0.

- Determinar qué variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda cierto evento.

La probabilidad se expresa mediante la función de distribución logística:

$$P_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_p X_P)}} \qquad (1)$$

Para facilidad de la explicación, haremos $Z_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

Es fácil verificar que a medida que Zi se encuentra dentro de un rango de -∞ a +∞, P_i se encuentra dentro de un rango de 0 a1, y que P_i no está relacionado linealmente con Zi (es decir con los Xi).

Si por ejemplo Pi es la probabilidad de aprobar un curso. Entonces (1-Pi), la probabilidad de no aprobar es:

$$1 - P_i = \frac{1}{1+e^{Zi}}$$

Por consiguiente podemos escribir: $\frac{P_i}{1-P_i} = \frac{1+e^{Zi}}{1+e^{-Zi}} = e^{Zi}$

que es sencillamente la razón de las probabilidades, para nuestro caso, la razón de la probabilidad de que un alumno apruebe un curso, respecto de la probabilidad que no lo apruebe.

Si tomamos el logaritmo natural:

$$L_i = \ln\left(\frac{P_i}{1-P}\right)_i = Z_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Vemos que Li, el logaritmo de la razón de las probabilidades, es lineal con las Xi. Li se llama logit, de aquí el nombre de modelo logit.

La determinación de los coeficientes es realizada a través del método de máxima verosimilitud. Estos coeficientes son interpretados en términos de <<odd-ratios>>, y la selección de variables puede realizarse mediante tres métodos: <<forward>>, <<backward>> o <<stepwise>>. El método <<stepwise>> es el más comúnmente utilizado. En él se selecciona una variable en cada paso para ser incluida o excluida del modelo según criterios estadísticos.

Estimación de máxima verosimilitud para el modelo de regresión logística

El cálculo de la probabilidad, por ejemplo, de que un alumno apruebe un curso, se realiza a partir de la Ecuación (1).

En realidad no observamos P_i, sino sólo el resultado $Y=1$, si un alumno aprueba, y $Y=0$ si no aprueba.

Como cada Y_i es una variable aleatoria Bernoulli, se expresa

$$\Pr(Y_i = 1) = P_i \tag{2}$$

$$\Pr(Y_i = 0) = (1 - P_i) \tag{3}$$

Supongamos que tenemos una *muestra aleatoria* de *n* observaciones. Sea la función $f_i(Y_i)$ tal que denote la probabilidad de que $Y_i = 1$ o 0; la probabilidad conjunta de observar los *n* valores Y, es decir, $f(Y_1, Y_2, \ldots, Y_n)$ se expresa como:

$$f(Y_1, Y_2, \ldots, Y_n) = \prod_1^n f_i(Y_i) = \prod_1^n P_i^{Y_i}(1 - P_i)^{1-Y_i} \tag{4}$$

donde $\prod$ es el operador producto; observe que escribimos la función de densidad de probabilidad conjunta como producto de las funciones de densidad individuales, pues cada Y_i se obtiene de manera independiente y cada Y_i tiene la misma función de densidad (logística). La probabilidad conjunta dada en la ecuación (4) se conoce como **función de verosimilitud (FV)**.

Es un poco difícil manipular la ecuación (4). Pero si tomamos su logaritmo natural, obtenemos lo que se conoce como **función log de verosimilitud (FLV)**.

$$\begin{aligned}\ln f(Y_1, Y_2, \ldots, Y_n) &= \sum_1^n [Y_i \ln P_i + (1 - Y_i)\ln(1 - P_i)] \\ &= \sum_1^n [Y_i \ln P_i - Y_i \ln(1 - P_i) + \ln(1 - P_i)] \\ &= \sum_1^n \left[Y_i \ln\left(\frac{P_i}{1 - P_i}\right)\right] + \sum_1^n \ln(1 - P_i)\end{aligned} \tag{5}$$

De (1) resulta fácil comprobar que

$$(1 - P_i) = \frac{1}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}} \tag{6}$$

así como

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X \tag{7}$$

Mediante (6) y (7) expresamos la FLV (5) como:

$$\ln f(Y_1, Y_2, \ldots, Y_n) =$$

$$= \sum_{1}^{n} \left[Y_i (\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p) \right] - \sum_{1}^{n} \ln \left[1 + e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)} \right] \qquad (8)$$

Como se observa de (8), la función log de verosimilitud es una función en los parámetros $\beta_0, \beta_1, \ldots, \beta_p$, pues las $X_1, \ldots, X_p$ se conocen.

En MV, el objetivo consiste en maximizar la FV (o la FLV), es decir, en obtener los valores de los parámetros desconocidos de forma que la probabilidad de observar las Y dadas sea tan grande (máximo) como sea posible. Con este propósito, diferenciamos (8) parcialmente respecto de cada incógnita, igualamos las expresiones resultantes a cero y resolvemos las expresiones así obtenidas. Luego aplicamos la condición de maximización de segundo orden a fin de verificar que los valores de los parámetros obtenidos en verdad maximicen la FV.

2.1.2.4 Regresión Múltiple

La *Regresión múltiple* tiene como objetivo analizar un modelo que pretende explicar el comportamiento de una variable explicada (variable endógena o dependiente), que designaremos como *Y*, utilizando la información proporcionada por los valores tomados por un conjunto de variables explicativas (exógenas o independientes), que designaremos por $X_1, X_2, \ldots, X_k$.

El modelo lineal (modelo econométrico) viene dado de la forma:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + u$$

Los coeficientes (parámetros) b_1, b_2,..., b_k denotan la magnitud del efecto que las variables explicativas (exógenas o independientes) X_1, X_2,..., X_k tienen sobre la variable explicada (endógena o dependiente) Y. El coeficiente b_0 se denomina término constante del modelo. El término u se denomina término de error del modelo.

Disponemos de un conjunto de T observaciones para cada una de las variables endógena y exógenas. Entonces, podremos escribir el modelo de la forma:

$$Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \ldots + b_k X_{kt} + u_t \qquad t = 1, 2, 3, \ldots, T.$$

La aparición (no necesaria) de un término constante en el modelo puede interpretarse como la presencia de una primera variable X_0 cuyo valor sea siempre 1.

El problema fundamental que se aborda es el siguiente: suponiendo que la relación entre la variable Y con el conjunto de variables X_1, X_2,..., X_k es como se ha descrito en el modelo, y que se dispone de un conjunto de T observaciones para cada una de las variables, la endógena y las exógenas, ¿cómo pueden asignarse valores numéricos a los parámetros b_0, b_1, b_2,..., b_k basándose en la información muestral? Estos valores se llamarán estimaciones de los parámetros.

Una vez encontradas las estimaciones de los parámetros del modelo, podremos hacer predicciones acerca del comportamiento futuro de la variable *Y*.

Formularemos el modelo lineal bajo las siguientes hipótesis:

- Las variables X_1, X_2,..., X_k son deterministas (no son variables aleatorias), ya que su valor es un valor constante proveniente de una muestra tomada.

- La variable *u* (término de error) es una variable aleatoria con esperanza nula y matriz de covarianzas constante y diagonal (matriz escalar). Es decir que, para todo *t*, la variable u_t tiene medida cero y varianza σ^2 no dependiente de *t*, y además $Cov(u_i, u_j) = 0$ para todo *i* y para todo *j* distintos entre sí. El hecho de que la varianza de u_t sea constante para todo *t* (que o depende de *t*), se denomina hipótesis de *homoscedasticidad*. El hecho de que $Cov(u_i, u_j) = 0$ para todo i distinto de j se denomina hipótesis de *no autocorrelación*.

- La variable *Y* es aleatoria, ya que depende de la variable aleatoria *u*.

- También se supone la ausencia de errores de especificación, es decir, que suponemos que todas las variables *X* que son relevantes para la explicación de la variable *Y*, están incluidas en la definición del modelo lineal.

- Las variables X_1, X_2,..., X_k son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Esta hipótesis se

denomina hipótesis de *independencia*, y cuando no se cumple, decimos que el modelo presenta *multicolinealidad.*

- A veces también se considera la hipótesis de normalidad de los residuos, consistente en que las variables u_t sean normales para todo *t*.

La estimación del modelo lineal de regresión múltiple se basa en el criterio de mínimos cuadrados, que considera que la función que mejor se ajusta a los datos es la que minimiza la varianza del error "e", lo que es equivalente a minimizar:

$$S(b_0,b_1,...,b_k = \sum_{t=1}^{T} e_t^2 Y_t = \sum_{t=1}^{T} (y_1 - (b_0 + b_1 X_{1t} + b_2 X_{2t} + \ldots + b_k X_{kt}))^2$$

Que al derivarla respecto a los parámetros b_0, b_1,...,b_k, e igualando a cero, nos da un sistema de (k+1) ecuaciones lineales, que al resolverla nos da los valores de b_0, b_1,...,b_k.

2.2 Metodología de la Investigación

2.2.1 Tipo de Investigación

La presente investigación es propositiva ya que trata de ayudar a los estudiantes a tomar mejores decisiones al inscribirse en un nuevo semestre académico.

Es experimental, pues utilizará un conjunto de datos correspondientes a la base histórica Académica de notas de la Facultad de Ingeniería Química y Textil, pertenecientes a todos los alumnos de la especialidad de Ingeniería Química. Estos datos servirán para la construcción de los modelos

predictivos basados en redes neuronales, regresión logística y regresión múltiple.

2.2.2 Población y Muestra

2.2.2.1 Población

Se tomará como población a los estudiantes de la especialidad de Ingeniería Química de la Facultad de Ingeniería Química y Textil de la Universidad Nacional de Ingeniería, que reportan haber llevado cursos entre 1993 y 2010.

2.2.2.2 Muestra

Se tomará a toda la población, debido al número no tan alto de estudiantes por semestre académico.

2.2.3 Técnicas de procesamiento

Para el tratamiento de los datos se realizará un programa en Lenguaje JAVA, para obtener los valores de las variables (dependiente e independientes) a partir de la base histórica académica de notas de la Facultad de Ingeniería Química y Textil, correspondiente a todos los alumnos de la especialidad de Ingeniería Química.

Para obtener los modelos predictivos se usará para Redes Neuronales el Solver del Excel, para Regresión Logística el Software SPSS y para Regresión Múltiple la herramienta de análisis de datos "regresión" del Excel.

2.2.4 Análisis y Tratamientos de Datos

El objetivo de la metodología consiste en construir un modelo a partir de los datos históricos de los estudiantes que sea capaz de representar la realidad de manera precisa. Este modelo permitirá la predicción del rendimiento

académico del estudiante en un curso en que se desea inscribir, en base a su propio rendimiento académico y al rendimiento académico de otros estudiantes con perfiles similares que hayan llevado el curso. Se realizará la aplicación para la especialidad de Ingeniería Química de la Facultad de Ingeniería Química y Textil de la UNI.

2.2.4.1 Descripción de los datos de entrada

Los datos de partida para este estudio son:

a) La Base histórica académica de notas de la Facultad de Ingeniería Química y Textil, correspondientes a todos los estudiantes de las especialidades de Ingeniería Química e Ingeniería Textil desde el año 1993 hasta el 2010 (un extracto de estos datos se da en tabla 2.1).

b) El Currículo o Plan de Estudios de la Especialidad de Ingeniería Química

El currículo actual de la Especialidad de Ingeniería Química (que se muestra en la Anexo 1) consta de:

- 63 cursos obligatorios (191 créditos)
- 19 cursos electivos de la especialidad (57 créditos)
- 06 cursos electivos complementarios (21 créditos)

De los cuales se le exige al estudiante para completar sus estudios:

- Número total de créditos obligatorios: 191
- Número mínimo de créditos electivos: 20
- Número mínimo de créditos electivos de la especialidad: 10

2.2.4.2 Selección de datos

a) Extracción del listado de cursos con sus pre-requisitos

En primer lugar se puede extraer del currículo la relación de cursos con sus pre-requisitos, que se presentan en la tabla 2.2.

Debido a que en el presente estudio se utilizan los datos históricos académicos desde el año 1993 hasta el 2010, es necesario tomar en cuenta los cambios habidos en el currículo de la especialidad de Ingeniería Química. Solo se han agregado al currículo en el período académico 2005-2 los cursos de Gestión Ambiental Empresarial (PI912) y Gas Natural y Condensados (PI824).

b) Extracción de datos solo para los estudiantes de la especialidad de Ingeniería Química

La Base histórica académica de notas de la Facultad de Ingeniería Química y Textil, es aquella que maneja la FIQT a partir de los datos proporcionados por ORCE. Esta Base histórica de notas se encuentra en archivo Excel y consta de 9 campos y 207191 registros (ver Tabla 2.3), dispuestos en el siguiente orden:

1. Código de alumno (CODALU)
2. Especialidad (ESPEC)
3. Código del Curso (CODCUR)
4. Sección (SEC)
5. Período Académico (PERACD)
6. Número de créditos (CRD)
7. Nota del Curso (NOTA)
8. Situación (SIT)
9. Condición del alumno (CND)

CODALU	ESPEC	CODCUR	SEC	PERACD	CRD	NOTA	SIT	CND
19721280F	Q1	PI523	B	2009-1	4	10.3	A	N
19721280F	Q1	EC618	B	2009-2	5	4.1	D	N
19721280F	Q1	EP307	B	2009-2	4	10.6	A	N
19721280F	Q1	PI140	C	2009-2	3	0.9	D	N
19721280F	Q1	QU526	B	2009-2	2	10.6	A	N
19721280F	Q1	QU527	A	2009-2	1	11.3	A	N
19721280F	Q1	EC618	A	2010-1	5	11.9	A	N
19721280F	Q1	PI140	A	2010-1	3	8.3	D	N
19721280F	Q1	PI140	A	2010-2	3	10.3	A	N
19721280F	Q1	PI513	B	2010-2	2	6.2	D	N
19740377A	Q1	PI135	6	1993-1	2	0.0	D	N
19740377A	Q1	PI345	6	1993-1	2	0.0	D	N
19740377A	Q1	PI531	6	1993-1	3	0.0	D	N
19740377A	Q1	PI135	6	1993-2	2	7.1	D	N
19740377A	Q1	PI345	6	1993-2	2	0.0	D	N
19740377A	Q1	PI531	6	1993-2	3	6.3	D	N
19740377A	Q1	PI135	6	1997-1	2	1.1	D	N
19740377A	Q1	PI515	6	1997-1	3	0.5	D	N
19740377A	Q1	PI826	6	1997-1	3	1.3	D	N
19740377A	Q1	PI911	6	1997-1	4	2.0	D	N
19740377A	Q1	PI135	7	1997-2	2	1.3	D	N
19740377A	Q1	PI826	6	1997-2	3	0.5	D	N
19740377A	Q1	PI911	6	1997-2	4	0.7	D	N
19740954I	Q1	PI142	6	1993-1	3	5.8	D	N
19740954I	Q1	PI318	6	1993-1	5	10.5	A	N
19740954I	Q1	QU511	7	1993-1	4	10.1	A	N
19744105F	Q1	PA136	6	1993-1	4	1.0	D	N
19744105F	Q1	PI135	6	1993-1	2	8.6	D	N
19744105F	Q1	PI136	6	1993-1	2	7.5	D	N
19744105F	Q1	PI415	7	1993-1	3	10.1	A	N
19744105F	Q1	PI425	7	1993-1	4	7.2	D	N
19744105F	Q1	PI510	6	1993-1	3	9.3	D	N

Tabla 2.1 Extracto del archivo Excel de la base histórica académica de notas de la Facultad de ingeniería Química y Textil

Nº	CODIGO CURSO	PRE-REQUISITOS
1	AU511	Ninguno
2	FI203	Ninguno
3	MA113	Ninguno
4	MA114	Ninguno
5	PI100	Ninguno
6	PI118	Ninguno
7	QU116	Ninguno
8	QU117	Ninguno
9	EM711	AU511
10	FI204	FI203
11	MA123	MA113
12	MA124	MA114
13	MA713	MA113 MA114
14	QU118	QU116
15	QU119	QU116 QU117
16	EP307	MA124
17	FI403	FI204
18	MA133	MA123
19	QU214	QU118 QU119
20	QU215	QU118 QU119
21	EE102	FI403
22	FI152	FI403
23	MA143	MA133 MA713
24	MA612	MA133
25	QU425	MA133 QU118
26	QU426	MA133 QU215
27	PI111	QU425
28	PI523	MA143
29	QU324	QU118 QU426
30	QU325	QU118 QU426
31	QU434	QU425 QU426
32	QU435	QU425 QU426
33	QU516	QU214 QU426
34	QU517	QU214 QU426
35	EC618	FI403
36	PA714	MA612
37	PI140	MA143 PI111
38	PI216	PI111 QU434
39	QU334	QU324 QU325
40	QU335	QU324 QU325
41	QU526	QU516 QU517
42	QU527	QU516 QU517
43	PA113	MA612
44	PI142	PI140
45	PI217	PI216 PI523

Tabla 2.2 Listado de cursos y sus pre-requisitos

Nº	CODIGO CURSO	PRE-REQUISITOS
46	PI318	PI140 QU334
47	PI513	QU526
48	EP818	EP307
49	PI143	PI142
50	PI144	PI142
51	PI146	PI142
52	PI515	PI513
53	PI135	PI143 PI146
54	PI225	PI217
55	PI415	EE102 PI144
56	PI510	EP818 PI144
57	PI612	PI143 PI144
58	PI911	EP307
59	AHD65	Ninguno
60	PA136	PA113 PA714
61	PI136	PI135 PI144
62	PI426	PI225 PI415
63	PI525	PI510 PI415
64	PI322	QU434
65	PI355	PI318 QU526
66	PI365	PI144 PI318
67	PI366	PI365
68	PI516	PI515
69	PI826	PI146
70	PI147	PI144
71	PI226	PI225
72	PI345	PI318
73	PI376	PI135 PI144
74	PI475	PI144 PI143
75	PI531	PI144 PI318
76	PI613	PI612
77	PI381	PI143 PI510
78	PI721	QU334 QU526
79	PI722	PI721
80	PI823	PI225
81	PI912	PI318
82	PI824	PI217 PI142
83	ME425	QU434
84	HC443	PI318
85	QU565	QU526
86	SA633	PA113
87	PA425	PI510
88	PA515	EP307

Tabla 2.2 Listado de cursos y sus pre-requisitos (continuación)

En este archivo Excel el Campo CODALU (Código del alumno) corresponde a un total de 3145 estudiantes de la FIQT. El Campo Especialidad, contiene dos tipos de elementos, que son Q1 y Q2, y que corresponden a las dos carreras profesionales que ofrece la FIQT y que son Ingeniería Química e Ingeniería Textil respectivamente. Y como nuestra metodología consiste en predecir en base a estudiantes que tengan los mismos perfiles, seleccionamos solo a los de la especialidad de Ingeniería Química, a pesar que existen cursos comunes a ambas especialidades, sobre todo en los primeros ciclos de ambas carreras. Retirando a los estudiantes de Ingeniería Textil, nos quedamos con un nuevo Archivo Excel que contiene 149825 registros correspondientes a 2230 estudiantes de la especialidad de Ingeniería Química.

2.2.4.3 Limpieza de datos

Empezaremos examinando los campos e iremos descartando los datos atípicos. El campo Período Académico está compuesto por los períodos de estudio regulares, que se dan entre marzo y diciembre (140626 registros), períodos de estudios acelerados, que se dan en los meses de verano (9171 registros) y período de estudios anuales (28 registros).

Los períodos académicos de verano son muy diferentes a los regulares, primero porque son acelerados y segundo porque tiene una reglamentación especial, ya que si el alumno desaprueba en ese período, automáticamente se le retira del curso. Esto hace que no pueda ser considerado en nuestro estudio. De igual forma los períodos académicos anuales, tenían todo un enfoque diferente (desaparecía el currículo flexible), sobre todo en la reglamentación

para pasar de un ciclo de la carrera al siguiente. Es por ello que tampoco consideraremos a los períodos académicos anuales en nuestro estudio.

El campo Condición, se refiere a la condición del estudiante en un período académico con respecto a un curso específico. Así tenemos cinco rubros diferentes:

Normal	: 135821 registros
Retiro Reglamentario	: 257 registros
Retiro Total	: 730 registros
Retiro Voluntario	: 3816
Retiro Reglamentario de Verano	: 02

Nos quedaremos solo con la Condición Normal, ya que los demás son diferentes formas de Retiros.

Finalmente examinemos el Campo Nota, que contiene los valores:

"0"	: 3261 registros
"30"	: 53 registros
"De 0.1 a 20"	: 132507 registros

Las Notas que corresponden a "30" se refieren a los cursos convalidados, y como no dan la nota verdadera, no los consideraremos para nuestro estudio.

Las notas "0" corresponden a los estudiantes que abandonaron completamente el curso sin realizar ningún tipo de retiro. Como este grupo de estudiantes tienen un comportamiento anómalo, no será considerado en este estudio.

Realizando los descartes o eliminaciones en la Hoja Excel, nos quedamos con 132507 registros que corresponden a 2211 estudiantes de la especialidad de Ingeniería Química.

Con ayuda de la opción "quitar duplicados" de "datos" en la Hoja Excel, eliminamos 132 duplicados del último archivo Excel mencionado.

Finalmente reemplazamos en la columna código de curso, los códigos antiguos de 11 cursos por su nuevo código del currículo actual. En total se realizaron 1858 reemplazos.

Con la eliminación de los duplicados y el reemplazo de códigos de cursos antiguos por los nuevos, el archivo Excel se reduce a 132375
registros, que corresponden a 2211 estudiantes de la especialidad de Ingeniería Química. A este archivo Excel le llamaremos Base Histórica de Notas Limpia.

CAMPO	SIGNIFICADO	ELEMENTOS DIFERENTES
CODALU	Código de alumno: Es una cadena de caracteres formada por 8 dígitos y una letra, que se le asigna a cada estudiante al momento de su ingreso a la UNI. Los cuatro primeros dígitos corresponden al año de ingreso del estudiante a la UNI.	Número de estudiantes
ESPEC	Especialidad: Es una cadena de 2 caracteres (la letra Q y un digito, que puede ser 1 ó 2) y que corresponden a las dos carreras profesionales que ofrece la FIQT.	Q1 : Ingeniería Química Q2: Ingeniería Textil
CODCUR	Código del Curso: Es una cadena de 5 caracteres formada por 2 ó 3 letras y 3 ó 2 dígitos, y que corresponden a los cursos que han llevado los estudiantes en la UNI.	Número de cursos
SEC	Sección: Es una letra ó un número de un digito, que corresponde a las diferentes secciones que puede tener un curso en un período académico.	Número de secciones
PERACD	Período Académico: Es una cadena de cinco dígitos, separados los cuatro primeros del último por un guión, y que corresponden a los dos períodos regulares (de marzo a diciembre) y al período acelerado que se dan en el verano. Los primeros cuatro dígitos corresponden al año en que se realizó y el último dígito es: 1 si se realiza entre marzo y julio. 2 si se realiza entre setiembre y diciembre, y 3 si se realiza entre enero y febrero.	Número de Períodos Académicos
CRD	Crédito: Es un número asignado a un curso, que representa una medida del número de horas de teoría y de práctica por semana que tiene el curso.	Del 1 al 10 sin incluir el 9
NOTA	Es la calificación final que obtiene el estudiante después de haber llevado un curso. Esta nota, que es un número con el primer decimal truncado, va de cero a veinte y existe la nota 30 para el caso de los cursos convalidados.	0, 0.1, 0.2,.....20.0 y 30
SIT	Situación: Corresponde a la situación del estudiante en un período académico con respecto a un curso específico.	A : Aprobado D: Desaprobado Vacío: Retirados por (R,T.V,X)
CND	Condición: Corresponde a la condición del estudiante en un período académico con respecto a un curso específico.	N: Normal R: Retiro reglamentario T: Retiro total V: Retiro voluntario X: Retiro reglamentario de verano

Tabla 2.3 Campos de la base histórica académica de notas

2.2.4.4 Descripción de la Base Histórica de Notas Limpia

En la Tabla 2.4 se muestran los diferentes elementos que conforman cada uno de los nueve campos de la Base Histórica de Notas Limpia. Es importante mencionar que solo 87 de los 160 cursos diferentes pertenecen al actual currículo de la especialidad, los demás pertenecen a currículos antiguos o a cursos de otras especialidades

CAMPO	ELEMENTOS DIFERENTES
CODALU	2211 estudiantes
ESPEC	Una sola Q1
CODCUR	171 cursos
SEC	27 secciones diferentes, a saber: De 8 digitos: 2, 3, 4, 5, 6, 7, 8, 9 De 19 letras: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, R, S, T
PERACD	66 Períodos Académicos regulares de 1975-1 a 2010-2
CRD	7 creditajes: 1, 2, 3, 4, 5, 6, 7
NOTA	198 valores que van de 0.1 a 20
SIT	Solo 2 , a saber A: Aprobado D: Desaprobado
CND	Una sola N: Normal

Tabla 2.4 Campos de la base histórica académica de notas limpia

CAPITULO III

DISEÑO DE LOS MODELOS DE PREDICCION

3.1 Elección de los Modelos Predictivos

Para poder verificar las cuatro hipótesis planteadas en el sección 1.8, debemos predecir si el estudiante aprueba o no un curso en que se desea inscribir en un nuevo ciclo, y también la nota que obtendrá el estudiante. Para el primer tipo de predicción, que es del tipo cualitativo o discreto, se empleará las técnicas de redes neuronales y de regresión logística; mientras que para el segundo tipo de predicción, que es del tipo cuantitativo o continuo, se empleará las técnicas de redes neuronales y de regresión múltiple. Esto se puede ver mejor en la tabla 3.1

OBJETIVO	TÉCNICA DE PREDICCIÓN
PREDICCIÓN DE SI APRUEBA O NO UN CURSO	REDES NEURONALES
	REGRESIÓN LOGÍSTICA
PREDICCIÓN DE LA NOTA QUE OBTENDRÁ EN UN CURSO	REDES NEURONALES
	REGRESIÓN MÚLTIPLE

Tabla 3.1 Técnicas de predicción empleadas

3.2 Selección de variables

En base a la experiencia de los autores en la enseñanza universitaria, se establecerán un conjunto de variables predictoras relacionadas a nuestro estudio, las que luego de una discusión se reducen a siete. Estas siete

variables son comparadas, con las utilizadas en estudios publicados relacionados al tema.

