Universidad Autónoma del Estado de Hidalgo

Instituto de Ciencias Básicas e Ingeniería

Área Académica de Computación y Electrónica

Minería de Datos Educativa: predicción y caracterización del promedio académico, índice de inscripción y conclusión de estudios, caso de estudio UTXJ

Tesis

Que para obtener el grado de

Maestro en Ciencias Computacionales

Presenta:

José Abdiel Domínguez León.

Director de Tesis:

Félix A. Castro Espinoza.

Pachuca de Soto, Hidalgo enero 2015

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BASICAS E INGENIERIA

Area Academica de Computación y Electronica

Maestria en Ciencias Computacionales

Oficio No: MCC-04-2015

Lic. José Abdiel Dominguez Leon
P R E S E N T E

Por este conducto le comunico que el jurado asignado para la revisión de su trabajo de tesis titulado *"Mineria de Datos Educativa: predicción y caracterización del promedio académico, índice de inscripción y conclusión de estudios, caso de estudio U.A.E."* que para obtener el grado de Maestro en Ciencias Computacionales fue presentado por usted, ha tenido a bien, en reunión de sinodales, autorizarlo para su impresión despues de verificar las correcciones que fueron acordadas.

A continuación se integran las firmas de conformidad de los integrantes del jurado:

PRESIDENTE: DR. OMAR LOPEZ ORTEGA

VOCAL: M. en C. FELIX AGUSTIN CASTRO ESPINOZA

SECRETARIO: M. en C. ARTURO CURIEL ANAYA

SUPLENTE: M. en C. MARIANO JAVIER POZAS CARDENAS

ATENTAMENTE
"AMOR, ORDEN Y PROGRESO"
Mineral de la Reforma, Hgo., a 16 de enero de 2015

M.C.C. Iliana Castillo Pérez
Coordinadora de la Maestría en Ciencias Computacionales

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
SISTEMA BIBLIOTECARIO

c.c.p. Archivo

"Soy de los que piensan que la ciencia tiene una gran belleza. Un científico en su laboratorio no es sólo un técnico: es también un niño colocado ante fenómenos naturales que le impresionan como un cuento de hadas. "

Marie Curie

Agradecimientos

La presente tesis es resultado de esfuero y dedicación, que sin la participación de algunas personas no hubiese sido posible llevarla a