3.2.1 Algunas definiciones previas

Es importante realizar algunas definiciones para entender cada una de las variables:

Curso en estudio: Curso del cual se desea realizar la predicción.

Período académico ó semestre de estudio: Período académico en el cual se desea predecir.

Período académico ó semestre previo: Período académico anterior al período académico de estudio.

Curso pre-requisito: Curso que aparece en Plan de Estudio como pre-requisito del Curso en estudio.

Efectividad de aprobación 1: Es la relación entre el número de cursos aprobados y el número de cursos matriculados de un estudiante.

Efectividad de aprobación 2: Es la relación entre el número de créditos aprobados y el número de créditos matriculados de un estudiante.

Grado de Dificultad 1 de aprobación de un curso por período académico o Semestre: Es la diferencia de veinte menos el promedio de notas del curso en un Período Académico o Semestre.

Grado de Dificultad 2 de aprobación de un curso por período académico o Semestre: Es la relación del número de estudiantes desaprobados entre el número de estudiantes matriculados en un curso en un Período Académico o Semestre.

3.2.2 Clasificación de las variables

Las variables que se discutirán las podemos clasificar en cuatro grupos:

A) Variables relacionadas al rendimiento global del alumno

En estas variables se toman en cuenta el desempeño académico del alumno en base a todos los cursos llevados.

B) Variables vinculadas al rendimiento del estudiante relacionado al curso

Estas variables se refieren a los rendimientos de los estudiantes, pero más orientados al curso. Para ello se consideraran el o los cursos pre-requisitos del curso en estudio.

C) Variables relacionadas al Grado de Dificultad de aprobación de un curso

Se trata de establecer variables que tomen en cuenta en forma global, la influencia de los diferentes factores que hacen que un curso sea más difícil de aprobar que otros.

D) Variables que midan la influencia de los otros cursos que se quiere llevar en el período académico en estudio.

Se busca establecer variables que tomen en cuenta la carga académica del estudiante, es decir, el total de cursos en los que se quiere matricular.

Ya que cuanto mayor es la carga académica, menor será la probabilidad de aprobar uno de los cursos en particular.

En la tabla 3.2 se muestran las distintas variables que serán discutidas.

3.2.3 Discusión de las variables

Considerando los cuatro grupos:

A. Con respecto al Rendimiento Global del Estudiante

Se establece las siguientes variables:

1. Promedio ponderado del semestre previo.
2. Promedio ponderado acumulado al semestre previo.
3. Efectividad de aprobación 1 del semestre previo.
4. Promedio de Efectividad de aprobación 1 por semestre hasta el semestre previo.
5. Efectividad de aprobación 2 del semestre previo.
6. Promedio de Efectividad de aprobación 2 por semestre hasta el semestre previo.
7. Antigüedad en años del alumno desde el año de ingreso a la UNI hasta el año del semestre previo.

Las dos primeras variables toman en cuenta el rendimiento del estudiante a través de su promedio ponderado en todos los cursos llevados. Con la diferencia, que la primera solo considera el promedio ponderado del semestre previo, mientras que la segunda, nos da el promedio ponderado acumulado hasta el semestre previo. Los valores de estas dos variables van desde 0.1 hasta 20.

En la figura 3.1 se ve el comportamiento de estas dos variables para tres alumnos, uno con un promedio ponderado acumulado de aproximadamente 13 (Gráficos 1 y 2), otro con 10 (Gráficos 2 y 3) y el último con 07 (Gráficos 5 y 6).

Mientras que la segunda variable nos da el comportamiento histórico del estudiante, con valores que se mantienen relativamente constantes a pesar de la influencia de los factores no académicos coyunturales que se pueden presentar, la primera variable tiene valores solo de un período académico, por lo que cambia más apreciablemente, debido precisamente a los mismos factores mencionados.

Como estas dos variables nos dan el comportamiento global del estudiante y se desean usar para predecir la nota de un curso en particular, conviene que la variable tenga la mayor regularidad posible y no que no esté oscilante con los períodos académicos. Es por ello, que seleccionamos de las dos, la variable ***Promedio ponderado acumulado al semestre previo.***

La tercera y cuarta variable consideran también el rendimiento del estudiante pero en función del porcentaje del número de cursos aprobados con respecto a los cursos matriculados. Mientras en la tercera variable este porcentaje se calcula para el semestre previo, en la cuarta variable se toma el promedio de estos porcentajes de todos los semestres hasta semestre previo. Los valores de estas dos variables van de 0 a 1 (de 0 % a 100 %).

En la figura 3.2 se ve el comportamiento de estas dos variables 3 y 4 para tres estudiantes, uno con un promedio ponderado acumulado de

aproximadamente 13 (Gráficos 7 y 8), otro con 10 (Gráficos 9 y 10) y uno último con 07 (Gráficos 11 y 12).

Al igual que las dos primeras variables, existe aquí un comportamiento histórico, de cambio gradual para la cuarta variable, mientras para la tercera variable el comportamiento es muy cambiante. Por las mismas razones expuestas para las dos primeras variables, aquí también seleccionamos la variable ***Promedio de Efectividad de aprobación 1 por semestre hasta el semestre previo.***

La quinta y sexta variables son parecidas a la tercera y cuarta variable, solo que en vez del número de cursos se utiliza el número de créditos de cada curso.

En la figura. 3.3 se ve el comportamiento de estas dos variables 5 y 6 para tres alumnos, uno con un promedio ponderado acumulado de aproximadamente 13 (Gráficos 13 y 14), otro con 10 (Gráficos 15y 16) y uno último con 07 (Gráficos 17 y 18).

Por tanto, por las mismas razones expuestas para las variables 3 y 4, seleccionamos entre las variables 5 y 6, a la variable ***Promedio de Efectividad de aprobación 2 por semestre hasta el semestre previo.***

Hasta aquí nos estamos quedando con la segunda, cuarta y sexta variable.
Tomando en cuenta lo siguiente dos puntos:

(1) Entre la cuarta y la sexta variable, la única diferencia está en el uso del número de cursos ó el del número de créditos de los cursos

CLASIFICACION	VARIABLES						
RENDIMIENTO GLOBAL DEL ESTUDIANTE	PROMEDIO PONDERADO DEL SEMESTRE PREVIO (0-20)	PROMEDIO PONDERADO ACUMULADO AL SEMESTRE PREVIO (0 – 20)	EFECTIVIDAD 1 DE APROBACION 1 DEL SEMESTRE PREVIO (0 – 1)	PROMEDIO DE EFECTIVIDAD DE APROBACION 1 POR SEMESTRE HASTA EL SEMESTRE PREVIO (0 – 1)	EFECTIVIDAD DE APROBACION 2 DEL SEMESTRE PREVIO (0 – 1)	PROMEDIO DE EFECTIVIDAD DE APROBACION 2 POR SEMESTRE HASTA EL SEMESTRE PREVIO (0 – 1)	ANTIGUEDAD EN AÑOS DESDE EL AÑO DE INGRESO A LA UNI HASTA EL AÑO DEL SEMESTRE PREVIO
RENDIMIENTO DEL ESTUDIANTE RELACIONADO AL CURSO	NOTA DE PRE-REQUISITO 1 (0-20)		NOTA DE PRE-REQUISITO 2 (0-20)				
GRADO DE DIFICULTAD DE APROBACION DE UN CURSO	GRADO DE DIFICULTAD 1 DE APROBACION DE UN CURSO EN EL SEMESTRE PREVIO (0 – 20)		PROMEDIO DEL GRADO DE DIFICULTAD 1 DE APROBACION POR SEMESTRE HASTA EL SEMESTRE PREVIO (0-20)		GRADO DE DIFICULTAD 2 DE APROBACION DE UN CURSO EN EL SEMESTRE PREVIO (0 – 20)	PROMEDIO DEL GRADO DE DIFICULTAD 2 DE APROBACION POR SEMESTRE HASTA EL SEMESTRE PREVIO (0 - 1)	
INFLUENCIA DE LOS OTROS CURSOS QUE SE QUIERE LLEVAR EN EL SEMESTRE EN ESTUDIO	NUMERO DE CURSOS TOTALES QUE SE QUIERE LLEVAR EN EL SEMESTRE EN ESTUDIO (0 – 10)		NUMERO DE CREDITOS TOTALES QUE SE QUIERE LLEVAR EN EL SEMESTRE EN ESTUDIO (1 – 28)		SUMATORIA DEL PROMEDIO DEL GRADO DE DIFICULTAD (POR SEMESTRE HASTA EL SEMESTRE PREVIO) , EVALUADO PARA TODOS LOS CURSOS QUE LLEVARAN EN EL SEMESTRE EN ESTUDIO (0 - 200)		

Tabla 3.2 Variables consideradas para predecir la nota de un curso

(2) Sabemos que el creditaje (número de créditos de un curso) nos da información adicional relacionada con las horas semanales de teoría y de práctica de cada curso, es decir, en el creditaje cada curso está siendo ponderado por el número de horas que tiene.

Podemos seguir seleccionando y considerando por lo expuesto, que la variable 6 tiene un mayor valor agregado (número de horas de cada curso), y escogemos entre las variables 4 y 6, a la variable ***Promedio de Efectividad de aprobación 2 por semestre hasta el semestre previo.***

Hasta aquí están quedando la segunda y sexta variable.

La segunda variable da el rendimiento en función de las notas obtenidas por el estudiante, dando valores de 0 a 20, mientras la sexta variable lo da en función del número de cursos aprobados, sin interesar la nota. Esto significa, por ejemplo si se tienen los siguientes casos:

Estudiante A: Llevó 5 cursos y los desaprobó con 9.9
Estudiante B: Llevó 5 cursos y los aprobó con 10.0
Estudiante C: Llevó 5 cursos y los aprobó con 15.0

Se tendrían los siguientes resultados

	Promedio Ponderado (0 a 20)	Efectividad de Aprobación 2 (0 a 1)
A	9.9	0.0
B	10.0	1.0
C	15.0	1.0

Lo que significa que:

- A pesar que los estudiantes A y B tienen promedios ponderados parecidos, A tiene una Efectividad de Aprobación del 0 %, mientras B tiene de 100 %.

- A pesar que los estudiantes B y C tienen promedios ponderados diferentes, su efectividad de aprobación es la misma.

El resultado de la comparación entre el Promedio Ponderado y la Efectividad de Aprobación 2, es la misma que para el Promedio Ponderado Acumulado y el Promedio de Efectividad de Aprobación 2.

Es por ello, que de estas dos comparaciones, podemos seguir seleccionando, esta vez entre las variables 2 y 6, escogemos a la variable ***Promedios ponderado acumulado.***

La variable 7, Antigüedad en años desde el año de ingreso a la UNI hasta el año del semestre previo, nos indica que cuanto mayor es este valor, menor es el rendimiento global del alumno. Como esta variable 7 no está relacionada directamente con la nota obtenida por el estudiante en cada curso llevado, también será seleccionado.

Conclusión:
Las variables que se considerarán para el ***Rendimiento Global del alumno*** son el ***Promedio Ponderado Acumulado*** **al Semestre previo** y la **Antigüedad en años del alumno desde el año de ingreso a la UNI hasta el año del semestre previo.**

B. Con respecto al Rendimiento del estudiante relacionado al curso

Se establece las siguientes variables:

1. Nota del curso prerrequisito 1
2. Nota del curso prerrequisito 2

Las dos variables consideran el rendimiento del estudiante ya no en todos los cursos, sino en los cursos más relacionados con el Curso en

estudio, es decir, los cursos Pre-requisitos. Estas variables emplean la nota de estos cursos pre-requisitos. Cuanto mayor sea la Nota en estos cursos, mayor será la probabilidad de aprobar el Curso en estudio. Se usará la nota aprobatoria del curso prerrequisito que obtuvo el estudiante. Esto significa que si el estudiante ha llevado dos o más veces el curso, se tomará la última nota. Los valores para estas variables van de 10 a 20.

Conclusión:

Se seleccionan las variables **Nota del curso pre-requisito 1** y **Nota del curso pre-requisito 2,** este último en caso el Curso en estudio tuviera dos cursos pre-requisitos.

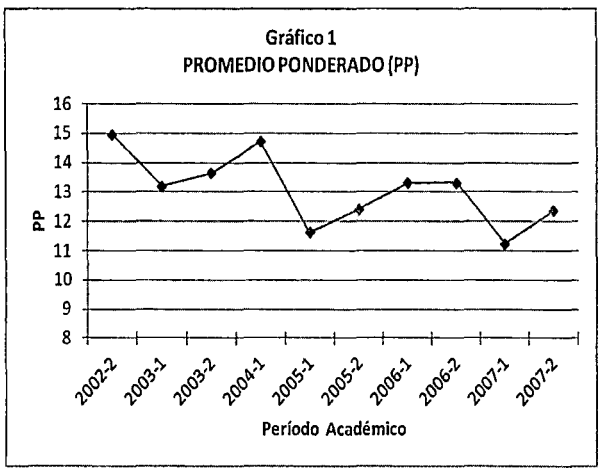
C. Con respecto al Grado de Dificultad de aprobación de un curso

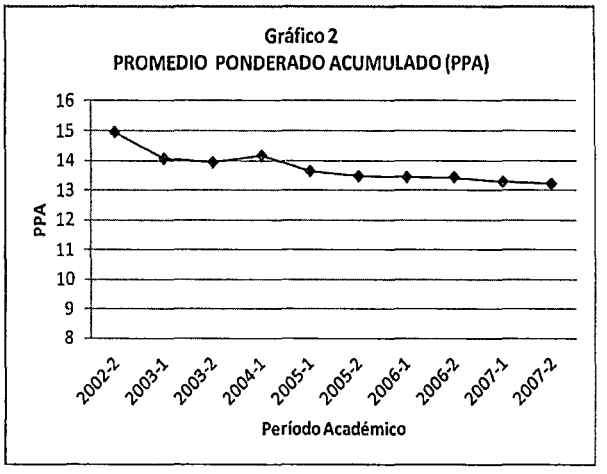
El grado de dificultad de aprobación de un curso, depende de cada curso, sabemos que algunos son más difíciles de aprobar que otros. Los factores que influyen en el valor del grado de dificultad son:

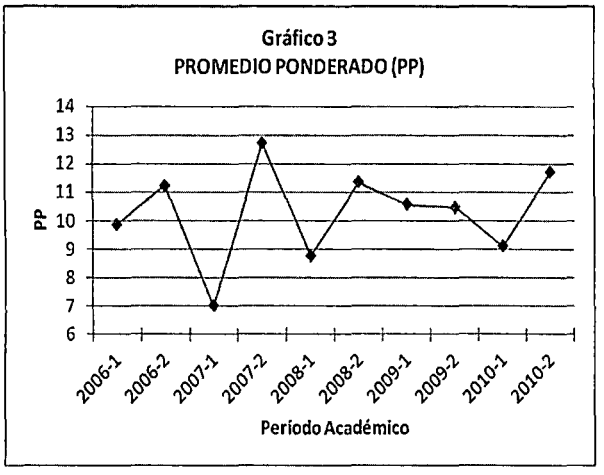
a. Las características propias de cada curso (complejidad, extensión, etc.)
b. Las características propias de cada profesor del curso (capacidad de hacerse entender, de llegar al estudiante, sus criterios para la corrección de las pruebas, etc.)
c. Rendimiento del conjunto de estudiante que llevó el curso en ese período académico.

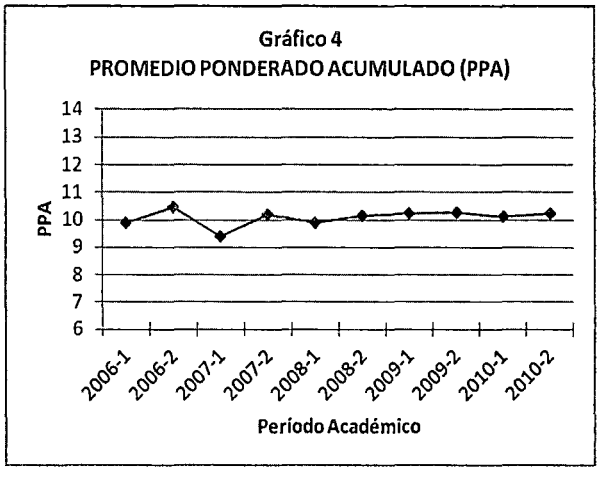
Otros factores secundarios:

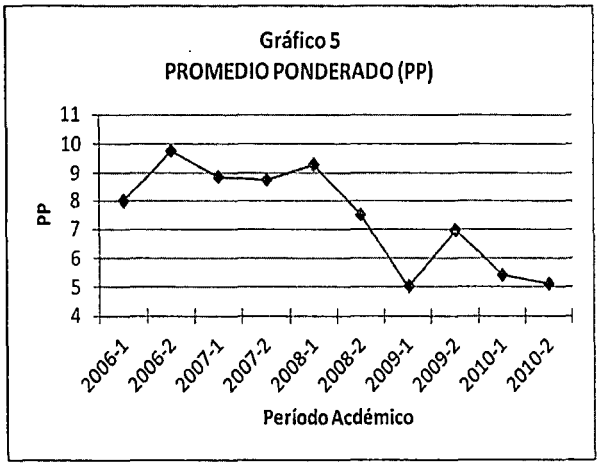
d. Si el período académico es del primer o segundo semestre del año.
e. Situación económica, política o social del país.











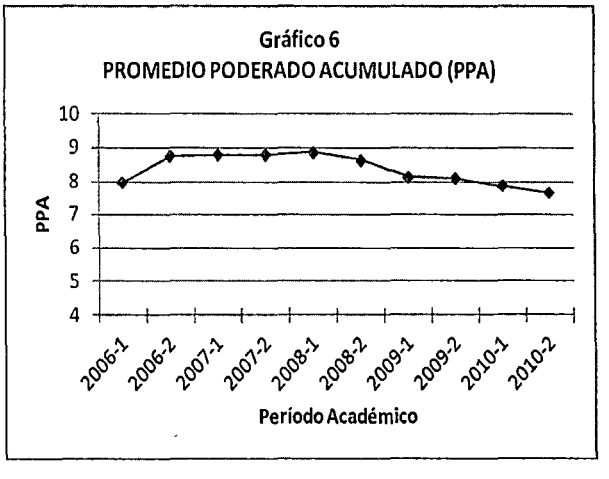
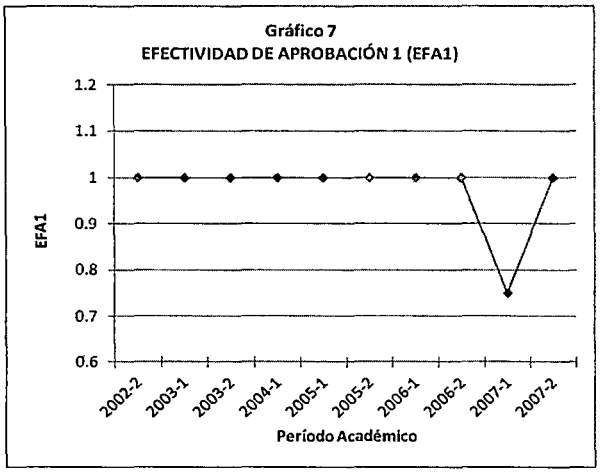
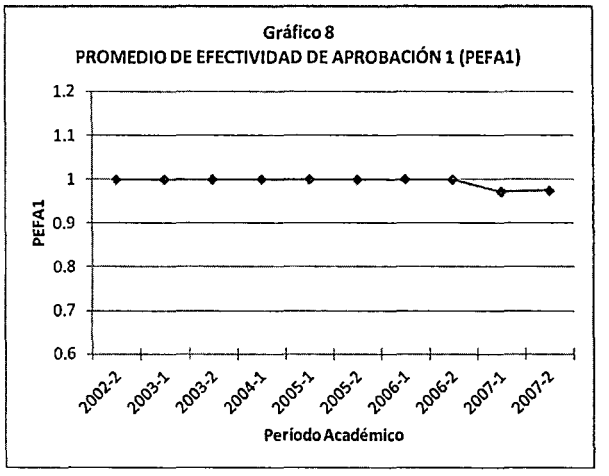
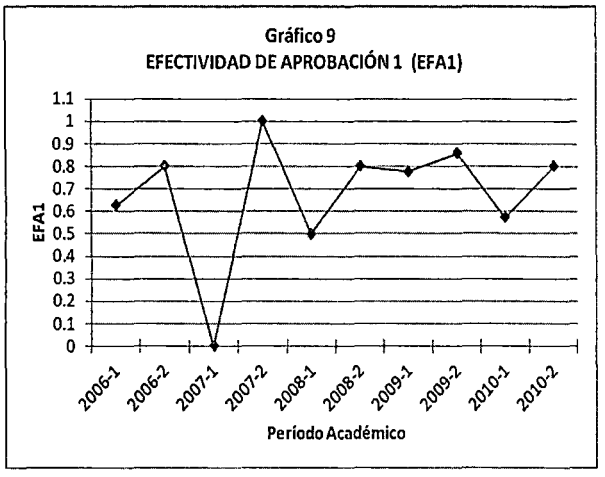
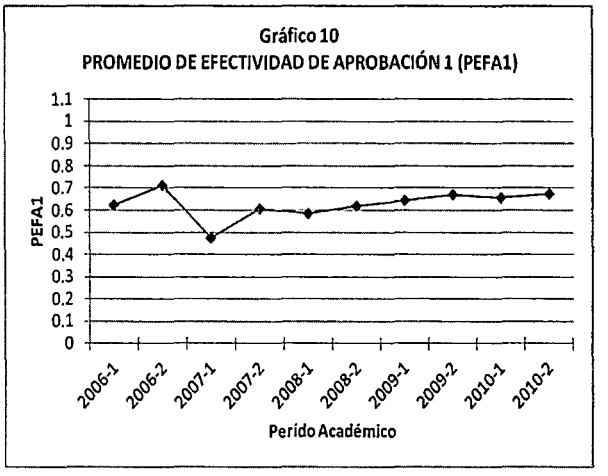


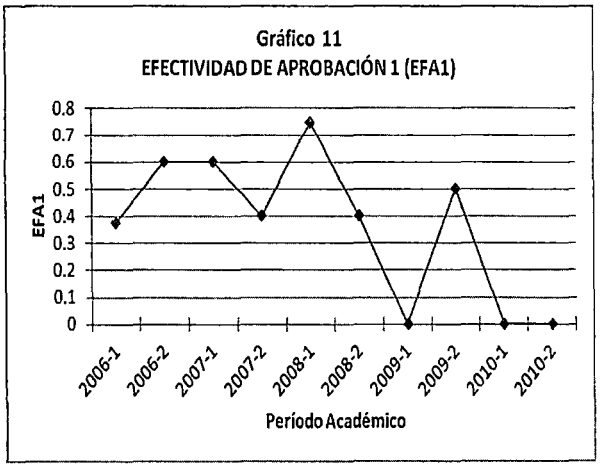
Figura 3.1 Promedio ponderado y promedio ponderado acumulado para tres alumnos











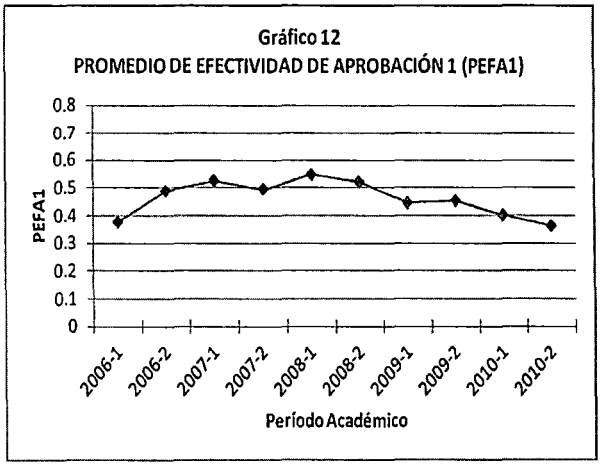
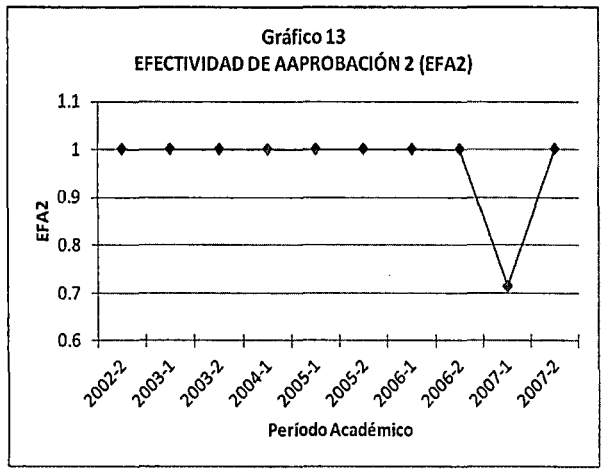
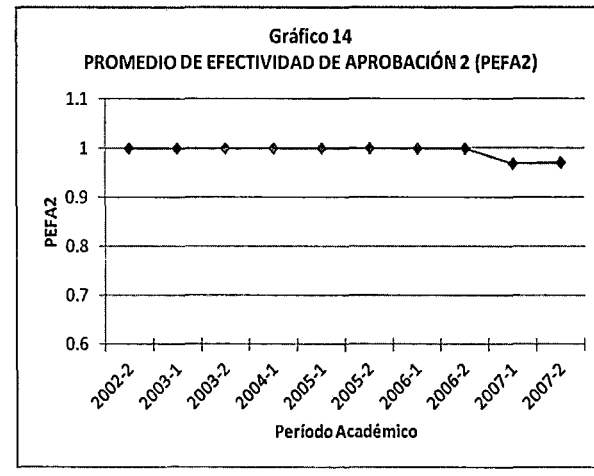
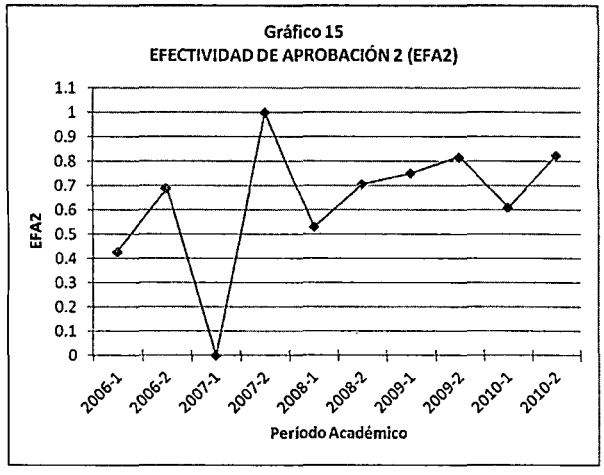
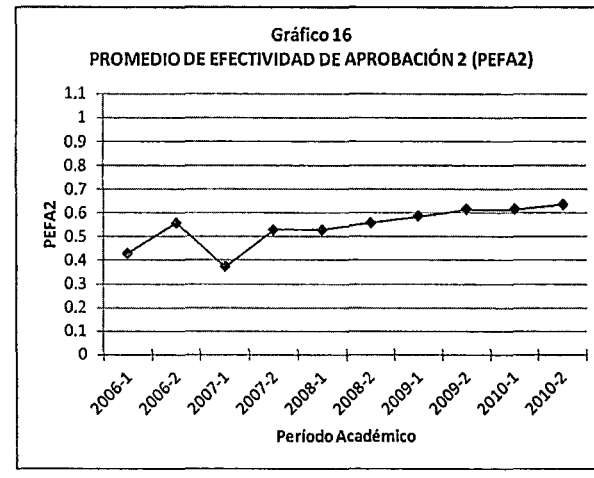


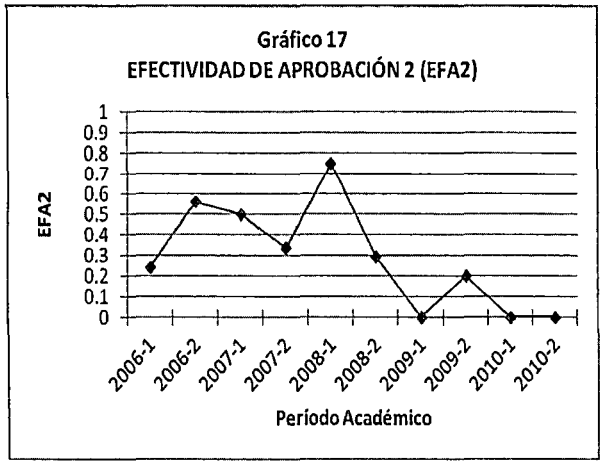
Figura 3.2 Efectividad de aprobación 1 y promedio de efectividad de aprobación 1 para tres alumnos











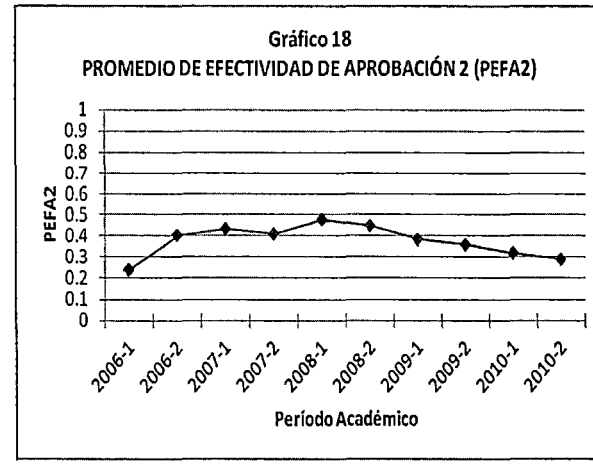


Figura 3.3 Efectividad de aprobación 2 y promedio de efectividad de aprobación 2 para tres alumnos

El grado de dificultad puede variar con el período académico, principalmente por los factores b, c, d y e, mencionados líneas arriba. Esto significa que para cada período académico y para cada curso-sección existe un valor definido de grado de dificultad de aprobación del curso.

Hemos considerado cuatro variables para representar el grado de dificultad de aprobación de un curso:

1.- Grado de Dificultad 1 de aprobación de un curso en el semestre previo (GD1)
2.- Promedio del Grado de Dificultad 1 de aprobación por semestre hasta el semestre previo.
3.- Grado de Dificultad 2 de aprobación de un curso en el semestre previo.
4.- Promedio del Grado de Dificultad 2 de aprobación por semestre hasta el semestre previo

Para el caso de aquellos cursos que tienen varias secciones en un período académico, haremos el siguiente análisis:

En la figura 3.4 se muestra el Grado de Dificultad 1 para cada una de las tres secciones de un curso (gráficos 19, 20 y 21), tomando como ejemplo para la explicación al curso MA133. Vemos que existen diferencias en el GD1 entre las tres secciones, sin embargo como deseamos que esta variable sea independiente de la sección (y por lo tanto de las características propias de cada profesor), haremos el cálculo tomando a las tres secciones como si fuera una sola (ver gráfico 22). Este mismo razonamiento (ver figura 3.5, gráficos 23, 24, 25 y 26) lo podemos aplicar para la variable Grado de Dificultad 2.

Conclusión previa: El cálculo de los dos Grados de Dificultad para un Período Académico o Semestre se realizará juntando a los estudiantes de todas las secciones en una sola.

Realizando ahora la discusión de las cuatro variables:
Las dos primeras variables consideran las notas de los estudiantes, por lo que sus valores van de 0 a 20. La primera variable considera sólo el semestre previo, es decir, toma en cuenta el valor más reciente; mientras la segunda variable toma el promedio de valores de todos los semestres hasta el semestre último, es decir, toma en cuenta el comportamiento histórico.

En las figuras 3.6 y 3.7 se muestran a las variables 1 y 2 para los cursos MA133 y PI216, respectivamente. Vemos que los valores de la variable 1 son cambiantes, mientras que para la variable 2, los valores cambian gradualmente. Como para realizar el pronóstico, necesitamos regularidad en nuestras variables, seleccionamos de estas dos, la variable **Promedio del Grado de Dificultad 1 de aprobación por semestre hasta el semestre previo.**

También en las mismas figuras 3.6 y 3.7 se muestran a las variables 3 y 4. La tercera y cuarta variables toman en cuenta el porcentaje de alumnos desaprobados con respecto a los matriculados, por lo que sus valores van de 0 a 1 (0 % a 100%). La tercera variable considera sólo el semestre previo, es decir, toma en cuenta el valor más reciente; mientras la cuarta variable toma el promedio de valores de todos los semestres hasta el del semestre previo, es decir, toma en cuenta el comportamiento histórico. El comportamiento de las variables 3 y 4, es muy similar al de las variables 1 y 2, por lo que de las variables 3 y 4, seleccionamos la variable **Promedio del Grado de Dificultad 2 de aprobación por semestre hasta el semestre previo.**

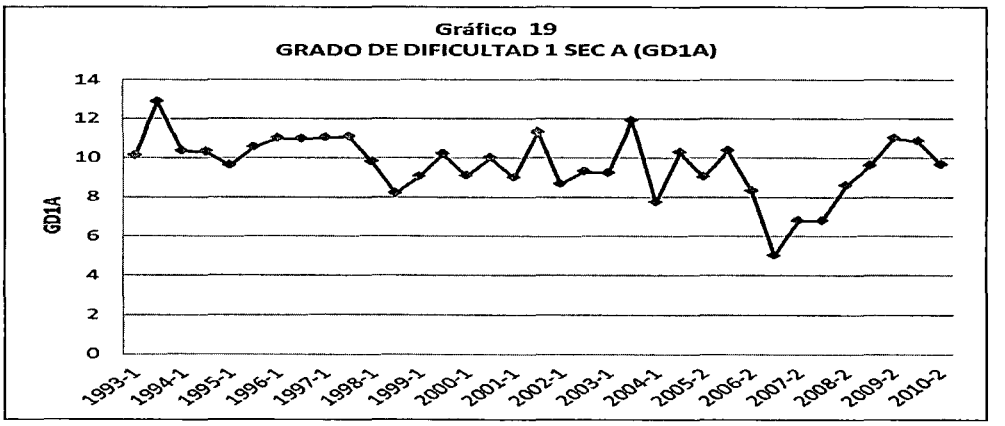
Hasta este punto nos estamos quedando con las variables 2 y 4.
Mientras que la variable 2 toma en cuenta la nota de cada uno de los estudiantes, por lo que es un valor totalmente cuantitativo, la variable 4 toma solo en cuenta de cada estudiante si ha aprobado o no, y por lo tanto representa a una expresión cualitativa. Porque consideramos que existe un mayor valor agregado en la variable 2, que en la variable cuatro, seleccionamos a la variable **Promedio del Grado de Dificultad 1 de aprobación por semestre hasta el semestre previo.**

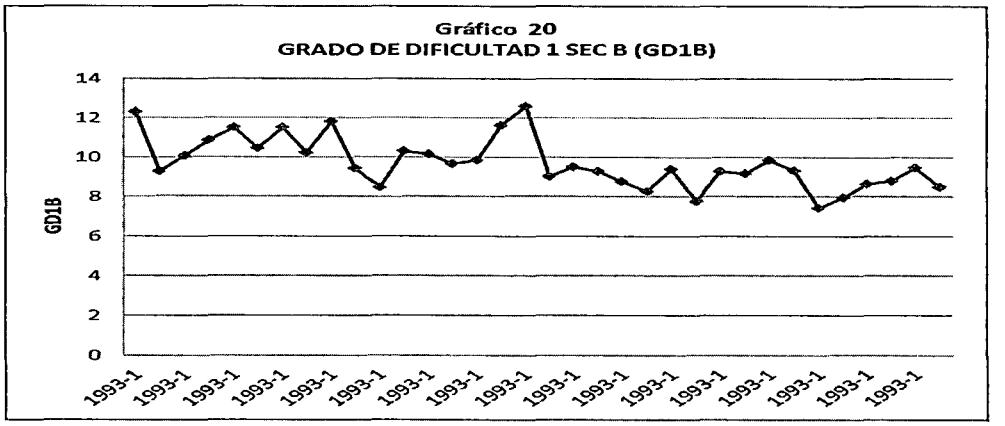
Conclusión:
Para este caso, se selecciona una sola variable, a saber, el **Promedio del Grado de Dificultad 1 de aprobación por semestre hasta el semestre previo,** que a partir de ahora, lo renombraremos como: **Promedio del Grado de Dificultad de aprobación por semestre hasta el semestre previo.**

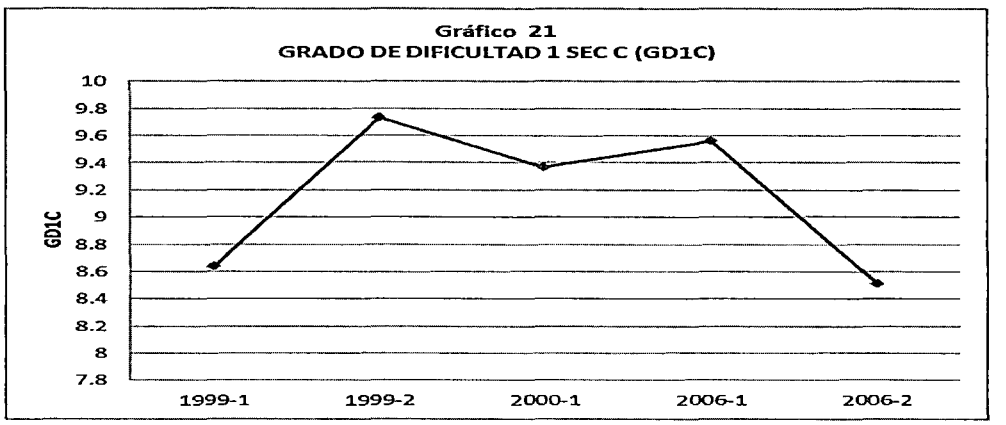
D. Con respecto a la influencia de los otros cursos que se quiere llevar en el Período Académico o Semestre en estudio.

Se establece las siguientes variables:

1.- Número de cursos totales que se quiere llevar en el Semestre en estudio.
2.- Número de créditos totales que se quiere llevar en el Semestre en estudio.
3.- Sumatoria del Promedio del Grado de Dificultad (por semestre hasta el semestre previo), evaluado para todos los cursos que llevará en el Semestre en estudio.







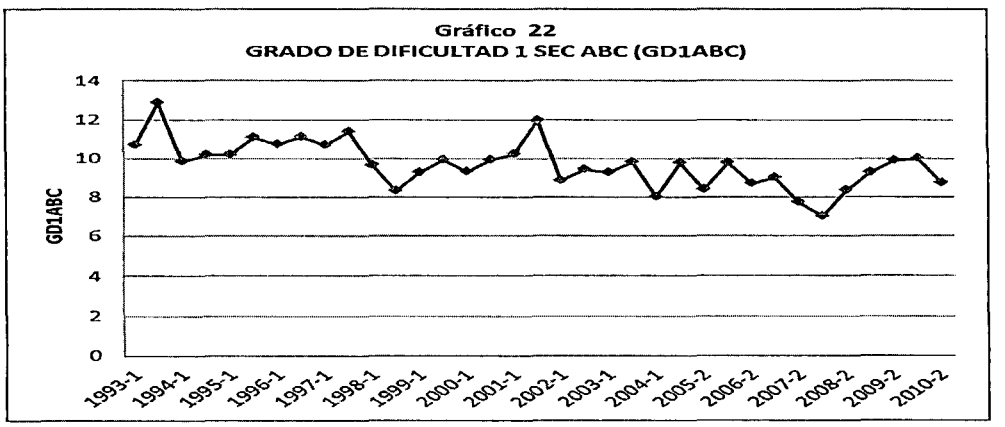
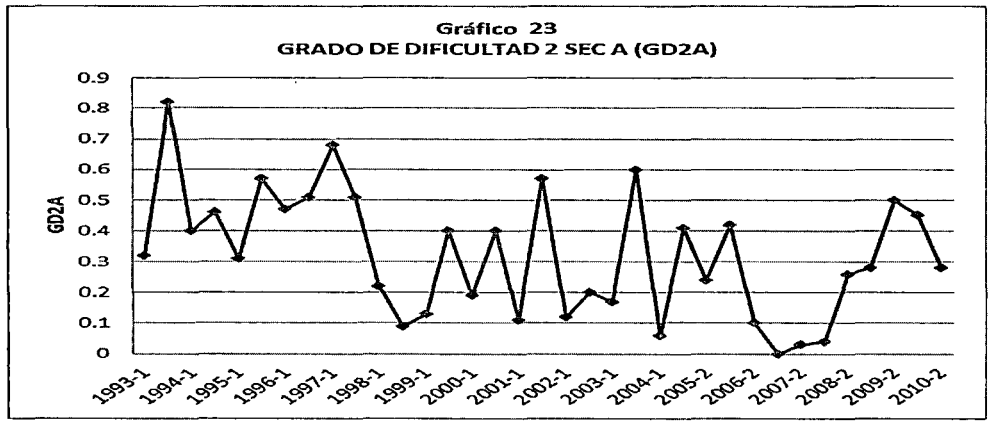
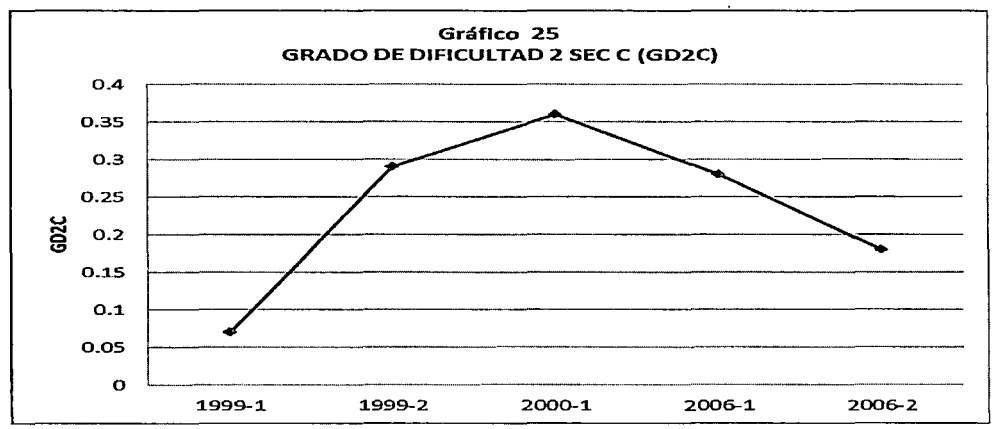


Figura 3.4 Grado de dificultad 1 por sección y en conjunto





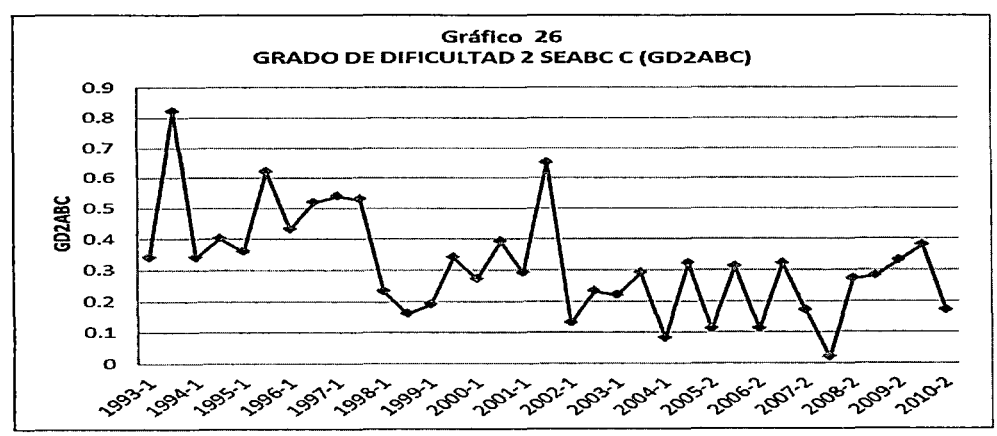


Figura 3.5 Grado de dificultad 2 por sección y en conjunto

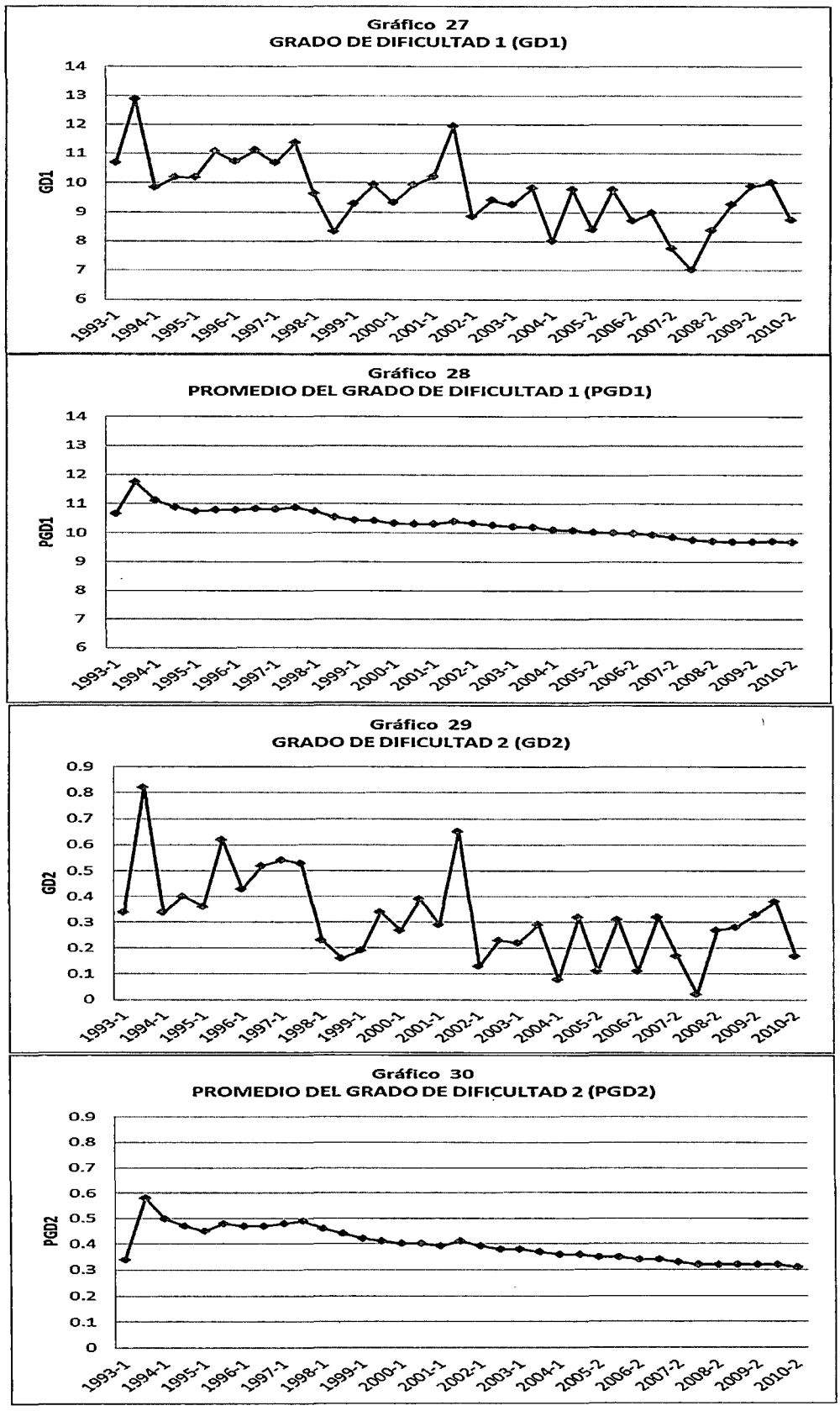


Figura 3.6 Grado de dificultad 1 y 2 y sus promedios de MA133

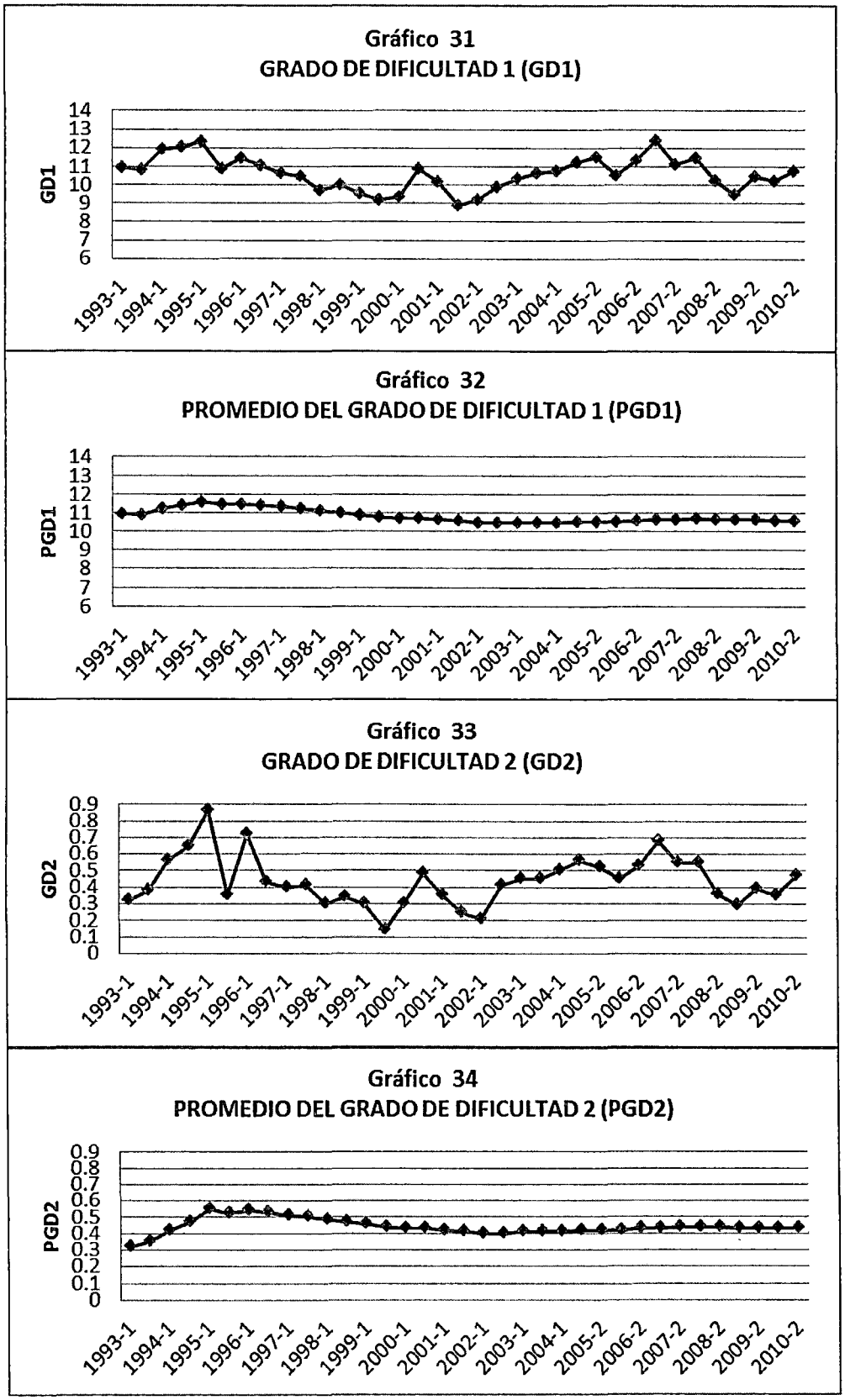


Figura 3.7 Grado de dificultad 1 y 2 y sus promedios de PI216

Cuantos más cursos lleve el alumno en el Semestre en estudio, menos tiempo va a tener para la dedicación a cada curso y por lo tanto mayor será la probabilidad de que no apruebe alguno de los cursos en particular. De aquí nace la variable 1. Los valores de esta variable van desde 2 a 8.

Pero una forma de considerar más apropiadamente la carga académica del estudiante en un Semestre, es a través del uso del concepto de creditaje, que toma en cuenta el número de horas de teoría y de práctica que tiene cada curso a la semana, y lo usa como una ponderación para cada uno de los cursos que desea llevar. De aquí nace la variable 2. Los valores de esta variable van desde 1 hasta 28. Tomando en cuenta que existe mayor información en la variable 2 que en la 1, seleccionamos a la variable Número de créditos totales que se quiere llevar en el Semestre en estudio.

En la variable 3, a diferencia de la variable 2, se está considerando como ponderación el Promedio del Grado de Dificultad de aprobación de un curso, para cada uno de los cursos que se desea llevar y de esta manera tener otra forma de representar la influencia de todos los cursos. Al ser el creditaje un concepto bastante diferente al Grado de Dificultad de aprobación de un curso, seleccionamos también a la variable **Sumatoria del Promedio del Grado de Dificultad (por semestre hasta el semestre previo), evaluado para todos los cursos que llevará en el Semestre en estudio.**

Conclusión:

Se seleccionan las variables **Número de créditos totales que se quiere llevar en el Semestre en estudio y Sumatoria del Promedio del Grado de Dificultad (por semestre hasta el semestre previo), evaluado para todos los cursos que llevará en el Semestre en estudio.**

3.2.4 Comparación con variables utilizadas en otros estudios

En el Trabajo de Romero, Ventura, Espejo y Hervás [17] las variables utilizadas, tomadas de datos y primeras evaluaciones del curso para predecir la nota final en el curso son:

i. Número de identificación del curso
ii. Número de asignaciones hechas
iii. Número de test tomados
iv. Número de test aprobados
v. Número de test desaprobados
vi. Número de mensajes enviados al fórum
vii. Número de mensajes leídos en el fórum
viii. Tiempo total usado en las asignaciones
ix. Tiempo total usado en los test
x. Tiempo total usado en el forum

Y la variable dependiente Nota fue discretizada en cuatro rangos:

i. MALO : Nota < 5
ii. REGULAR : Nota >= 5 y < 7
iii. BUENO : Nota >= 7 y < 9
iv. EXCELENTE : Nota >= 9

En el trabajo de Vialardi, Bravo, Shafti y Ortigosa [21] las variables para la predicción de la Nota de un curso son:

i. Número de cursos tomados simultáneamente
ii. Nombre del curso
iii. Promedio ponderado acumulado hasta el semestre previo

Y la variable dependiente Nota fue discretizada en dos rangos:

i. DESAPROBADO : Nota de 0 a 10.99
ii. APROBADO : Nota de 11.00 a 20

En el trabajo de Vialardi [22] las variables consideradas para la predicción de la Nota de un curso son:

i. Nombre del curso
ii. Número de veces que lleva el curso
iii. Promedio ponderado acumulado con que inicia el ciclo
iv. Potencial: que depende de las notas de los pre-requisitos inmediatos y no inmediatos.
v. Número de créditos del curso
vi. Número de créditos matriculados
vii. Dificultad: indica la dificultad del curso en base al promedio de los alumnos que la han llevado.

Y la variable dependiente Nota fue discretizada de la siguiente forma:

i. SUSPENSO: Nota < 11
ii. APROBADO: Nota >= 11

En nuestro trabajo vemos que nuestras variables predictoras, coinciden con muchas de las variables de la referencia [22], con las siguientes observaciones:

La variable "potencial", para nosotros son las notas de solo los cursos pre-requisitos.

Con respecto a la influencia de los otros cursos inscritos, además del Número de créditos matriculados, que ha considerado Vialardi, nosotros hemos tomado adicionalmente la sumatoria de los promedios del grado de dificultad de los cursos matriculados.

Dentro del rendimiento global del estudiante estamos considerando la antigüedad del estudiante en años.

Con respecto a la variable dependiente Nota del Curso, ésta se considera de dos maneras, como variable cuantitativa, que sería precisamente la Nota; y como variable categórica, que se discretiza en dos rangos:

i. DESAPROBADO : Nota de 0 a 9.9
ii. APROBADO : Nota de 10 a 20

3.3 Generación de los datos para las variables

Consiste en calcular los valores de las variables que se utilizaran en este estudio. Las variables que se considerarán en este estudio y que fueron listadas en el apartado 1.8.4 son:

1. Promedio ponderado acumulado al semestre previo.
2. Antigüedad en años del estudiante desde ingreso a la UNI hasta el semestre previo inclusive.
3. Nota del curso pre-requisito 1.
4. Nota del curso pre-requisito 2.
5. Promedio del Grado de Dificultad de aprobación de un curso por semestre hasta el semestre previo.
6. Número de créditos totales que se quiere llevar en el semestre actual.
7. Sumatoria del promedio del Grado de Dificultad de aprobación de un curso (por semestre hasta el semestre previo), evaluado para todos los cursos que se llevaran en el semestre actual.
8. Nota del Curso en el semestre actual.

Estas variables son calculadas mediante un programa realizado en JAVA que se da en el Anexo 2 y que se explica a continuación.

3.3.1 Descripción del Programa en Java

Este Programa se conecta a la Base de Datos (BD) en Microsoft Office Access, llamada "BASE", la cual ha sido extraída del archivo Excel "Base Histórica de Notas Limpia". La BD cuenta con las siguientes tablas:

- HIST, que tiene los campos: código del alumno (CODALU), Especialidad (ESPEC), código del curso (CODCUR), sección (SEC), período académico (PERACD), número de créditos del curso (CRD), nota del curso (NOTA), situación (SIT) y condición (CND). Estos campos son los mismos que se presentaron en la tabla 2.5. Es por ello que la tabla HIST contiene la información de todos los alumnos de la especialidad de Ingeniería Química (Q1) desde el período académico 1993-1 hasta el periodo académico 2010-2. Un extracto de este archivo se da en la tabla 3.3.

- CURSO, que tiene los campos: código del curso (CODCUR), código del pre-requisito 1 (PREREQ1) y código del pre-requisito 2 (PREREQ2). La tabla CURSO contiene todos los cursos y sus pre-requisitos de la especialidad Q1. La información para esta tabla es la misma que se halla en la Tabla 2.2.

El programa está dividido en dos bloques:

Primer bloque (figura 3.8): Aquí el programa genera la tabla llamada GD, que es insertada a la BD. La tabla GD contiene los grados de dificultad de todos los cursos de la especialidad Q1. La tabla GD tiene los campos: número de registro (Registro), código del curso (CODCUR), período académico (PERACD), promedio de notas del período académico (PN), grado de dificultad de un curso en un periodo académico (GD) y el promedio del grado de dificultad hasta el periodo

académico previo (PGD). Una parte de esta tabla se brinda en la tabla 3.4.

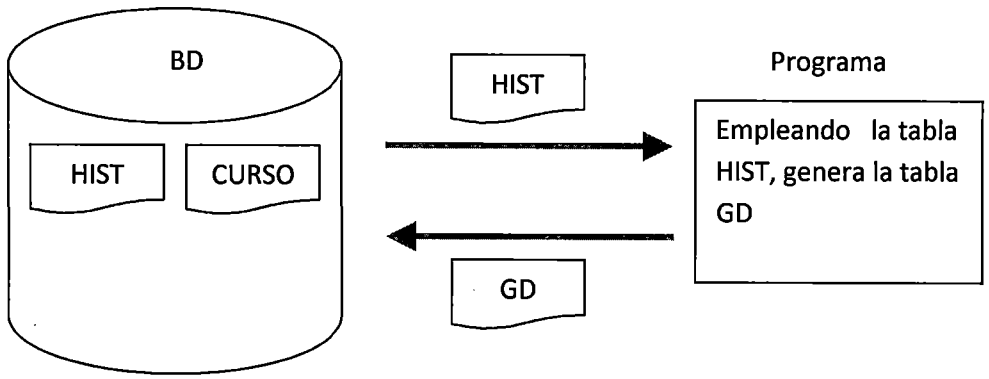


Figura 3.8 Esquema del Primer Bloque

Segundo bloque (figura 3.9): En este bloque el programa solicita el código de un curso (por ejemplo PI216) de la carrera de la especialidad Q1, con el cual se genera la tabla llamada PGDPI216, que es insertada a la BD. Esta Tabla (por ejemplo, PGDPI216) tiene los campos: número de registro (Registro), período académico (perAca), código del alumno (cod), código del curso (cur), nota del curso (notCur), promedio ponderado acumulado (ppa), código del pre-requisito 1 (curPreReq1), nota del pre-requisito 1 (notPreReq1), código del pre-requisito 2 (curPreReq2), nota del pre-requisito 2 (notPreReq2), promedio del grado de dificultad del curso hasta el período académico previo (pGD), suma de los pGD de todos los cursos inscritos en el período académico en estudio (spgd), la suma de los créditos de los cursos que lleva en el período académico en estudio (sumCre) y la antigüedad del alumno que lleva el curso, en el momento del semestre previo (antAlu). Ver extracto de esta tabla en la tabla 3.5.

CODALU	ESPEC	CODCUR	SEC	PERACD	CRD	NOTA	SIT	CND
19962630E	Q1	PI415	A	2001-1	3	10.2	A	N
19962630E	Q1	PI475	A	2001-1	4	12.5	A	N
19962630E	Q1	PI510	A	2001-1	3	13.2	A	N
19962630E	Q1	PI612	A	2001-1	2	12.5	A	N
19962630E	Q1	PI911	B	2001-1	4	10.3	A	N
19962632H	Q1	AU511	6	1996-2	2	3.6	D	N
19962632H	Q1	FI203	6	1996-2	5	10.2	A	N
19962632H	Q1	MA113	6	1996-2	4	8.4	D	N
19962632H	Q1	MA114	6	1996-2	3	9.8	D	N
19962632H	Q1	QU116	6	1996-2	3	10.3	A	N
19962632H	Q1	QU117	6	1996-2	1	12.4	A	N
19962632H	Q1	AU511	7	1997-1	2	6.8	D	N
19962632H	Q1	FI204	6	1997-1	5	12.5	A	N
19962632H	Q1	MA113	8	1997-1	4	10.8	A	N
19962632H	Q1	MA124	7	1997-1	3	9.6	D	N
19962632H	Q1	QU118	7	1997-1	3	6.3	D	N
19962632H	Q1	QU119	6	1997-1	1	12.4	A	N
19962632H	Q1	EE102	7	1998-2	3	2.3	D	N
19962632H	Q1	EP307	7	1998-2	4	7.6	D	N
19962632H	Q1	MA133	7	1998-2	6	12.2	A	N
19962632H	Q1	PI100	6	1998-2	1	11.5	A	N
19962632H	Q1	PI118	6	1998-2	2	7	D	N
19962632H	Q1	FI403	7	1997-2	5	7.6	D	N
19962632H	Q1	MA123	7	1997-2	4	9.2	D	N
19962632H	Q1	QU118	6	1997-2	3	16.3	A	N
19962632H	Q1	EC618	B	2000-2	5	10.8	A	N
19962632H	Q1	EM711	B	2000-2	3	9.5	D	N
19962632H	Q1	MA143	B	2000-2	4	11	A	N
19962632H	Q1	PI111	A	2000-2	3	11.1	A	N
19962632H	Q1	QU516	B	2000-2	3	10.3	A	N
19962632H	Q1	FI152	B	2001-2	4	6.8	D	N
19962632H	Q1	PA515	B	2001-2	2	11	A	N
19962632H	Q1	PI143	A	2001-2	3	11.3	A	N
19962632H	Q1	PI146	A	2001-2	3	10.5	A	N
19962632H	Q1	PI513	A	2001-2	2	12.1	A	N
19962632H	Q1	PI523	B	2001-2	4	12.6	A	N
19962632H	Q1	QU334	A	2001-2	4	7.6	D	N
19962632H	Q1	EE102	7	1999-1	3	11.4	A	N
19962632H	Q1	EP307	8	1999-1	4	11	A	N
19962632H	Q1	MA612	6	1999-1	4	12.8	A	N
19962632H	Q1	MA713	6	1999-1	3	7.1	D	N

Tabla 3.3 Tabla HIST

Registro	CODCUR	PERACD	PN	GD	PGD
640	MA113	2005-2	10.24	9.76	9.557084371
641	MA113	2006-1	9.594827586	10.4051724	9.580005669
642	MA113	2006-2	8.950793651	11.0492063	9.618668845
643	MA113	2007-1	9.780701754	10.2192982	9.634069599
644	MA113	2007-2	10.05333333	9.94666667	9.641884526
645	MA113	2008-1	10.1921875	9.8078125	9.645931549
646	MA113	2008-2	10.80655738	9.19344262	9.635158003
647	MA113	2009-1	9.946808511	10.0531915	9.644879712
648	MA113	2009-2	9.830769231	10.1692308	9.656796782
649	MA113	2010-1	11.5372093	8.4627907	9.630263313
650	MA113	2010-2	9.708571429	10.2914286	9.644636471
651	MA114	1975-1	13.6	6.4	6.4
652	MA114	1977-1	12.2	7.8	7.1
653	MA114	1978-2	10.9	9.1	7.766666667
654	MA114	1983-1	10.8	9.2	8.125
655	MA114	1984-2	11.15	8.85	8.27
656	MA114	1985-1	10.15	9.85	8.533333333
657	MA114	1986-1	11.6	8.4	8.514285714
658	MA114	1988-1	11.8	8.2	8.475
659	MA114	1989-1	10.2	9.8	8.622222222
660	MA114	1993-1	9.475	10.525	8.8125
661	MA114	1993-2	9.841975309	10.1580247	8.934820426
662	MA114	1994-1	10.32133333	9.67866667	8.996807613
663	MA114	1994-2	9.994736842	10.0052632	9.074381117
664	MA114	1995-1	9.727659574	10.2723404	9.159949639
665	MA114	1995-2	10.28275862	9.71724138	9.197102421
666	MA114	1996-1	9.016071429	10.9839286	9.308779056
667	MA114	1996-2	9.031521739	10.9684783	9.406408421
668	MA114	1997-1	8.073493976	11.926506	9.546413843
669	MA114	1997-2	9.11322314	10.8867769	9.616959265
670	MA114	1998-1	8.636507937	11.3634921	9.704285905
671	MA114	1998-2	9.377987421	10.6220126	9.747987175
672	MA114	1999-1	9.09375	10.90625	9.800635485
673	MA114	1999-2	9.828703704	10.1712963	9.816751173
674	MA114	2000-1	9.509859155	10.4901408	9.844809076
675	MA114	2000-2	10.475	9.525	9.832016713
676	MA114	2001-1	9.366037736	10.6339623	9.862860772
677	MA114	2001-2	9.829310345	10.1706897	9.874261842
678	MA114	2002-1	10.7942029	9.2057971	9.850388101
679	MA114	2002-2	10.78088235	9.21911765	9.828620155

Tabla 3.4 Tabla GD

Registro	perAca	cod	cur	notCur	ppa	curPreReq1	notPreReq1	curPreReq2	notPreReq2	pGD	spgd	sumCre	antAlu
362	1995-2	19930464B	PI216	10.6	14.2823529		0	QU434	17.3	10.379968	64.8634959	19	2
363	1996-1	19930471I	PI216	10	13.86	PI111	14.8	QU434	13.3	10.42583	87.2623461	27	3
364	1997-1	19930474H	PI216	8.3	10.1681818	PI111	10.1	QU434	10.3	10.555029	66.2783519	22	4
365	1997-2	19930474H	PI216	10.7	10.1863636	PI111	10.1	QU434	10.3	10.561151	72.2844528	22	4
366	1996-1	19930477G	PI216	9.4	9.59166667		0	QU434	10	10.42583	59.8566018	21	3
367	1996-2	19930477G	PI216	12.3	9.25238095		0	QU434	10	10.510483	77.1340198	22	3
368	1996-1	19930483G	PI216	7.7	10.1176471	PI111	10.1	QU434	12	10.42583	66.3769597	19	3
369	1996-2	19930483G	PI216	10.2	8.17894737	PI111	10.1	QU434	12	10.510483	63.354753	20	3
370	1996-1	19930519A	PI216	8.5	9.65454545	PI111	12.8	QU434	11.3	10.42583	69.0638608	21	3
371	1996-2	19930519A	PI216	12.6	10.2857143	PI111	12.8	QU434	11.3	10.510483	70.6765346	21	3
372	1997-1	19930533D	PI216	12.3	10.2238095	PI111	10.6	QU434	11	10.555029	76.7173862	22	4
373	1996-1	19930561H	PI216	8.5	11.152381	PI111	11.2	QU434	10.3	10.42583	67.0287972	22	3
374	1996-2	19930561H	PI216	12.7	9.64545455	PI111	11.2	QU434	10.3	10.510483	68.9391232	21	3
375	1996-1	19930610I	PI216	7.7	9.84090909	PI111	14.2		0	10.42583	73.2244683	21	3
376	1996-2	19930610I	PI216	10.2	9.93333333	PI111	14.2		0	10.510483	63.4849902	21	3
377	1998-1	19930617C	PI216	11.8	9.86190476		0	QU434	10.6	10.554386	58.2991271	17	5
378	1997-1	19930643D	PI216	11.5	9.365		0	QU434	11.6	10.555029	75.9589612	20	4
379	1997-1	19930644K	PI216	8.4	7.51111111		0	QU434	10.3	10.555029	65.3464198	21	4
380	1997-2	19930644K	PI216	10	9.73809524		0	QU434	10.3	10.561151	70.0158047	22	4
381	1998-1	19930651G	PI216	12.2	10.1714286	PI111	14	QU434	12	10.554386	70.3994011	18	5
382	1997-1	19930655B	PI216	8.4	9.22777778	PI111	11		0	10.555029	70.8790394	22	4
383	1997-2	19930655B	PI216	12.1	8.10454545	PI111	11		0	10.561151	60.0544783	20	4
384	1996-2	19930659H	PI216	10	11.075	PI111	15.3	QU434	10	10.510483	69.2529548	17	3
385	1997-2	19930668G	PI216	8.8	12.71875	PI111	10.1	QU434	13.3	10.561151	87.7815124	26	4
386	1998-1	19930668G	PI216	12.3	10.2576923	PI111	10.1	QU434	13.3	10.554386	64.9410615	21	5
387	1999-1	19930673K	PI216	8.4	7.07368421		0		0	10.47255	62.1693097	18	6

Tabla 3.5 Tabla PGDPI216

Esta Tabla por ejemplo PGDPI216 que se halla en una Base de Datos de Microsoft Office Access, se exporta a un archivo en Excel.

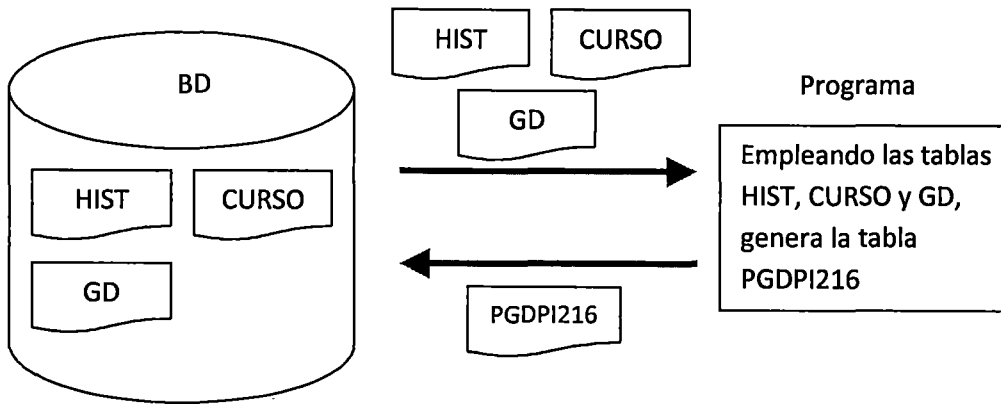


Figura 3.9 Esquema del Segundo Bloque

3.3.2 Aplicación del Programa en Java

El Programa en Java explicado se aplica a 7 cursos del Currículo de la especialidad de Ingeniería Química, que se listan en la tabla 3.6.

En la última columna de la tabla 3.6, se indica el número de registros obtenido para cada curso luego de la aplicación del Programa en Java.

3.3.3 Limpieza adicional de los datos obtenidos para las variables

Cuando examinamos las columnas de Notas de Pre-requisito1 y 2, en los archivos resultantes del Programa en Java, vemos que existen sitios vacíos. Esto es debido a que algunos alumnos han aprobado estos Pre-requisitos en los Períodos Académicos de Verano, por lo que su nota no aparece. Con el fin de evitar la perturbación de nuestro estudio, debido a la nota obtenida en un período académico no regular, se eliminaran los registros correspondientes.

NOMBRE DEL CURSO	CÓDIGO DEL CURSO	NÚMERO DE PRE-REQUISITOS	CICLO	AREA ACADÉMICA	NÚMERO DE REGISTROS
MATEMÁTICAS III	MA133	1	3	AACB	1865
FENÓMENOS DE TRANSPORTE	PI140	2	6	AAIQ	2411
TERMODINÁMICA PARA INGENIERÍA QUIMICA I	PI216	2	6	AAIQ	2118
CORROSIÓN I	PI515	1	8	AAIQ	1765
CINÉTICA QUÍMICA Y DISEÑO DE REACTORES I	PI225	1	9	AAIQ	1821
ECONOMÍA DE PROCESOS	PI510	2	9	AAIQ	1566
PLANEAMIENTO Y CONTROL DE LA PRODUCCION	PA136	2	10	AACC	1315

Tabla 3.6 Relación cursos estudiados y el número de registros

Por otro lado en la columna antigüedad del alumno, vemos que existen alumnos hasta con 17 años de antigüedad, lo que también perturba nuestro estudio, que busca predecir en base a estudiantes que tengan los mismos perfiles. Es por ello que eliminaremos a los estudiantes cuya antigüedad sea mayor de 10 años.

Finalmente solo consideraremos a los estudiantes cuyos registros en los cursos en estudio hayan empezado en el Período Académico 1993-1.

Luego de realizar la eliminación de los registros correspondientes a los tres casos mencionados para los 7 cursos, se obtiene tabla 3.7, donde se dan los número de registros antes y después de esta limpieza adicional.

NOMBRE DEL CURSO	CÓDIGO DEL CURSO	NÚMERO DE REGISTROS ANTES	NÚMERO DE REGISTROS DESPUÉS DE LIMPIEZA ADICIONAL
MATEMÁTICAS III	MA133	1865	1276
FENÓMENOS DE TRANSPORTE	PI140	2411	1262
TERMODINÁMICA PARA INGENIERÍA QUIMICA I	PI216	2118	1302
CORROSIÓN I	PI515	1765	1497
CINÉTICA QUÍMICA Y DISEÑO DE REACTORES I	PI225	1821	1273
ECONOMÍA DE PROCESOS	PI510	1566	994
PLANEAMIENTO Y CONTROL DE LA PRODUCCION	PA136	1315	851

Tabla 3.7 Números de registros antes y después de la limpieza

Los archivos Excel para los 7 cursos con la limpieza adicional realizada, poseen 13 campos o columnas para los cursos que tienen 2 pre-requisitos y 12 campos para los que tienen un solo pre-requisito.

A estos siete archivos, que están ordenados de menor a mayor, primero por orden del período académico y luego por orden de código de alumno, se les elimina las columnas, código de alumno, código de curso, curso pre-requisito 1, curso pre-requisito 2, con el fin de

quedarnos con el campo período académico y los campos que representan a la variable dependiente y las variables predictoras.

Solamente para la aplicación de las técnicas que permiten predecir si el alumno aprueba o no un curso en el que se desea inscribir, la columna Nota se cambiará por la columna Condición (Aprobado, representado por 1 , cuando la Nota es mayor o igual a diez ó Desaprobado, representado por 0, cuando la nota es menor a 10). En la tabla 3.8, se dan los nombres de cada uno de los archivos y los campos que poseen. Estos archivos están listos para la aplicación de las técnicas de predicción.

3.4 Transformación de los datos para las variables (normalización)

Para mejorar el trabajo de las funciones de transferencia y la habilidad de la red para generalizar, los valores de las variables se normalizan con la función:

$$X' = \frac{X - X\min}{X\max - X\min} \qquad \text{(Ec. 3.1)}$$

que produce una salida en el rango (-1, 1) y donde:

X : valor de la variable
X' : valor de la variable después de la normalización
$X\max$: valor máximo de la variable
$X\min$: valor mínimo de la variable

OBJETIVO	TÉCNICA DE PREDICCIÓN	N° de Pre-Requisitos	CURSOS/ NOMBRE DE ARCHIVO	C A M P O S O COLUMNAS								
PREDICCION DE SI APRUEBA O NO UN CURSO	REDES NEURONALES Y REGRESION LOGISTICA	1	MA133 /EXCEL MA133 A/D PI515 /EXCEL PI515 A/D PI225 /EXCEL PI225 A/D	perAca	ppa	notPreReq1		pGD	spgd	sumCre	antAlu	CND
		2	PI140 /EXCEL PI140 A/D PI216 /EXCEL PI216 A/D PI510 /EXCEL PI510 A/D PA136 /EXCEL PI136 A/D	perAca	ppa	notPreReq1	notPreReq2	pGD	spgd	sumCre	antAlu	CND
PREDICCIÓN DE LA NOTA QUE OBTENDRÁ EN UN CURSO	REDES NEURONALES Y REGRESIÓN MÚLTIPLE	1	MA133 /EXCEL MA133 NOTA PI515 /EXCEL PI515 NOTA PI225 /EXCEL PI225 NOTA	perAca	ppa	notPreReq1		pGD	spgd	sumCre	antAlu	notCur
		2	PI140 /EXCEL PI140 NOTA PI216 /EXCEL PI216 NOTA PI510 /EXCEL PI510 NOTA PA136 /EXCEL PA136 NOTA	perAca	ppa	notPreReq1	notPreReq2	pGD	spgd	sumCre	antAlu	notCur

Tabla 3.8 Relación de campos existentes en los archivos para aplicar las técnicas de predicción

3.5 Elección de los conjuntos de datos para obtención del Modelo y Pronóstico

Para la obtención del modelo en los cuatro casos se utilizarán los registros obtenidos desde el período académico 1993-1 hasta 2009-2, y para realizar el pronóstico se usará los registros correspondientes a los períodos académicos 2010-1 y 2010-2. Esto se encuentra descrito en la tabla 3.9, donde se también se dan los números de registros para cada conjunto.

CURSO	N° TOTAL DE REGISTROS	PARA EL MODELO		PARA EL PRONÓSTICO	
		RANGO 1	N° DE REGISTROS	RANGO2	N° DE REGISTROS
MA133	1276	1993-2 AL 2009-2	1217	2010-1 AL 2010-2	59
PI140	1262	1994-1 AL 2009-2	1199	2010-1 AL 2010-2	63
PI216	1302	1994-2 AL 2009-2	1237	2010-1 AL 2010-2	65
PI515	1497	1993-2 AL 2009-2	1401	2010-1 AL 2010-2	96
PI225	1273	1993-2 AL 2009-2	1205	2010-1 AL 2010-2	68
PI510	994	1993-2 AL 2009-2	930	2010-1 AL 2010-2	64
PA136	851	1993-2 AL 2009-2	819	2010-1 AL 2010-2	32

Tabla 3.9 Conjunto de datos para el modelo y para el pronóstico

CAPITULO IV

APLICACIÓN DE LOS MODELOS PREDICTIVOS

4.1 Modelo de red neuronal. Diseño

Se aplicará tanto a la predicción de si aprobará o no un curso como a la predicción de la nota que obtendrá en el curso

El diseño de la red neuronal se ha realizado en varios pasos: selección de la red, del algoritmo de entrenamiento, del número y tamaño de las capas, construcción de los conjunto de entrenamiento, validación y pronóstico, y determinación del número de neuronas en las capas ocultas. Las secciones citadas arriba se describen a continuación.

4.1.1 Arquitectura de la Red

Se utiliza la arquitectura de red alimentada hacia adelante o feedforward, ya que las conexiones entre neuronas se establecen en un único sentido, en el siguiente orden: capa de entrada, capa(s) oculta(s) y capa de salida. Es una red totalmente interconectada, ya que los nodos de cada capa están conectados con todos los nodos de la capa siguientes. La capa de entrada tiene como única misión distribuir la información que llega a la red neuronal para su procesamiento en la primera capa oculta. La función de activación f(x) de las capas ocultas y de salida es de tipo tangente hiperbólica:

$$f(x) = \frac{2}{1 + e^{(-2x)}} - 1$$

La inicialización de los pesos y umbrales de cada capa se realiza tomando valores en el intervalo de [-1; 1].

4.1.2 Algoritmo de Entrenamiento

El algoritmo que se utiliza es el de retropropagación, y la actualización de los pesos se realiza aplicando el algoritmo de Gradiente Conjugado. Éste asegura rapidez en el proceso de convergencia comparado con otros algoritmos de entrenamiento para un mayor volumen de datos [5].

4.1.3 Número y tamaño de las capas

El número de datos que se ingresan para las variables predictoras son 6 (en el caso de cursos con un solo pre-requisito) o de 7 (en caso de cursos con dos pre-requisitos). Así para el caso de cursos con 2 pre-requisitos las variables predictoras son: **ppa**, **notPreReq1**, **notPreReq2**, **pGD**, **spgd**, **sumCre** y **antAlu**. Mientras se tiene una sola variable dependiente, la nota del curso (**notCur**) cuando se desea predecir la nota de un curso, ó la condición de aprobado/desaprobado (**CND**), cuando se desea predecir si aprobará o no un curso.

La red seleccionada contiene dos capas de neuronas ocultas entre las capas de entrada y de salida. El número de neuronas para las capas ocultas ha sido determinado a través de un método heurístico de prueba y error, asignando siempre a la primera capa oculta el doble de neuronas que la segunda capa oculta [5]. En la figura 4.1 se puede ver la estructura de la red.

4.1.4 Construcción de los conjuntos de entrenamiento, validación y pronóstico

Se considera que una red neuronal artificial ha sido entrenada con éxito si se ajusta bien a los valores de los patrones de entrenamiento y proporciona interpolaciones suaves para el conjunto de datos no entrenados.

Con los archivos que se mencionan en la tabla 3.8, de la sección 3.3.3, donde se hallan los datos para todas las variables independientes y

dependiente, se procede a establecer los conjuntos de entrenamiento, validación y pronóstico.

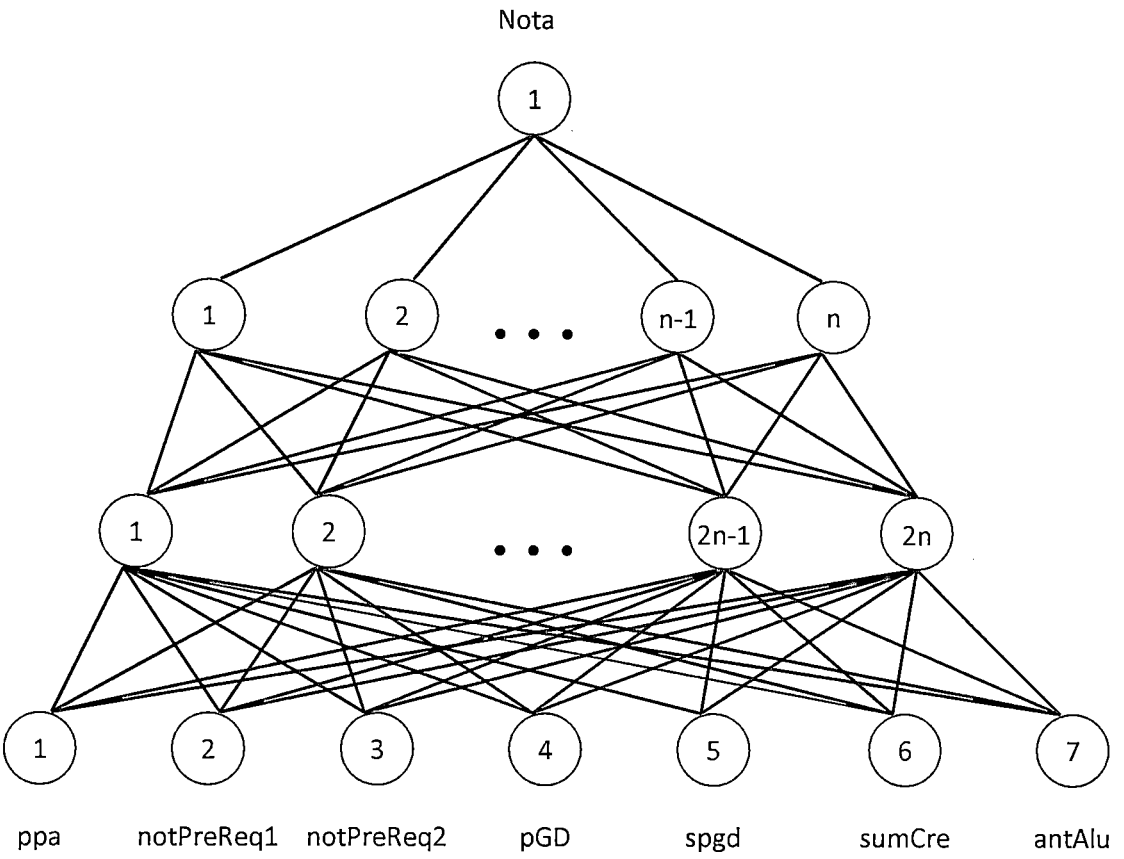


Figura 4.1 Estructura de la Red Neuronal

Para el conjunto de entrenamiento y validación se han considerado los registros existentes en los períodos académicos desde el 1993-2 hasta el 2009-2, de los cuales el 66% serán tomados para entrenamiento y el 34% para validación. Para el pronóstico se han considerado los registros de los períodos académicos 2010-1 y 2010-2.

En las siguientes tablas 4.1 y 4.2 se muestran la estructura de la matriz de entradas y de la matriz de salida de la red neuronal.

Entrada	ppa	notPreReq1	notPreReq2	pGD	spgd	sumCre	antAlu
Muestra 1							
Muestra 2							
.							
.							
.							
Muestra N							

Tabla 4.1: Matriz de entrada de la red neuronal

Entrada	Nota ó CND
Muestra 1	
Muestra 2	
.	
.	
.	
Muestra N	

Tabla 4.2: Matriz de salida de la red neuronal

4.1.5 Determinación del número de neuronas en las capas ocultas

Para determinar el número de neuronas en las capas ocultas se han considerado tres posibles configuraciones y para cada configuración se han realizado siete simulaciones (uno para cada curso en estudio). En la tabla 4.3 se resumen los resultados obtenidos para el caso de la predicción de la nota de un curso.

A la vista de los resultados, podemos observar que la variación en el error (raíz cuadrada del error cuadrático medio) entre el primer modelo ([7, 6, 3, 1]) y los otros dos modelos ([7, 8, 4, 1] y [7, 10, 5, 1]) es menor al 0.5%, por lo tanto para optimizar el tiempo de ejecución del modelo podemos elegir al primer modelo ([7, 6, 3, 1].

Esto también se da para el caso de la predicción de si aprobará o desaprobará un curso. En la figura 4.2, se muestra la estructura final de la red.

Red neuronal	Curso	Error	Error promedio
[7, 6, 3, 1]	MA133	0,1397	0,1577
	PI140	0,1998	
	PI216	0,1834	
	PI515	0,1655	
	PI225	0,1558	
	PI510	0,1251	
	PA136	0,1344	
[7, 8, 4, 1]	MA133	0,1396	0,1570
	PI140	0,1991	
	PI216	0,1813	
	PI515	0,1650	
	PI225	0,1559	
	PI510	0,1243	
	PA136	0,1336	
[7, 10, 5, 1]	MA133	0,1394	0,1571
	PI140	0,1981	
	PI216	0,1818	
	PI515	0,1657	
	PI225	0,1562	
	PI510	0.1244	
	PA136	0,1341	

Tabla 4.3 Determinación de neuronas en la capa oculta

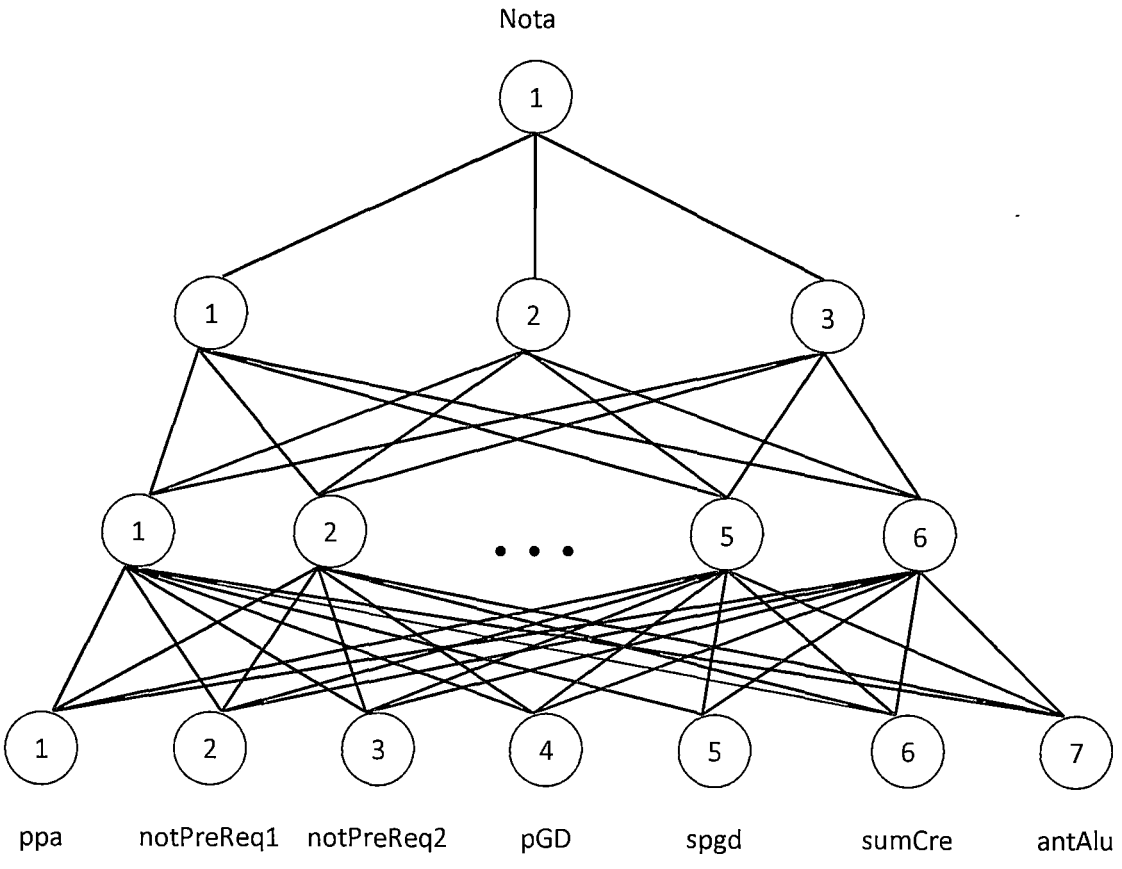


Figura 4.2 Estructura final de la red

4.1.6 Esquema de la Red Neuronal para el Excel

La figura 4.3, presenta el esquema de funcionamiento de la red neuronal para predecir la nota mediante el Excel [7, 6, 3, 1], en donde se presentan las siguientes zonas: Pesos (incluye el valor del umbral), Datos, Capa de entrada (con los datos normalizados), las dos Capas Ocultas, la Capa de Salida y el Error. El entrenamiento de la red mediante la actualización de pesos se realiza empleando el **Solver** del Excel. El complemento Solver permite actualizar la zona de los pesos minimizando el error de entrenamiento (E1) , sujeto a que el error de validación (E2), sea menor ó igual a una tolerancia pre-establecida.

Luego del entrenamiento y validación del modelo neuronal (estableciendo los pesos, incluyendo el umbral), se realizan los cálculos de los errores del modelo (E1+E2) y el error del pronóstico (E3), y del cálculo del porcentaje de aciertos para caso de predecir si aprobará o desaprobará un curso.

El esquema de funcionamiento de la red neuronal para la predicción de si aprobará o no un curso, es similar al esquema de la figura 4.3, con la diferencia, que una vez definido los pesos, incluyendo el umbral, se realizan los cálculos del porcentaje de aciertos tanto para el conjunto de entrenamiento y validación, como para el conjunto separado para pronóstico (ver figura 4.4).

4.2 Modelo de Regresión Múltiple

Este modelo será aplicado para predecir la nota de un curso en el cual el alumno desea inscribirse.

La hipótesis de trabajo ha sido planteada en la Sección 1.8.2.4, denominada Hipótesis Específica N° 4, que a letra dice "La técnica de regresión múltiple aplicadas a la universidad peruana, permitiría al estudiante predecir la nota que obtendrá en un curso en que se desea inscribir en un nuevo ciclo"

4.2.1 Selección de las variables

Las variables han sido seleccionadas en la sección 3.2 y son las que aparecen en la tabla 3.8 . En esta misma tabla aparece también el nombre de los archivos para cada uno de los 7 cursos (por ejemplo archivo EXCEL MA133 NOTA), donde se encuentra los datos ya preparados, para la aplicación de cualquiera de las técnicas de predicción

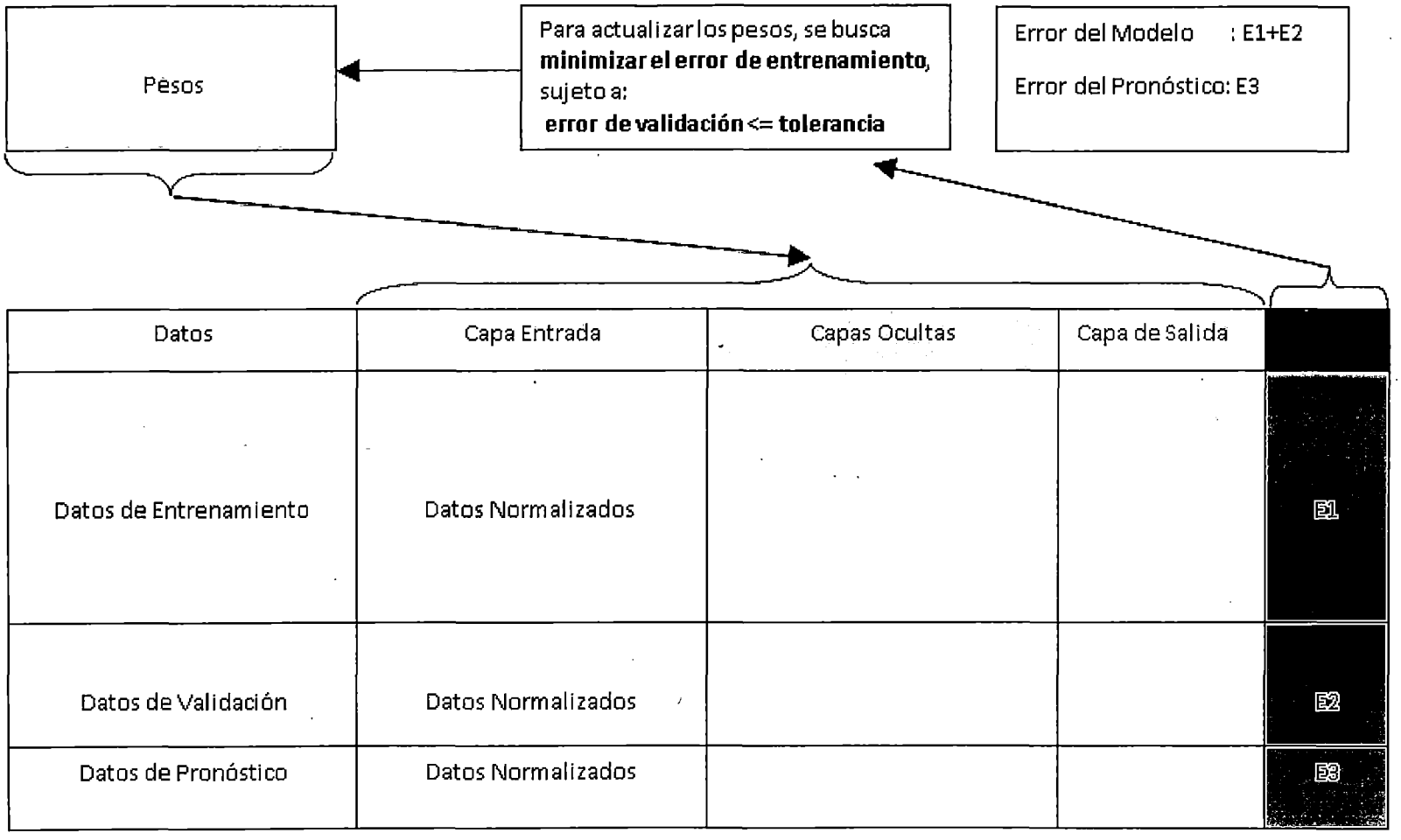


Figura 4.3: Muestra el esquema de funcionamiento de la red neuronal para predecir la nota mediante el Excel

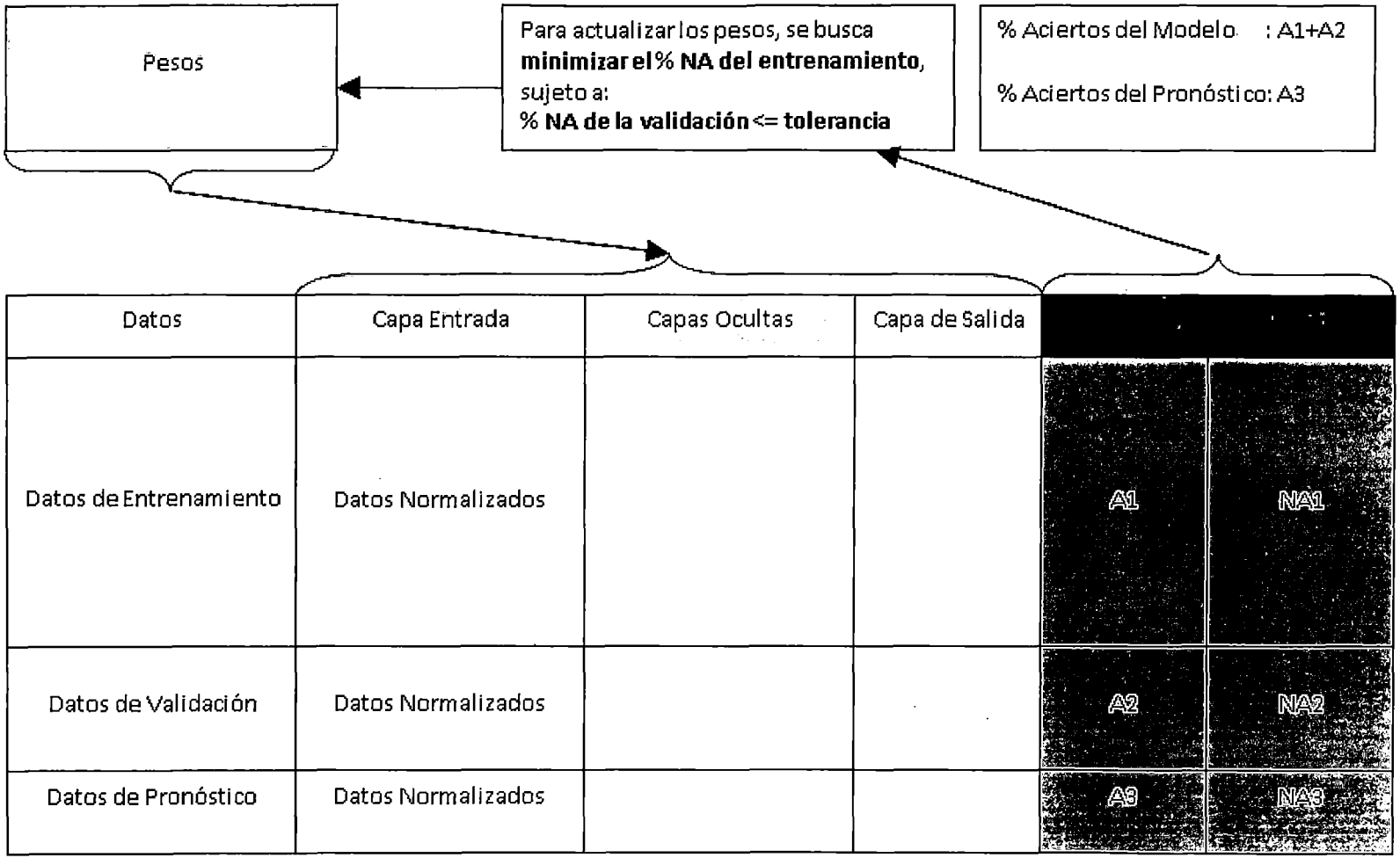


Figura 4.4: Muestra el esquema de funcionamiento de la red neuronal mediante el Excel para el caso cualitativo

4.2.2 Elección de los conjuntos para el modelo y para el pronóstico

En la Sección 3.5 se realizó la elección de estos conjuntos, y se encuentra especificado en la tabla 3.9.

4.2.3 Esquema de trabajo usando Excel

En una hoja Excel se pega la información del archivo Excel (por ejemplo archivo EXCEL PI140 NOTA), que contiene los datos de las 8 variables (7 predictoras y una dependiente). Estos datos son normalizados utilizando la Ecuación 3.1 y son usados como datos de entrada en la Herramienta de Análisis de Datos, llamada "Regresión" del Excel. De la aplicación de "Regresión" se obtiene un resumen de la regresión, que tiene como resultados principales:

El coeficiente de correlación R^2.
El estadístico F de la regresión y el valor p.
Los estadísticos t y el valor p para cada variable.
Los coeficientes estimados para cada variable y el intercepto.
Estos resultados permiten contrastar la hipótesis de trabajo y determinar el grado de significancia de las variables independientes. Así tenemos que R^2 es la relación entre la variabilidad explicada por el modelo y la variabilidad total. Esto significa que un valor de R^2 cercano a uno, indica que las variables predictoras explican bien la variabilidad observada en la variable dependiente, mientras valores bajos de R^2 indican que lo explican parcialmente. Esto último en muchos casos indica que no se han tomado en cuenta otras variables, que pueden tener fuerte influencia en la predicción.

En el caso de que algunas variables, no sean significativas, se debe reformular el modelo con un menor número de variables independientes

y volver correr la Herramienta de Análisis "Regresión". Este esquema se encuentra descrito en la figura 4.5.

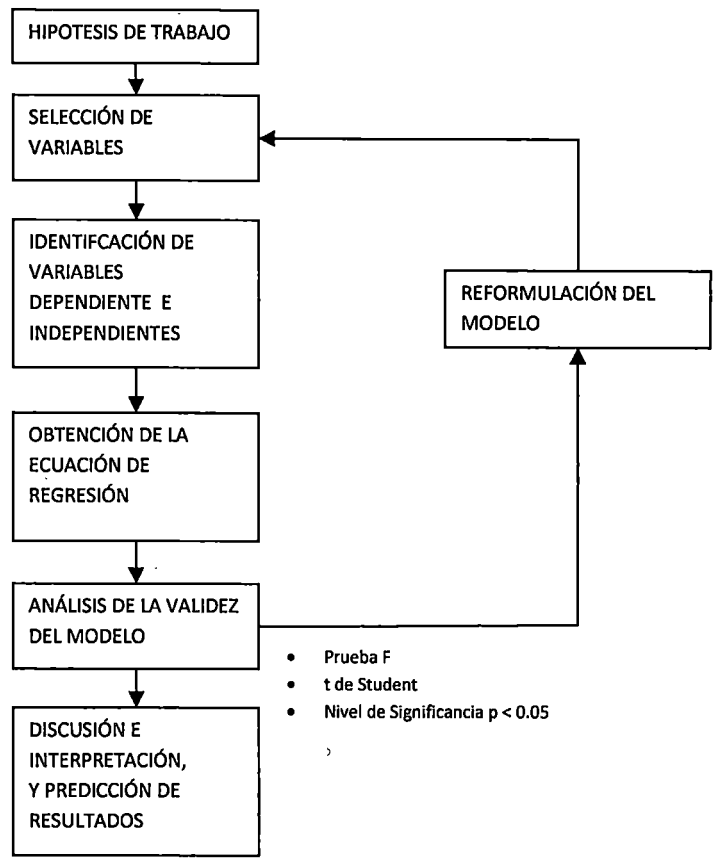


Figura 4.5 Pasos para el Modelo de Regresión

Finalmente en la misma hoja Excel con ayuda de los coeficientes estimados se calcula la nota para el conjunto de datos elegidos para pronosticar.

4.2.4 Pruebas de contrastes

4.2.4.1 Prueba de significancia general de una regresión múltiple: La Prueba F

Con el modelo de regresión con k variables

$Yi = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$

Para probar la hipótesis

Ho: $\beta_2 = \beta_3 = \ldots = \beta_k = 0$

(es decir, todos los coeficientes de pendiente son simultáneamente cero) frente a

H1: no todos los coeficientes de pendiente son simultáneamente cero

Se debe calcular

$$F = \frac{SCE/gl}{SCR/gl} = \frac{SCE/(k-1)}{SCR/(n-k)}$$

Si F > F_α(k-1, n-k), rechace H_0; de lo contrario, no la rechace, donde F_α(k-1, n-k) es el valor F crítico en el nivel de significancia α, y (k-1) es "gl" en el numerador y (n-k) es el "gl" en el denominador. Por otra parte, si el valor p del F obtenido es lo bastante bajo se puede rechazar H_0.

Según Leamer y Schwartz, la hipótesis nula se rechaza cuando el valor de F calculado sea superior al logaritmo natural del tamaño muestral.

n = número de observaciones

k = número de variables totales (variables dependientes + variables independientes)

4.2.4.2 Prueba de hipótesis sobre los coeficientes de regresión individuales

Con el modelo de regresión con k variables

$Yi = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$

Para probar la hipótesis

Ho: $\beta_k = 0$ frente a

H_1: $\beta_k \neq 0$

Se debe calcular

$$t = \frac{\hat{\beta}_k - \beta_k}{ee(\hat{\beta}_k)}$$

Si el valor de t calculado es mayor que el t crítico en el nivel de confianza α escogido, se debe rechazar la hipótesis nula, de lo contrario, la hipótesis nula es válido.

En la práctica, no se tiene que suponer un valor particular de α para llevar a cabo la prueba de hipótesis. Lo que se hace es consultar la tabla estadística para "t" y encontrar la probabilidad real de obtener un valor del estadístico de prueba tan grande como el valor de t calculado. Esta probabilidad se denomina valor p (es decir, valor de probabilidad), también conocido como nivel de significancia. Más técnicamente, el valor p se define como el nivel de significancia más bajo al cual puede rechazarse la hipótesis nula. Lo entenderemos mejor si definimos p como la probabilidad de cometer error si se rechaza la hipótesis nula.

4.3 Modelo de Regresión Logística

Este modelo será aplicado para predecir si un alumno aprobará o no un curso en el cual desea inscribirse.

La hipótesis de trabajo ha sido planteada en la Sección 1.8.2.2, denominada Hipótesis Específica N° 2, que a letra dice "La técnica de regresión logística aplicada a la universidad peruana, permitiría al estudiante predecir si aprobará o no un curso en que se desea inscribir en un nuevo ciclo"

4.3.1 Selección de las variables

Las variables han sido seleccionadas en la Sección 3.2 y son las que aparecen en la tabla 3.8. En esta misma tabla aparece también el nombre de los archivos para cada uno de los 7 cursos (por ejemplo archivo EXCEL MA133 A/D) donde se encuentra los datos ya preparados, para la aplicación de cualquiera de las técnicas de predicción.

4.3.2 Elección de los conjuntos para el modelo y para el pronóstico

En la sección 3.5 se realizó la elección de estos conjuntos, y se encuentra especificado en la tabla 3.9

4.3.3 Esquema de trabajo

En una hoja Excel se pega la información del archivo Excel (por ejemplo archivo EXCEL PI140 NOTA), que contiene los datos de las 8 variables (7 predictoras y una dependiente). Estos datos son normalizados utilizando la Ecuación 3.1 y son usados como datos de entrada en "Analizar" del Software SPSS, para realizar Regresión Logística binaria. De esta aplicación se obtiene resultados sobre el modelo y las variables, siendo los principales:

- Prueba del Chi cuadrado sobre las variables del modelo y su nivel de significancia
- Porcentaje de aciertos o Cuenta R^2
- El estadístico Wald para cada variable y su nivel de significancia.
- Los coeficientes estimados para cada variable y el intercepto.

Estos resultados permiten contrastar la hipótesis de trabajo y determinar el grado de significancia de las variables independientes. Finalmente esta aplicación también permite realizar la estimación de si

aprueba o no un curso, para el conjunto de datos elegidos para pronosticar.

4.3.4 Pruebas de contrastes

4.3.4.1 Prueba de contraste de los coeficientes del modelo en su conjunto

El Estadístico Chi cuadrado contrasta la hipótesis nula, de si los coeficientes del modelo en su conjunto son estadísticamente distintos de cero. El "valor p" o "nivel de significancia" se obtiene a partir del valor de Chi cuadrado y sus grados de libertad correspondiente. Si el nivel de significancia tiene un valor por debajo 0.05, entonces se puede rechazar la hipótesis nula y por lo tanto los coeficientes en conjunto son diferentes de cero.

4.3.4.2 Prueba de contraste sobre los coeficientes de las variables

El estadístico Wald permite contrastar la hipótesis de si los coeficientes de las variables son iguales a cero, y sigue una distribución χ^2, que con ayuda de su correspondiente grado de libertad, se puede hallar su "valor p" o "nivel de significancia". Este estadístico se halla a partir de:

$$Estadístico\ de\ Wald = \left(\frac{Coeficiente}{Error\ típico} \right)^2$$

Si el nivel de significancia tiene un valor por debajo 0.05, entonces se puede rechazar la hipótesis que el valor del coeficiente es igual a cero, y por lo tanto la variable correspondiente es significativa.

CAPITULO V

RESULTADOS DE LA INVESTIGACIÓN

5.1 Modelo de Red Neuronal aplicado para predecir si aprobará o no un curso

5.1.1 Resultados

En la tabla 5.1 (a) se muestran los porcentajes de aciertos de la aplicación de la Red neuronal de retropropagación a 7 cursos.

PORCENTAJE DE ACIERTOS O INSTANCIAS CORRECTAMENTE CLASIFICADAS		
CURSO	EN EL MODELO	EN EL PRONÓSTICO
MA133	69.43%	74.58%
PI140	64.39%	65.08%
PI216	65.64%	63.08%
PI515	70.66%	73.96%
PI225	67.22%	70.59%
PI510	81.29%	81.25%
PA136	74.48%	81.25%
PROMEDIO	70.45%	72.83%

Tabla 5.1 (a) Porcentaje de aciertos de la predicción de Aprobado/Desaprobado en siete cursos

En la tabla 5.1 (b) se dan los valores de las Tablas de Clasificación para cada uno de los 7 cursos.

MODELO DE REDES NEURONALES

1 = APROBADO
0 = DESAPROBADO

TABLA DE CLASIFICACION DEL MODELO

MA133

1217	0	1	PORCENTAJE
0	19	358	5.04
1	14	826	98.33
PORCENTAJE GLOBAL			69.43

PI140

1199	0	1	PORCENTAJE
0	318	292	52.13
1	135	454	77.08
PORCENTAJE GLOBAL			64.39

PI216

1237	0	1	PORCENTAJE
0	267	245	52.15
1	180	545	75.17
PORCENTAJE GLOBAL			65.64

PI515

1401	0	1	PORCENTAJE
0	95	355	21.11
1	56	895	94.11
PORCENTAJE GLOBAL			70.66

PI225

1205	0	1	PORCENTAJE
0	88	315	21.84
1	80	722	90.02
PORCENTAJE GLOBAL			67.22

PI510

930	0	1	PORCENTAJE
0	33	146	18.44
1	28	723	96.27
PORCENTAJE GLOBAL			81.29

PA136

819	0	1	PORCENTAJE
0	2	207	0.96
1	2	608	99.67
PORCENTAJE GLOBAL			74.48

TABLA DE CLASIFICACION DEL PRONOSTICO

MA133

59	0	1	PORCENTAJE
0	3	12	20.00
1	3	41	93.18
PORCENTAJE GLOBAL			74.58

PI140

63	0	1	PORCENTAJE
0	10	12	45.45
1	10	31	75.61
PORCENTAJE GLOBAL			65.08

PI216

65	0	1	PORCENTAJE
0	19	13	59.38
1	17	22	56.41
PORCENTAJE GLOBAL			57.75

PI515

96	0	1	PORCENTAJE
0	5	13	27.78
1	12	66	84.62
PORCENTAJE GLOBAL			73.96

PI225

68	0	1	PORCENTAJE
0	7	18	28.00
1	2	41	95.35
PORCENTAJE GLOBAL			70.59

PI510

64	0	1	PORCENTAJE
0	2	10	16.67
1	2	50	96.15
PORCENTAJE GLOBAL			81.25

PA136

32	0	1	PORCENTAJE
0	0	6	0.00
1	0	26	100.00
PORCENTAJE GLOBAL			81.25

Tabla 5.1 (b) Tabla de Clasificación o Matriz de confusión

5.1.2 Análisis de los resultados

El porcentaje de aciertos de la técnica de Redes Neuronales aplicado a los 7 cursos, tomando en cuenta los datos para el modelo (de entrenamiento y validación) se hallan entre el 64.4% y el 81.3%, mientras en el pronóstico varía entre el 63.1% y el 81.3%.

Estos niveles de aciertos, nos permite probar la Hipótesis Específica Nº 1, de la Sección 1.8.2.1, que a la letra dice: "Las redes neuronales de retropropagación aplicadas a la universidad peruana, permitiría al estudiante predecir si aprobará o no un curso en que se desea inscribir en un nuevo ciclo".

El promedio del porcentaje de aciertos en los 7 cursos para los pronósticos es mayor que para el modelo. Esto nos asegura que, efectivamente los datos de los períodos académicos pasados, utilizados como entrenamiento y validación, pueden ser utilizados para pronosticar bien en los períodos académicos más recientes.

Los cursos de PI510 y PA136, que pertenecen al noveno y décimo ciclo respectivamente, tienen los mayores porcentajes de aciertos, que el resto de los cursos que pertenecen al octavo ciclo o más bajos. Esta diferencia del porcentaje de aciertos entre los 7 cursos, nos indica que cada curso tiene características bastante propias.

5.2 Modelo de Regresión logística aplicado a predecir si aprobará o no un curso

5.2.1 Resultados

De la aplicación de la técnica de Regresión logística se han obtenido las siguientes Tablas:

- Tabla 5.2: Prueba sobre los coeficientes del modelo en conjunto en la predicción de Aprobado/Desaprobado en siete cursos.
- Tabla 5.3: Porcentaje de aciertos de la predicción de Aprobado/Desaprobado en siete cursos.
- Tabla 5.4: Estadístico de Wald para cada variable para 7 cursos.
- Tabla 5.5: Coeficientes de Regresión para 7 cursos.
- Tabla 5.6 (a) Tabla de Clasificación o Matriz de Confusión
- Tabla 5.6 (b) Comparación del modelo con el corregido

que se muestran a continuación:

PRUEBA SOBRE LOS COEFICIENTES DEL MODELO EN CONJUNTO			
CURSO	CHI CUADRADO	GRADOS DE LIBERTAD	NIVEL DE SIGNIFICANCIA
MA133	120.0	6	0.000
PI140	78.7	7	0.000
PI216	140.6	7	0.000
PI515	187.6	6	0.000
PI225	126.5	6	0.000
PI510	160.2	7	0.000
PA136	77.1	7	0.000

Tabla 5.2 Prueba sobre los coeficientes del modelo en conjunto en la predicción de Aprobado/Desaprobado en siete cursos

PORCENTAJE DE ACIERTOS O INSTANCIAS CORRECTAMENTE CLASIFICADAS		
CURSO	EN EL MODELO	EN EL PRONÓSTICO
MA133	71.16%	76.27%
PI140	60.55%	68.25%
PI216	65.40%	69.23%
PI515	70.59%	76.04%
PI225	68.38%	69.12%
PI510	81.94%	81.25%
PA136	74.73%	78.13%
PROMEDIO	70.39%	74.04%

Tabla 5.3 Porcentaje de aciertos de la predicción de Aprobado/Desaprobado en siete cursos

	MA133		PI140		PI216		PI515		PI225		PI510		PA136	
	Wald	Sign.	Wald	Sign.	Wald	Sign.	Wald	Sign.	Wald	Sign.	Wald	Sign.	Wald	Sign.
X1ppa	37,654	0,000	18,052	0,000	4,431	0,035	45,486	0,000	33,047	0,000	24,518	0,000	17,994	0,000
X2notPreReq1	5,401	0,020	2,566	0,109	15,565	0,000	1,446	0,229	1,369	0,242	1,717	0,190	1,204	0,272
X3notPreReq2			7,697	0,006	3,723	0,054					3,046	0,081	0,532	0,466
X4pGD	16,909	0,000	0,021	0,885	7,641	0,006	0,703	0,402	2,556	0,110	3,199	0,074	5,576	0,018
X5spgd	7,330	0,007	3,961	0,047	5,331	0,021	5,338	0,021	0,000	0,998	5,313	0,021	13,246	0,000
X6sumCre	9,684	0,002	8,537	0,003	0,926	0,336	4,730	0,030	0,244	0,622	5,241	0,022	11,052	0,001
X7antAlu	2,466	0,116	0,720	0,396	40,575	0,000	33,994	0,000	14,498	0,000	41,661	0,000	6,201	0,013

Tabla 5.4 Estadístico de Wald para cada variable para 7 cursos

	COEFICIENTES DE REGRESION						
	MA133	PI140	PI216	PI515	PI225	PI510	PA136
X1ppa	3,632	2,209	1,031	3,517	3,332	4,072	2,926
X2notPreReq1	1,155	0,512	1,454	0,475	0,463	0,732	0,641
X3notPreReq2		0,951	0,694			1,358	0,415
X4pGD	1,893	0,045	1,023	0,326	0,735	-0,789	-0,679
X5spgd	-2,837	1,694	-2,040	-2,285	0,003	-3,436	-4,064
X6sumCre	3,619	-2,406	0,772	2,108	0,515	3,294	3,662
X7antAlu	-1,054	-0,349	-2,560	-1,716	-1,195	-3,009	-1,108

Tabla 5.5 Coeficientes de Regresión para 7 cursos

DEL MODELO				DEL PRONOSTICO			
MA133				**MA133**			
1217	0	1	PORCENTAJE	59	0	1	PORCENTAJE
0	75	302	19.89	0	8	7	53.33
1	49	791	94.17	1	7	37	84.09
PORCENTAJE GLOBAL			71.16	PORCENTAJE GLOBAL			76.27
PI140				**PI140**			
1199	0	1	PORCENTAJE	63	0	1	PORCENTAJE
0	403	207	66.07	0	13	9	59.09
1	266	323	54.84	1	11	30	73.17
PORCENTAJE GLOBAL			60.55	PORCENTAJE GLOBAL			68.25
PI216				**PI216**			
1237	0	1	PORCENTAJE	65	0	1	PORCENTAJE
0	217	295	42.38	0	14	12	53.85
1	133	592	81.66	1	8	31	79.49
PORCENTAJE GLOBAL			65.40	PORCENTAJE GLOBAL			69.23
PI515				**PI515**			
1401	0	1	PORCENTAJE	96	0	1	PORCENTAJE
0	125	325	27.78	0	5	13	27.78
1	87	864	90.85	1	10	68	87.18
PORCENTAJE GLOBAL			70.59	PORCENTAJE GLOBAL			76.04
PI225				**PI225**			
1205	0	1	PORCENTAJE	68	0	1	PORCENTAJE
0	105	298	26.05	0	5	20	20.00
1	83	719	89.65	1	1	42	97.67
PORCENTAJE GLOBAL			68.38	PORCENTAJE GLOBAL			69.12
PI510				**PI510**			
930	0	1	PORCENTAJE	64	0	1	PORCENTAJE
0	40	139	22.35	0	4	8	33.33
1	29	722	96.14	1	4	48	92.31
PORCENTAJE GLOBAL			81.94	PORCENTAJE GLOBAL			81.25
PA136				**PA136**			
819	0	1	PORCENTAJE	32	0	1	PORCENTAJE
0	27	182	12.92	0	0	6	0.00
1	25	585	95.90	1	1	25	96.15
PORCENTAJE GLOBAL			74.73	PORCENTAJE GLOBAL			78.13

Tabla 5.6 (a) Tabla de Clasificación o Matriz de Confusión

5.2.2 Análisis de los resultados

De la tabla 5.2 los valores de Chi cuadrado obtenido de la Regresión logística para los 7 cursos se hallan entre 77.1 y 160.2, todos ellos con

valores del nivel de significancia cercanos a cero, muy por debajo de 0.05. Esto nos permite rechazar la hipótesis nula y por lo tanto probar la Hipótesis Específica Nº 2, de la sección 1.8.2.2, que a la letra dice: "La técnica de regresión logística aplicada a la universidad peruana, permitiría al estudiante predecir si aprobará o no un curso en que se desea inscribir en un nuevo ciclo".

PORCENTAJE DE ACIERTOS O INSTANCIAS CORRECTAMENTE CLASIFICADAS		
CURSO	EN EL MODELO	EN EL MODELO CORREGIDO
MA133	71.16%	70.91%
PI140	60.55%	60.13%
PI216	65.40%	65.40%
PI515	70.59%	70.38%
PI225	68.38%	69.13%
PI510	81.94%	81.18%
PA136	74.73%	74.24%
PROMEDIO	70.39%	70.20%

Tabla 5.6 (b) Comparación del modelo con el corregido

De la tabla 5.3, se obtiene que el porcentaje de aciertos en la aplicación de la regresión logística para los 7 cursos, en el modelo se halla entre el 60.6% y el 81.9%, mientras en el pronóstico varía entre el 68.3% y el 81.3%. Estos niveles de aciertos, nos indica la buena bondad de ajuste del modelo de Regresión logística para predecir si un alumno aprobará o no un curso en que se desea inscribir.

El promedio del porcentaje de aciertos en los 7 cursos para los pronósticos es mayor que para el modelo. Esto nos asegura que, efectivamente los datos de los períodos académicos pasados, utilizados para estimación de los coeficientes, pueden ser utilizados para pronosticar bien en los períodos académicos más recientes.

Como una observación particular indicamos, que los cursos de PI510 y PA136, que pertenecen al noveno y décimo ciclo respectivamente, tienen los dos mayores porcentajes de aciertos, comparado con el resto de cursos que pertenecen al octavo ciclo o más bajos.

La tabla 5.4 nos proporciona el estadístico de Wald y su nivel de significancia, cuando este nivel de significancia es menor que 0.05, podemos inferir que la variable es significativa. En esta misma tabla, los valores de nivel de significancia de las variables que son significativas, han sido resaltados en amarillo Si bien es cierto algunas variables no son significativas, el alto valor del estadístico Chi cuadrado y los valores bajísimos de sus niveles de significancia, nos indican que estas variables no significativas de todas formas contribuye, aunque en menor grado a la bondad del ajuste de la regresión. Esto último mencionado se verifica de la tabla 5.6 (b), donde el porcentaje de aciertos de la regresión considerando solo las variables significativas disminuyen muy ligeramente para cada curso, comparándolo con la regresión que considera a todas las variables.

Además podemos sacar algunas observaciones adicionales de la tabla 5.4:

- Las 7 variables son siempre significativas para uno o más cursos de los siete en estudio.
- La variable "promedio ponderado acumulado" es significativa para los 7 cursos.

- Los cursos PI515 y PI225 tienen como variables significativas solamente al "promedio ponderado acumulado" y a la "antigüedad del alumno".

- El curso de MA133 tiene como variables significativas a todas las variables excepto la "antigüedad del alumno".

En la tabla 5.5 se dan los valores de los coeficientes de la ecuación de la regresión para el cálculo de la probabilidad, para cada uno de los 7 cursos, estando resaltado aquellos que corresponden a las variables significativas. En el Anexo 3, se da para el curso PI140 un ejemplo de cálculo de estimación de la Nota, considerando la regresión usando todas las variables, como usando solo las variables significativas.

5.3 Modelo de Red Neuronal aplicado a predecir la nota de un curso

5.3.1 Resultados

En la tabla 5.7 se presentan los errores, tomando como error "la raíz del error medio cuadrático" al aplicar la técnica de Redes Neuronales de Retropropagación a 7 cursos.

5.3.2 Análisis de los resultados

De la tabla 5.7 se ve que los errores de la aplicación de Redes Neuronales a 7 cursos, utilizando el conjunto de datos para entrenamiento y validación (en el modelo), se hallan entre 0.127 y 0.188, mientras en el pronóstico varía entre 0.100 y 0.182. Estos niveles bajos de los errores, nos permite probar la Hipótesis Específica Nº 3, de la Sección 1.8.2.3, que a la letra dice: "Las redes neuronales de retropropagación aplicadas a la universidad peruana, permitiría al estudiante predecir la nota que obtendrá en un curso en que se desea inscribir en un nuevo ciclo".

E R R O R		
CURSO	EN EL MODELO	EN EL PRONÓSTICO
MA133	0.1377	0.1369
PI140	0.1883	0.1818
PI216	0.1622	0.1654
PI515	0.1570	0.1320
PI225	0.1381	0.1266
PI510	0.1266	0.1353
PA136	0.1363	0.0997
Promedio	0.1495	0.1397

Tabla 5.7 Error (*) en la predicción de la Nota de siete cursos
(*) Raíz del error medio cuadrático

De la misma tabla, vemos también que el promedio de los errores de los 7 cursos para los pronósticos es menor que para el modelo. Esto nos asegura que, efectivamente los datos de los períodos académicos pasados, utilizados como entrenamiento y validación, pueden ser utilizados para pronosticar bien en los períodos académicos más recientes.

Una observación particular , es que los cursos de PI140 y PI216, que pertenecen al sexto ciclo y que se dictan en tres y dos secciones respectivamente, tienen los mayores niveles de error, que el resto de los cursos que tienen un nivel de errores esencialmente constante.

5.4 Modelo de Regresión Múltiple aplicado para predecir la nota de un curso

5.4.1 Resultados

Los resultados proporcionado por el modelo de regresión múltiple se presentan en las siguientes tablas:

5.4.2 Análisis de los resultados

De la tabla 5.8 los valores del Estadístico F para los 7 cursos se hallan entre 21.78 y 64.40, todos ellos con valores del nivel de significancia cercanos a cero (el valor más alto es 8.38×10^{-28}). Esto nos permite rechazar la hipótesis nula y por lo tanto probar la Hipótesis Específica Nº 4, de la Sección 1.8.2.4, que a la letra dice: "La técnica de regresión múltiple aplicada a la universidad peruana, permitiría al estudiante predecir la nota que obtendrá en un curso en que se desea inscribir en un nuevo ciclo".

De la misma tabla 5.8, los valores del Coeficiente de determinación van desde 0.113 y 0.281, esto nos indica, que las variables predictoras explican en un porcentaje que va del 11.3% al 28.1 % la variabilidad observada en la variable dependiente.

De la tabla 5.9, se obtiene que los errores en la aplicación de la regresión múltiple para los 7 cursos, en el conjunto de datos tomado para el modelo se hallan entre 0.121 y 0.185, mientras en el pronóstico varía entre el 0.101 y 0.168. Estos niveles bajos de los errores, nos indica la buena bondad de ajuste del modelo de Regresión múltiple para predecir si aprobará o no un curso en que un alumno se desea inscribir.

ESTADÍSTICO F Y COEFICIENTE DE DETERMINACION			
CURSO	ESTADÍSTICO F	NIVEL DE SIGNIFICANCIA	COEFICIENTE DE DETERMINACIÓN R^2
MA133	44.29	4.1105E-49	0.180
PI140	21.78	8.3823E-28	0.113
PI216	36.90	4.4531E-47	0.174
PI515	64.40	1.0145E-70	0.217
PI225	44.01	8.9915E-49	0.181
PI510	51.58	4.0889E-62	0.281
PA136	24.05	7.6773E-30	0.172

Tabla 5.8 Estadístico F y Coeficiente de determinación para el modelo de Regresión múltiple aplicado a 7 cursos

ERROR EN LA PREDICCION DE LA NOTA DE 7 CURSOS		
CURSO	EN EL MODELO	EN EL PRONÓSTICO
MA133	0.134	0.147
PI140	0.185	0.168
PI216	0.154	0.168
PI515	0.147	0.137
PI225	0.135	0.122
PI510	0.121	0.121
PA136	0.124	0.101
PROMEDIO	0.143	0.138

Tabla 5.9 Error (*) en la predicción de la Nota de 7 cursos

(*) Raíz del error medio cuadrático

	MA133		PI140		PI216		PI515		PI225		PI510		PA136	
	t	Sign.	t	Sign.	t	Sign.	t	Sign.	t	Sign.	t	Sign.	t	Sign.
X1ppa	7.164	1.35E-12	4.258	0.00002	2.558	0.01064	9.465	1.2E-20	6.383	2.47E-10	6.595	7.16E-11	5.299	1.50E-07
X2notPreReq1	5.471	5.43E-08	1.203	0.22936	4.640	3.87E-06	0.674	0.5004	1.813	0.070	1.319	0.187	2.080	0.038
X3notPreReq2			2.512	0.01214	2.365	0.01818					2.419	0.016	1.793	0.073
X4pGD	4.094	0.000045	0.337	0.73599	2.307	0.02122	0.105	0.9162	2.700	0.007	-1.579	0.115	-5.028	6.10E-07
X5spgd	-4.387	0.000012	3.961	0.00007	-2.948	0.00325	-1.543	0.1232	0.665	0.506	-2.481	0.013	-2.958	0.003
X6sumCre	5.152	3.01E-07	-3.173	0.00154	1.996	0.04610	1.885	0.0596	0.464	0.643	3.712	0.0002	2.829	0.005
X7antAlu	-3.036	0.00245	-2.555	0.01072	-9.198	1.52E-19	-8.463	6.5E-17	-6.832	1.33E-11	-8.578	4.03E-17	-4.383	1.32E-05

Tabla 5.10 Estadístico t y su nivel de significancia

	COEFICIENTES DE REGRESION						
	MA133	PI140	PI216	PI515	PI225	PI510	PA136
X1ppa	0.248863532	0.195663259	0.089603688	0.311037985	0.218814982	0.24884713	0.195114755
X2notPreReq1	0.144598212	0.034185474	0.117627552	0.015540332	0.041371172	0.028478485	0.055963248
X3notPreReq2		0.076538845	0.05979949			0.065313508	0.048956767
X4pGD	0.111909901	0.009465898	0.061404388	0.002600316	0.079181571	-0.030701337	-0.073048401
X5spgd	-0.279833588	0.302462409	-0.184420346	-0.096953149	0.041243341	-0.160558958	-0.164335508
X6sumCre	0.362814777	-0.233703302	0.114162416	0.11654639	0.029437958	0.233742027	0.156181405
X7antAlu	-0.121080217	-0.09344558	-0.256727918	-0.165702962	-0.134763719	-0.183703794	-0.106713524

Tabla 5.11 Coeficientes de Regresión para 7 cursos

ERROR EN LA PREDICCION DE LA NOTA DE 7 CURSOS		
CURSO	EN EL MODELO	EN EL MODELO CORREGIDO
MA133	0.134	0.134
PI140	0.185	0.186
PI216	0.154	0.154
PI515	0.147	0.147
PI225	0.135	0.136
PI510	0.121	0.121
PA136	0.124	0.125
PROMEDIO	0.143	0.143

Tabla 5.12 Comparación del modelo con el corregido

De la misma tabla 5.9 también observamos que el promedio de los errores en los 7 cursos para los pronósticos es menor que para el modelo. Esto nos asegura que, efectivamente los datos de los períodos académicos pasados, utilizados para estimación de los coeficientes de la regresión, pueden ser utilizados para pronosticar bien en los períodos académicos más recientes.

Una observación adicional de la misma tabla, es que Los cursos de PI510 y PA136, que pertenecen al noveno y décimo ciclo respectivamente, tienen los menores valores de error, mientras el curso PI140 del sexto ciclo tiene el mayor valor de error.

La tabla 5.10 nos proporciona el estadístico "t" y su nivel de significancia, cuando este nivel de significancia es menor que 0.05, podemos inferir que la variable es significativa. En esta misma tabla, los valores de nivel de significancia de las variables que son significativas, han sido resaltados en amarillo. Si bien es cierto algunas variables no son significativas, el alto valor del estadístico F y los valores bajísimos de sus niveles de significancia, nos indican que estas variables no significativas de todas formas contribuye, aunque en menor grado a la bondad del ajuste de la regresión. Esto último mencionado se verifica de la tabla 5.12, donde el error de la regresión considerando solo las variables significativas aumenta muy ligeramente para cada curso, comparándolo con la regresión que considera todas las variables.

Además podemos sacar algunas observaciones adicionales de la tabla 5.10:

- Las 7 variables son siempre significativas para uno o más cursos de los siete en estudio.
- La variable "promedio ponderado acumulado" y la "antigüedad del alumno" son significativas para los 7 cursos.
- Para los cursos MA133 y PI216 todas las variables son significativas.
- El curso de PI515 tiene solo como variables significativas al "promedio ponderado acumulado" y a la "antigüedad del alumno".

En la tabla 5.11 se dan los valores de los coeficientes de la ecuación de la regresión para cada uno de los 7 cursos, estando resaltado aquellos que corresponden a las variables significativas. En el Anexo 3, se da para el curso PI140 un ejemplo de cálculo de estimación de la Nota, considerando la regresión usando todas las variables, como usando solo las variables significativas.

5.5 Análisis comparativo de las técnicas de predicción

Del análisis realizado en los puntos anteriores podemos obtener la tabla 5.13

Esta tabla nos indica que con las variables tomadas en cuenta, tanto para la predicción de aprobación o no de un curso, como de la nota que se obtendrá, la aplicación de la técnica de Redes Neuronales y de Regresión Logística para el primer caso; y de Redes Neuronales y de Regresión Múltiple para el segundo, nos arrojan resultados similares en cuanto a su porcentaje de aciertos o nivel de error.

PARA PREDECIR APROBACIÓN O NO DE UN CURSO	TÉCNICA DE PREDICCIÓN	PROMEDIO DEL PORCENTAJE DE ACIERTOS PARA 7 CURSOS	
		MODELO	PRONOSTICO
	RED NEURONAL	70.45%	72.83%
	REGRESIÓN LOGÍSTICA	70.39%	74.04%
PARA PREDECIR NOTA DE UN CURSO	TÉCNICA DE PREDICCIÓN	PROMEDIO DE LOS ERRORES PARA 7 CURSOS	
		MODELO	PRONOSTICO
	RED NEURONAL	0.1495	0.1397
	REGRESIÓN MÚLTIPLE	0.1430	0.1380

Tabla 5.13. Comparación de las técnicas de predicción

No existe una arquitectura ideal de red neuronal para todas las aplicaciones. La arquitectura más apropiada se va obteniendo a través de sucesivos entrenamientos de ensayo y error. Para nuestro estudio la tabla 4.3 muestra que el modelo más adecuado de red neuronal, por su menor error, es la configuración [7, 8, 4, 1], pero sin embargo se tomó la configuración [7, 6, 3, 1], debido a que su configuración es más simple y tiene un diferencia de menos del 0.5 % en el error con respecto a la primera configuración mencionada.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. El modelo de red neuronal de retropropagación aplicado a la universidad peruana, permite al alumno predecir si aprobará o no un curso en que se desea inscribir en un nuevo ciclo (5.1.2).

2. El modelo de regresión logística aplicado a la universidad peruana, permite al alumno predecir si aprobará o no un curso en que se desea inscribir en un nuevo ciclo (5.2.2).

3. El modelo de red neuronal de retropropagación aplicado a la universidad peruana, permite al alumno predecir la nota que obtendrá en un curso en que se desea inscribir en un nuevo ciclo (5.3.2).

4. El modelo de regresión múltiple aplicado a la universidad peruana, permite al alumno predecir la nota que obtendrá en un curso en que se desea inscribir en un nuevo ciclo (5.4.2).

5. Para los cuatro modelos se asegura que con los datos de los períodos académicos pasados, utilizados como entrenamiento y validación para los modelos de redes neuronales, y de estimación de coeficientes para los modelos de regresiones, se pueden realizar los pronósticos para los períodos académicos más recientes (5.1.2, 5.2.2, 5.3.2 y 5.4.2).

6. Del modelo de regresión múltiple y regresión logística, se desprende que las 7 variables (para cursos con 2 pre-requisitos) ó 6 variables (para cursos con 1 pre-requisito) son siempre significativas para uno o más cursos de los siete en estudio. Además las variables "promedio ponderado acumulado" y "antigüedad en años del alumno" son siempre significativos para los 7 cursos estudiados.
7. La aplicación de las técnicas de redes neuronales de retropropagación y de regresión logística para la predicción de la aprobación o no de un curso, arrojan promedios de porcentajes de aciertos similares, de 70.45 % y 70.39 % para los modelos, y de 72.83 % y 74.04 % para los pronósticos, respectivamente.

8. La aplicación de las técnicas de redes neuronales de retropropagación y de regresión múltiple para la predicción de la "nota" de un curso, arrojan promedios de raíz de errores medios cuadráticos similares, de 0.1495 y 0.1430 para los modelos, y de 0.1397 y 0.1380 para los pronósticos, respectivamente.

9. No se requiere de una herramienta sofisticada para la aplicación del modelo de redes neuronales de retropropagación. En este trabajo se ha utilizado el Excel de Microsoft con su complemento Solver para la implementación de la red neuronal con diferentes número de capas y neuronas por capas. La ventaja es que se trabaja con una herramienta de uso general.

RECOMENDACIONES

1. El uso del Excel del Microsoft y su complemento Solver en la aplicación de la técnica de minería de datos de red neuronal de retropropagación a diferentes casos donde se requiera realizar predicciones.

2. Continuar con el estudio de investigación, añadiendo otras variables predictoras relacionadas con el nivel social y económico del alumno, como por ejemplo, tipo de colegio donde realizó sus estudios secundarios, nivel económico de la familia, etc.

3. Para la aplicación de las técnicas de minería de datos para predicción en una institución u organización, se recomienda conocer bien su funcionamiento y características propias, para realizar una adecuada selección de las variables más adecuadas.

GLOSARIO DE TÉRMINOS

Árbol de decisión: Modelo de datos en forma de árbol que generan algunos métodos de minería de datos. Los árboles de decisión se pueden utilizar para la predicción.

Archivo de base de datos: Uno de los archivos físicos que forman una base de datos.

Coeficiente de correlación R^2: Es una medida de la bondad de ajuste de una regresión lineal.

Currículo flexible: Es un instrumento orientado hacia un cambio total desde el punto de vista didáctico en la planificación y usos de los objetivos, métodos, medios y formas de evaluación.

Curso pre-requisito: Es todo curso que debe llevarse obligatoriamente antes que otro, estando este orden de precedencia escrito en el plan de estudios.

Discretizar: Colocar valores de un conjunto continuo de datos en grupos para que haya un número discreto de posibles estados.

Estadístico F de una regresión: Es una prueba estadística que sirve para rechazar o aceptar una hipótesis nula.
Estadísticos t de la regresión: Es una prueba estadística que permite determinar si las variables independientes de una regresión son o no significativas.

Minería de datos: Proceso de identificar comercialmente patrones o relaciones útiles en bases de datos u otros repositorios del equipo mediante el uso de herramientas estadísticas avanzadas.

Modelo de minería de datos: Objeto que contiene la definición de un proceso de minería de datos y los resultados de la actividad de entrenamiento. Por ejemplo, un modelo de minería de datos puede especificar la entrada, salida, algoritmo y otras propiedades del proceso y albergar la información recopilada durante la actividad de aprendizaje, como un árbol de decisión.

Período académico: es el periodo de tiempo (en meses) que comprende un ciclo académico universitario.

Promedio ponderado acumulado: Es el cociente entre la sumatoria del producto de la nota de cada curso llevado por su creditaje, y el número total de créditos.

Regresión: Proceso estadístico de predecir una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos.

Valor p: Es el nivel de significancia más bajo, al cual puede rechazarse la hipótesis nula.

REFERENCIAS BIBLIOGRÁFICAS

1. Baker, R., Yacef K. : "The State of Educational Data Mining in 2009: A Review and Future Visions" EDM2009. Córdoba 2009.
2. Bonsón Ponte, E. "Tecnologías inteligentes para la Gestión Empresarial". Edit. Alfaomega/Ra-ma. 1996.
3. Caro Encalada, M.: "El uso de tecnologías de información y comunicación en el sector hotelero de la Península de Yucatán; Hacia un modelo explicativo" Tesis Doctoral, Universidad Politécnica de Madrid, Madrid 2008.
4. Freeman, J., Skapura, D., "Redes neuronales. Algoritmos, aplicaciones y técnicas de programación". Edit Addison-Wesley, Diaz de Santos. 1993.
5. Gallardo Quingatuña, C.: "Estabilidad y amortiguamiento de oscilaciones en sistemas eléctricos con alta penetración eólica" Tesis Doctoral, Universidad Carlos III de Madrid. Leganés/Getafe, 2009
6. García Esteban, L.,García Fernández,F. "MOE prediction in Abies pinsapo Boiss. timber: Application of an artificial neural network using non-destructive testing" An International Journal: Computers and Structures,Elsevier. Madrid 2009.
7. Gujarati, D., Porter D. : "Econometría" McGraw Hill. México 2010
8. Gutiérrez Pulido, H., De la Vara Salazar, R. : "Análisis y diseño de experimentos", McGraw Hill. México 2008.
9. Hernández Orallo, J., Ramírez Quintana, M., Ferri Ramirez, C., "Introducción a la Minería de Datos", Edit. Pearson – Prentice Hall. 2005.
10. Hilera Gonzáles, J., Martinez Hernando V., "Redes Neuronales Artificiales. Fundamentos, modelos y aplicaciones". Edit. Addison-Wesley Iberoamericana, de la edición ra-ma. 1995.
11. Kumar, V., Chadha, A. : "An empirical study of the application of Data Mining techniques in higher education" International Journal of Advanced Computer Science and Applications. Vol.2 No. 3, 2011

12. [Lye, Ch., Ng, L., Hassan, M., Goh, W., Law, Ch., Ismail, N.: "Predicting Pre-university Students` mathematics achievement" International Conference on Mathematics Education Research. Elsevier, Malaysia.
13. Martín del Brío, B., Sanz Molina, A. "Redes Neuronales y Sistemas Difusos". Edit. Alfaomega/Ra-Ma. 2001
14. Pérez Lopez, C., Santín González, D., Minería de Datos. Técnicas y Herramientas. Edit. Thomson. 2007.
15. Rocha, M., Cortez, P., Neves, J. : "Evolution of networks for classification and regression" University of Minho. Elsevier, Portugal 2007
16. Romero, C., Ventura, S. : "Educational data mining: A survey from 1995 to 2005", Departamento de Ciencia de la computación. Universidad de Córdoba, Córdoba 2006
17. Romero, C., Ventura, S., Espejo, P. G., Hervás, C. "Data Mining Algorithms to Classify Students". I International Conference on Educational Data Mining, 2008.
18. Russell, S., Norvig, P., "Inteligencia Artificial. Un Enfoque Moderno". Edit. Prentice Hall Hispanoamericana, S.A. 1996.
19. Skapura, D., "Building neural networks". Edit Addison-Wesley. 1995.
20. Valluru R., Hayagriva R., "Neural Networks & Fuzzy Logic". Edit MIS:Press. 1995.
21. Vialardi, C., Bravo, J., Shafti, J., Ortigosa, A., "Recomendation in Higher Education Using Data Mining Techniques". II International Conference on Educational Data Mining, 2009.
22. Vialardi Sacin, César: "Propuesta de una metodología de aplicación de técnicas de descubrimiento del conocimiento para la ayuda al estudiante en entornos de enseñanza superior" Tesis Doctoral, Universidad Autónoma de Madrid, Madrid 2010.

ANEXOS

ANEXO 1

CURRÍCULO DE LA ESPECIALIDAD DE INGENIERÍA QUÍMICA DE LA FIQT

UNI-FIQT CURRICULO DE LA ESPECIALIDAD DE INGENIERIA QUIMICA								
Ciclo	**Código**	**Nombre del Curso**	**HT**	**HP**	**HL**	**CR**	**SE**	**Pre-Requisito**
1	AU511	Dibujo Técnico	1	3	-	2	D	Ninguno
	FI203	Física I	4	3	-	5	G	Ninguno
	MA113	Matemáticas I	3	2	-	4	G	Ninguno
	MA114	Matemáticas Básicas I	2	2	-	3	G	Ninguno
	PI100	Formación Profesional del Ing. Químico y Textil	1	1	-	1	G	Ninguno
	PI118	Sistemas de Información y Reportes Técnicos	2	-	-	2	D	Ninguno
	QU116	Química I	3	-	-	3	M	Ninguno
	QU117	Laboratorio de Química I	-	-	3	1	D	Ninguno
2	EM711	Introducción al Diseño Mecánico	2	3	-	3	I	AU511
	FI204	Física II	4	3	-	5	G	FI203
	MA123	Matemáticas II	3	2	-	4	G	MA113
	MA124	Matemáticas Básicas II	2	2	-	3	G	MA114
	MA713	Programación Digital	2	3	-	3	F	MA113 MA114
	QU118	Química II	3	-	-	3	M	QU116
	QU119	Laboratorio de Química II	-	-	3	1	D	QU116 QU117
3	EP307	Economía de la Empresa I	4	-	-	4	B	MA124
	FI403	Física III	4	3	-	5	G	FI204
	MA133	Matemáticas III	5	2	-	6	G	MA123
	QU214	Química Inorgánica	4	-	-	4	B	QU118 QU119
	QU215	Laboratorio de Química Inorgánica	-	-	3	1	D	QU118 QU119
4	EE102	Circuitos e Instalaciones Eléctricas Industriales	2	2	-	3	F	FI403
	FI152	Introducción a la Física Moderna	4	1	-	4	G	FI403
	MA143	Matemáticas IV	3	2	-	4	F	MA133 MA713
	MA612	Estadística y Diseño de Experimentos	4	1	-	4	F	MA133
	QU425	Físico Química I	4	-	-	4	B	MA133 QU118
	QU426	Laboratorio de Físico Química I	-	-	3	1	D	MA133 QU215
5	PI111	Balance de Materia y Energía	2	2	-	3	D	QU425
	PI523	Cálculos en Ingeniería Química I	4	1	-	4	G	MA143
	QU324	Química Orgánica I	4	1	-	4	F	QU118 QU426
	QU325	Laboratorio de Química Orgánica I	-	-	3	1	D	QU118 QU426
	QU434	Físico Química II	4	-	-	4	B	QU425 QU426
	QU435	Laboratorio de Físico Química II	-	-	3	1	D	QU425 QU426
	QU516	Análisis Químico Cualitativo	3	-	-	3	B	QU214 QU426
	QU517	Laboratorio de Análisis Químico Cualitativo	-	-	3	1	D	QU214 QU426

CURRÍCULO DE LA ESPECIALIDAD DE INGENIERÍA QUÍMICA DE LA FIQT (CONTINUACIÓN)

	EC618	Mecánica y Resistencia de los Materiales	3	4	-	5	F	FI403
	PA714	Investigación de Operaciones I	2	2	-	3	F	MA612
	PI140	Fenómenos de Transporte	3	1	-	3	F	MA143 PI111
6	PI216	Termodinámica Para Ingeniería Química I	3	1	-	3	G	PI111 QU434
	QU334	Química Orgánica II	4	1	-	4	F	QU324 QU325
	QU335	Laboratorio de Orgánica II	-	-	3	1	D	QU324 QU325
	QU526	Análisis Químico Cuantitativo	2	-		2	B	QU516 QU517
	QU527	Laboratorio de Análisis Químico Cuantitativo	-	-	3	1	D	QU516 QU517
	PA113	Ingeniería de Métodos I	3	2	-	4	F	MA612
	PI142	Transferencia de Cantidad de Movimiento	3	1	-	3	G	PI140
7	PI217	Termodinámica Para Ingeniería Química II	3	1	-	3	G	PI216 PI523
	PI318	Industria de los Procesos Químicos	4	3	-	5	F	PI140 QU334
	PI513	Materiales Industriales	2	1	-	2	G	QU526
	EP818	Costos y Presupuestos	2	2	-	3	F	EP307
	PI143	Transferencia de Calor	3	1	-	3	G	PI142
8	PI144	Transferencia de Masa	3	1	-	3	G	PI142
	PI146	Operaciones en Ingeniería Química I	2	1	2	3	F	PI142
	PI515	Corrosión I	2	-	3	3	F	PI513
	PI135	Laboratorio de Operaciones Unitarias I	1	-	3	2	I	PI143 PI146
	PI225	Cinética Química y Diseño de Reactores I	3	1	-	3	G	PI217
9	PI415	Instrumentos de Control	2	3	-	3	F	EE102 PI144
	PI510	Economía de Procesos	3	1	-	3	G	EP818 PI144
	PI612	Seminarios en Ingeniería Química	1	2	-	2	D	PI143 PI144
	PI911	Gestión Tecnológica y Empresarial	3	2	-	4	G	EP307
	AHD65	Constitución y Derechos Humanos	2	-	-	2	B	Ninguno
	PA136	Planeamiento y Control de la Producción	3	-	2	4	F	PA113 PA714
10	PI136	Laboratorio de Operaciones Unitarias II	1	-	3	2	I	PI135 PI144
	PI426	Simulación y Control de Procesos	3	2	-	4	F	PI225 PI415
	PI525	Diseño de Plantas	3	2	-	4	A	PI510 PI415

CURRÍCULO DE LA ESPECIALIDAD DE INGENIERÍA QUÍMICA DE LA FIQT (CONTINUACIÓN)

Ciclo	Código	Nombre del Curso	HT	HP	HL	CR	SE	Pre-Requisito
	Electivos de la Especialidad							
	Área de Tecnología de Materiales							
	PI322	Electroquímica Industrial	2	3	-	3	D	QU434
	PI355	Tratamiento de Agua Industrial I	3	1	-	3	F	PI318 QU526
	PI365	Polímeros I	2	2	-	3	D	PI144 PI318
	PI366	Polímeros II	2	2	-	3	D	PI365
	PI516	Corrosión II	2	3	-	3	F	PI515
C	PI826	Tratamiento de Efluentes Industriales	2	2	-	3	F	PI146
U	***Área de Tecnología de Procesos***							
R	PI147	Transferencia de Masa II	2	3	-	3	I	PI144
S	PI226	Cinética Química y Diseño de Reactores II	3	1	-	3	G	PI225
O	PI345	Aceites y Grasas	2	-	-	2	D	PI318
S	PI376	Diseño, Selección y Mantenimiento de Equipos	2	2	-	3	D	PI135 PI144
	PI475	Procesos de Refinación de Petróleo y Gas	3	2	-	4	G	PI144 PI143
	PI531	Introducción a la Investigación en Procesos Químicos	2	3	-	3	D	PI144 PI318
E	PI613	Seminario de Tesis en Ingeniería Química	1	2	-	2	D	PI612
L								
E	***Área de Biomasa, Derivados y Energía***							
C	PI381	Conservación de la Energía	3	1	-	3	B	PI143 PI510
T	PI721	Bioquímica y Microbiología	2	3	-	3	F	QU334 QU526
I	PI722	Procesos Bioquímicos	2	3	-	3	F	PI721
V	PI823	Combustión y Combustibles en Industrias de Procesos	2	2	-	3	F	PI225
O	PI912	Gestión Ambiental Empresarial	3	1		3	D	PI318
S	PI824	Gas Natural y Condensados	3	2		4	G	PI217 PI142
	Electivos Complementarios							
	Área Tecnológica							
	ME425	Cerámica y Refractarios	2	3	-	3	G	QU434
	HC443	Lubricantes y Aceites Minerales	3	2	-	4	G	PI318
	QU565	Análisis Químico Instrumental I	4	3	-	5	H	QU526
	SA633	Higiene Industrial	2	2	-	3	G	PA113
	Área de Administración							
	PA425	Diseño y Evaluación de Proyectos	3	2	-	4	G	PI510
	PA515	Mercadotecnia	2	1	-	2	D	EP307

CURRÍCULO DE LA ESPECIALIDAD DE INGENIERÍA QUÍMICA DE LA FIQT (CONTINUACIÓN)

Requisitos para la obtención de la Constancia de Egresado

Número Total de Créditos Obligatorios	191
Número mínimo de Créditos Electivos	20

De los Créditos Electivos

Número mínimo de Créditos de Especialidad	10

HT : Horas Teoría
HP: Horas Práctica
HL: Horas Laboratorio
CR: Crédito
SE: Sistema de Evaluación

ANEXO 2

Programa fuente hecho en el lenguaje de programación JAVA para la generación de las Variables Predictivas

```
//-------------------------------------------------------------------------------------------------------
---
//Programa fuente para la generación de las variables predictivas
//Creadores:
//            Pedro Raúl Acosta De La Cruz
//            Pedro Arturo Pizarro Solís
//Fecha: 30/04/2011
//-------------------------------------------------------------------------------------------------------
---
//Inicio del Programa
//-------------------------------------------------------------------------------------------------------
---
package ejemplo06;
import java.sql.*;
class Consulta {
  private Connection objConnection = null;
  private int cnt = 0;
  private int intcntPGD = 0;
  Consulta() {
  }
  public void OpenConnection() {
    try {
      String strDriver = "sun.jdbc.odbc.JdbcOdbcDriver";
      String strConnection = "jdbc:odbc:BASE";
      Class.forName(strDriver);
      objConnection = DriverManager.getConnection(strConnection);
    } catch (ClassNotFoundException objClassNotFoundException) {
      System.out.println("Error del driver: " + objClassNotFoundException.getMessage());
    } catch (SQLException objSQLException) {
      System.out.println("Error de la ruta: " + objSQLException.getMessage());
    }
  }
  public void CloseConnection() {
    try {
      objConnection.close();
    } catch (SQLException objSQLException) {
      System.out.println("Error: " + objSQLException.getMessage());
```

```
    }
}
private boolean TablaExiste(String strTabla) {
    boolean Estado = false;
    try {
        DatabaseMetaData objDatabaseMetaData = objConnection.getMetaData();
        ResultSet objResultSetTables = objDatabaseMetaData.getTables(null, null, "%", null);
        String strNombreTabla = "";
        while (objResultSetTables.next()) {
            strNombreTabla = objResultSetTables.getString(3);
            if (strNombreTabla.equals(strTabla) == true) {
                Estado = true;
            }
        }
        objResultSetTables.close();
    } catch (Exception objException) {
        System.out.println("Error: " + objException.getMessage());
        Estado = false;
    } finally {
        return Estado;
    }
}
private void TablaCrea(String strTabla, int intTipo) {
    try {
        Statement objStatementCreateTable = null;
        String strSQL = "";
        switch (intTipo) {
            case 1:
                strSQL = "Create Table " + strTabla
                        + "("
                        + "Registro AUTOINCREMENT PRIMARY KEY NOT NULL,"
                        + "Codigo Text(9),"
                        + "Periodo Text(6),"
                        + "SumCre Double,"
                        + "SumNot Double,"
                        + "AcuCre Double,"
                        + "AcuNot Double,"
                        + "ProPonAcu Double"
                        + ")";
                break;
            case 2:
                strSQL = "Create Table " + strTabla
                        + "("
                        + "Registro AUTOINCREMENT PRIMARY KEY NOT NULL,"
                        + "PERACD Text(6),"
```

```
                + "CODCUR Text(9),"
                + "N_APROB Integer,"
                + "N_DESAP Integer"
                + ")";
            break;
        case 3:
            strSQL = "Create Table " + strTabla
                + "("
                + "Registro AUTOINCREMENT PRIMARY KEY NOT NULL,"
                + "CODCUR Text(9),"
                + "PERACD Text(6),"
                + "PN Double,"
                + "GD Double,"
                + "PGD Double"
                + ")";
            break;
        case 4:
            strSQL = "Create Table " + strTabla
                + "("
                + "Registro AUTOINCREMENT PRIMARY KEY NOT NULL,"
                + "perAca Text(6),"
                + "cod Text(9),"
                + "cur Text(9),"
                + "notCur Double,"
                + "ppa Double,"
                + "curPreReq1 Text(9),"
                + "notPreReq1 Double,"
                + "curPreReq2 Text(9),"
                + "notPreReq2 Double,"
                + "pGD Double,"
                + "spgd Double,"
                + "sumCre Double,"
                + "antAlu Integer"
                + ")";
            break;
      }
      objStatementCreateTable = objConnection.createStatement();
      objStatementCreateTable.execute(strSQL);
      objStatementCreateTable.close();
    } catch (Exception objException) {
      System.out.println("Error: " + objException.getMessage());
    }
  }
  private void TablaElimina(String strTabla) {
    try {
```

```
      String strSQL = "DROP TABLE " + strTabla;
      Statement objStatementTableElimina = null;
      objStatementTableElimina = objConnection.createStatement();
      objStatementTableElimina.execute(strSQL);
      objStatementTableElimina.close();
    } catch (Exception objException) {
      System.out.println("Error: " + objException.getMessage());
    }
  }
  public void CreaTabla(String strTabla) {
    try {
      if (TablaExiste(strTabla) == true) {
        TablaElimina(strTabla);
      }
      TablaCrea(strTabla, 3);
      String strSQLSelect = "SELECT "
          + "CODCUR, "
          + "PERACD, "
          + "Avg(NOTA) AS PN, "
          + "Avg(20-NOTA) AS GD "
          + "FROM HIST "
          + "GROUP BY CODCUR, PERACD "
          + "ORDER BY CODCUR, PERACD, Avg(20-NOTA) DESC ";
      String strSQLInsert;
      String strCodCursoActual = "";
      String strCodCursoSiguiente = "";
      String strPerAcademico;
      double dblPN;
      double dblGD;
      double dblPGD = 0;
      Statement objStatementSelect = objConnection.createStatement();
      ResultSet objResultSet = objStatementSelect.executeQuery(strSQLSelect);
      Statement objStatementInsert = objConnection.createStatement();
      int i = 0;
      System.out.println("Inicio de creación de la tabla principal");
      while (objResultSet.next()) {
        strCodCursoActual = objResultSet.getString("CODCUR");
        strPerAcademico = objResultSet.getString("PERACD");
        dblPN = objResultSet.getDouble("PN");
        dblGD = objResultSet.getDouble("GD");
        if (strCodCursoActual.equals(strCodCursoSiguiente) == false) {
          strCodCursoSiguiente = strCodCursoActual;
          dblPGD = 0;
          i = 0;
        }
```

```
            dblPGD += dblGD;
            strSQLInsert = "INSERT INTO " + strTabla
                + "("
                + "CODCUR,"
                + "PERACD,"
                + "PN,"
                + "GD,"
                + "PGD"
                + ")"
                + " VALUES "
                + "('"
                + strCodCursoActual + "','"
                + strPerAcademico + "',"
                + dblPN + ","
                + dblGD + ","
                + dblPGD / ++i
                + ")";

            System.out.print(".");
            objStatementInsert.execute(strSQLInsert);
        }
        System.out.println("Fin de creación de la tabla principal");
        objResultSet.close();
        objStatementSelect.close();
        objStatementInsert.close();
    } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
    }
}
public void CreaTablaFinal(String strTabla) {
    if (TablaExiste(strTabla) == true) {
        TablaElimina(strTabla);
    }
    TablaCrea(strTabla, 4);
}
private double promGD(String strTabla, String strPerAcd, String strCodCur) {
    try {
        Statement objStatementSelect = null;
        ResultSet objResultSet = null;
        String strSQLSelect = "SELECT TOP 1 Avg(GD) AS promGD "
            + "FROM " + strTabla + " "
            + "WHERE CODCUR='" + strCodCur + "' AND PERACD < '" + strPerAcd + "' "
            + "GROUP BY CODCUR ";
        objStatementSelect = objConnection.createStatement();
        objResultSet = objStatementSelect.executeQuery(strSQLSelect);
```

```
      double promGD = 0;
      while (objResultSet.next()) {
        promGD = objResultSet.getDouble("promGD");
      }
      objResultSet.close();
      objStatementSelect.close();
      return (promGD);
    } catch (SQLException objSQLException) {
      System.out.println("Error: " + objSQLException.getMessage());
      return (-1);
    }
  }
  private double promPonAcu(String strCodigo, String strPerAca) {
    try {
      Statement objStatementSelect = null;
      ResultSet objResultSet = null;
      String strSQLSelect = "SELECT TOP 1 SUM(CRD*NOTA)/SUM(CRD) AS TA FROM HIST "
          + "WHERE CODALU='" + strCodigo + "' AND PERACD < '" + strPerAca + "' "
          + "GROUP BY CODALU,PERACD "
          + "ORDER BY CODALU,PERACD DESC ";
      objStatementSelect = objConnection.createStatement();
      objResultSet = objStatementSelect.executeQuery(strSQLSelect);
      double promPonAcu = 0;
      while (objResultSet.next()) {
        promPonAcu = objResultSet.getDouble("TA");
      }
      objResultSet.close();
      objStatementSelect.close();
      return (promPonAcu);
    } catch (SQLException objSQLException) {
      System.out.println("Error: " + objSQLException.getMessage());
      return (-1);
    }
  }
  private String CurPreReq1(String strCodigo, String strPerAca, String strCurso) {
    try {
      Statement objStatementSelect = null;
      ResultSet objResultSet = null;
      String strSQLSelect = "SELECT HIST_1.CODCUR "
          + "FROM (HIST LEFT JOIN CURSO AS PR1 ON HIST.CODCUR = PR1.CODCUR) "
          + "LEFT JOIN HIST AS HIST_1 ON PR1.PREREQ1 = HIST_1.CODCUR "
          + "WHERE (((HIST.CODALU)='" + strCodigo + "') "
          + "AND ((HIST.PERACD)='" + strPerAca + "') "
          + "AND ((HIST.CODCUR)='" + strCurso + "') "
          + "AND ((HIST_1.CODALU)=[HIST].[CODALU]) "
```

```
            + "AND ((HIST_1.NOTA)>=10)) ";
        objStatementSelect = objConnection.createStatement();
        objResultSet = objStatementSelect.executeQuery(strSQLSelect);
        String curso = "";
        while (objResultSet.next()) {
          curso = objResultSet.getString("CODCUR");
        }
        objResultSet.close();
        objStatementSelect.close();
        return (curso);
      } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
        return ("Error");
      }
    }
    private double NotPreReq1(String strCodigo, String strPerAca, String strCurso) {
      try {
        Statement objStatementSelect = null;
        ResultSet objResultSet = null;
        String strSQLSelect = "SELECT HIST_1.NOTA "
            + "FROM (HIST LEFT JOIN CURSO AS PR1 ON HIST.CODCUR = PR1.CODCUR) "
            + "LEFT JOIN HIST AS HIST_1 ON PR1.PREREQ1 = HIST_1.CODCUR "
            + "WHERE (((HIST.CODALU)='" + strCodigo + "') "
            + "AND ((HIST.PERACD)='" + strPerAca + "') "
            + "AND ((HIST.CODCUR)='" + strCurso + "') "
            + "AND ((HIST_1.CODALU)=[HIST].[CODALU]) "
            + "AND ((HIST_1.NOTA)>=10)) ";
        objStatementSelect = objConnection.createStatement();
        objResultSet = objStatementSelect.executeQuery(strSQLSelect);
        double nota = 0;
        while (objResultSet.next()) {
          nota = objResultSet.getDouble("NOTA");
        }
        objResultSet.close();
        objStatementSelect.close();
        return (nota);
      } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
        return (-1);
      }
    }
    private String CurPreReq2(String strCodigo, String strPerAca, String strCurso) {
      try {
        Statement objStatementSelect = null;
        ResultSet objResultSet = null;
```

```
        String strSQLSelect = "SELECT HIST_1.CODCUR "
            + "FROM (HIST LEFT JOIN CURSO AS PR1 ON HIST.CODCUR = PR1.CODCUR) "
            + "LEFT JOIN HIST AS HIST_1 ON PR1.PREREQ2 = HIST_1.CODCUR "
            + "WHERE (((HIST.CODALU)='" + strCodigo + "') "
            + "AND ((HIST.PERACD)='" + strPerAca + "') "
            + "AND ((HIST.CODCUR)='" + strCurso + "') "
            + "AND ((HIST_1.CODALU)=[HIST].[CODALU]) "
            + "AND ((HIST_1.NOTA)>=10)) ";
        objStatementSelect = objConnection.createStatement();
        objResultSet = objStatementSelect.executeQuery(strSQLSelect);
        String curso = "";
        while (objResultSet.next()) {
          curso = objResultSet.getString("CODCUR");
        }
        objResultSet.close();
        objStatementSelect.close();
        return (curso);
      } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
        return ("Error");
      }
    }
    private double NotPreReq2(String strCodigo, String strPerAca, String strCurso) {
      try {
        Statement objStatementSelect = null;
        ResultSet objResultSet = null;
        String strSQLSelect = "SELECT HIST_1.NOTA "
            + "FROM (HIST LEFT JOIN CURSO AS PR1 ON HIST.CODCUR = PR1.CODCUR) "
            + "LEFT JOIN HIST AS HIST_1 ON PR1.PREREQ2 = HIST_1.CODCUR "
            + "WHERE (((HIST.CODALU)='" + strCodigo + "') "
            + "AND ((HIST.PERACD)='" + strPerAca + "') "
            + "AND ((HIST.CODCUR)='" + strCurso + "') "
            + "AND ((HIST_1.CODALU)=[HIST].[CODALU]) "
            + "AND ((HIST_1.NOTA)>=10)) ";
        objStatementSelect = objConnection.createStatement();
        objResultSet = objStatementSelect.executeQuery(strSQLSelect);
        double nota = 0;
        while (objResultSet.next()) {
          nota = objResultSet.getDouble("NOTA");
        }
        objResultSet.close();
        objStatementSelect.close();
        return (nota);
      } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
```

```
      return (-1);
    }
  }
  public void MuestraAlumnoPeriodo(String strTabla1, String strTabla2, String strCurso) {
    try {
      String strSQLSelect = "SELECT "
          + "CODALU,PERACD,CODCUR "
          + "FROM HIST "
          + "WHERE CODCUR = '" + strCurso + "'"
          + "ORDER BY CODALU,PERACD ";
      String strCODALU;
      String strPERACD;
      String strCODCUR;
      Statement objStatementSelect = objConnection.createStatement();
      ResultSet objResultSet = objStatementSelect.executeQuery(strSQLSelect);
      double spgd = 0;
      while (objResultSet.next()) {
        strCODALU = objResultSet.getString("CODALU");
        strPERACD = objResultSet.getString("PERACD");
        strCODCUR = objResultSet.getString("CODCUR");
        spgd = SPGD(strTabla1, strCODALU, strPERACD);
        Mostrar(strTabla1, strTabla2, strCODALU, strPERACD, strCODCUR, spgd);
      }
      objResultSet.close();
      objStatementSelect.close();
    } catch (SQLException objSQLException) {
      System.out.println("Error: " + objSQLException.getMessage());
    }
  }
  public double SPGD(String strTabla1, String strCodigo, String strPerAca) {
    try {
      String strSQLSelect = "SELECT "
          + "CODCUR "
          + "FROM HIST "
          + "WHERE CODALU = '" + strCodigo + "' AND PERACD = '" + strPerAca + "' "
          + "GROUP BY CODALU, PERACD, CODCUR ";
      String strCODCUR;
      Statement objStatementSelect = objConnection.createStatement();
      ResultSet objResultSet = objStatementSelect.executeQuery(strSQLSelect);
      //String strTabla = "GD";
      double pGD = 0;
      while (objResultSet.next()) {
        strCODCUR = objResultSet.getString("CODCUR");
        //CreaTablaPGD(strTabla, strPerAca, strCODCUR);
        pGD += promGD(strTabla1, strPerAca, strCODCUR);
```

```
        }
        objResultSet.close();
        objStatementSelect.close();
        return (pGD);
      } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
        return (-1);
      }
    }
    public double CreditosTotales(String cod, String perAca) {
      try {
        String strSQLSelect = "SELECT SUM(CRD) AS SumCrd "
            + "FROM HIST "
            + "WHERE CODALU='" + cod + "' AND PERACD='" + perAca + "' "
            + "GROUP BY CODALU,PERACD ";
        Statement objStatementSelect = objConnection.createStatement();
        ResultSet objResultSet = objStatementSelect.executeQuery(strSQLSelect);
        double sumCrd = 0;
        while (objResultSet.next()) {
          sumCrd = objResultSet.getDouble("SumCrd");
        }
        objResultSet.close();
        objStatementSelect.close();
        return (sumCrd);
      } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
        return (-1);
      }
    }
    public double NotaCurso(String cod, String perAca, String cur) {
      try {
        String strSQLSelect = "SELECT NOTA "
            + "FROM HIST "
            + "WHERE CODALU='" + cod
            + "' AND PERACD='" + perAca
            + "' AND CODCUR='" + cur + "' ";
        double notCur = 0;
        Statement objStatementSelect = objConnection.createStatement();
        ResultSet objResultSet = objStatementSelect.executeQuery(strSQLSelect);
        while (objResultSet.next()) {
          notCur = objResultSet.getDouble("NOTA");
        }
        objResultSet.close();
        objStatementSelect.close();
        return (notCur);
```

```
    } catch (SQLException objSQLException) {
      System.out.println("Error: " + objSQLException.getMessage());
      return (-1);
    }
  }
  public int AntiguedadAlumno(String cod, String perAca) {
    String si = String.format("%s", cod.substring(0, 4));
    String sf = String.format("%s", perAca.substring(0, 4));
    int ymin = Integer.parseInt(String.format("%s", cod.substring(0, 4)));
    int ymax = Integer.parseInt(String.format("%s", perAca.substring(0, 4)));
    return (ymax - ymin);
  }
  public void Mostrar(String strTabla1, String strTabla2, String cod, String perAca, String
cur, double spgd) {
    double notCur = NotaCurso(cod, perAca, cur);
    double ppa = promPonAcu(cod, perAca);
    String curPreReq1 = CurPreReq1(cod, perAca, cur);
    double notPreReq1 = NotPreReq1(cod, perAca, cur);
    String curPreReq2 = CurPreReq2(cod, perAca, cur);
    double notPreReq2 = NotPreReq2(cod, perAca, cur);
    int antAlu = AntiguedadAlumno(cod, perAca);
    double sumCre = CreditosTotales(cod, perAca);
    double pGD = promGD(strTabla1, perAca, cur);
System.out.printf("%10s%12s%10s%10.2f%10.2f%10s%10.2f%10s%10.2f%10.2f%10.2f%10.
2f%15d\n", perAca, cod, cur, notCur, ppa, curPreReq1, notPreReq1, curPreReq2,
notPreReq2, pGD, spgd, sumCre, antAlu);
    try {
      Statement objStatementInsert = objConnection.createStatement();
      String strSQLInsert = "INSERT INTO " + strTabla2 + " "
          + "(perAca, cod, cur, notCur, ppa, curPreReq1, notPreReq1, curPreReq2,
notPreReq2, pGD, spgd, sumCre, antAlu) "
          + "VALUES "
          + "('"
          + perAca + "', '"
          + cod + "', '"
          + cur + "', "
          + notCur + ", "
          + ppa + ", '"
          + curPreReq1 + "', "
          + notPreReq1 + ", '"
          + curPreReq2 + "','"
          + notPreReq2 + "',"
          + pGD + ", "
          + spgd + ", "
          + sumCre + ", "
```

```
            + antAlu
            + ")";
        objStatementInsert.execute(strSQLInsert);
        objStatementInsert.close();
      } catch (SQLException objSQLException) {
        System.out.println("Error: " + objSQLException.getMessage());
      }
    }
}
public class EjecutaConsulta {
  public static void main(String[] args) {
    Consulta objConsulta = new Consulta();
    objConsulta.OpenConnection();
                String curso = "PI510";
                String strTabla1 = "GD"+curso;
                 String strTabla2 = "PGD"+curso;
                 objConsulta.CreaTabla(strTabla1);
                objConsulta.CreaTablaFinal(strTabla2);


System.out.printf("%10s%12s%10s%10s%10s%10s%10s%10s%10s%10s%10s%10s%15s\n",
"PerAca", "CÃ³digo", "Curso", "Nota", "PPA", "CurPR1", "NotPR1", "CurPR1", "NotPR2",
"promGD", "SPGD", "Creditos", "Antiguedad");
    objConsulta.MuestraAlumnoPeriodo(strTabla1,strTabla2,curso);
    objConsulta.CloseConnection();
  }
}

//-----------------------------------------------------------------------------------------------------
---
//Fin del Programa
//-----------------------------------------------------------------------------------------------------
---
```

ANEXO 3

ECUACIONES PARA ESTIMAR LA NOTA PARA EL CURSO DE PI140

X1ppa = (ppa - 5.10) / (17.68-5.10)

X2notPreReq1 = (notPreReq1 – 10) / (18.4 – 10)

X3notPreReq2 = (notPreReq2 – 10) / (18.0 – 10)

X4Pgd = (Pgd – 9.10) / (11.02 - 9.10)

X5spgd = (spgd – 18.75) / (97.90 – 18.75)

X6sumCre = (sumCre – 4) / (27 – 4)

X7antAlu = (antAlu – 2) / (10 – 2)

Con todas las variables

notCur calc = 0.449139323 + 0.195663259 (X1ppa) + 0.034185474 (X2notPreReq1) + 0.076538845 (X3not PreReq2)
+ 0.009465898 (X4Pgd) + 0.302462409 (X5spgd) - 0.233703302 (X6sumCre) - 0.09344558 (X7antAlu)

Con las variables significativas

notCur calc = 0.460584043 + 0.207279948 (X1ppa) + 0.079434093 (X3notPreReq2) + 0.314063151 (X5spgd)
- 0.244755601 (X6sumCre) - 0.094567306 (X7antAlu)

NOTA = 0.1 + (15.7 – 0.1) (notCur calc)

ECUACIONES PARA ESTIMAR LA PROBABILIDAD DE APROBAR EL CURSO PI140

X1ppa = (ppa - 5.10) / (17.68-5.10)

X2notPreReq1 = (notPreReq1 – 10) / (18.4 – 10)

X3notPreReq2 = (notPreReq2 – 10) / (18.0 – 10)

X4Pgd = (Pgd – 9.10) / (11.02 - 9.10)

X5spgd = (spgd – 18.75) / (97.90 – 18.75)

X6sumCre = (sumCre – 4) / (27 – 4)

X7antAlu = (antAlu – 2) / (10 – 2)

Con todas las variables

Z = - 0.643 + 2.209 (X1ppa) + 0.512 (X2notPreReq1) + 0.951 (X3notPreReq2) + 0.045 (X4pGD) + 1,694 (X5spgd) - 2,406 (X6sumCre) – 0.349 (X7antAlu)

Con las variables significativas

Z = - 0.800 + 2.525 (X1ppa) +1.005 (X3notPreReq2) + 1.858 (X5spgd) – 2.404 (X6sumCre)

$$\Pr = \frac{1}{1+e^{-z}}$$

Donde Pr = Probabilidad de aprobar el curso PI140 de acuerdo a:

Pr >= 0.5 aprueba el curso ;

Pr < 0.5 desaprueba el curso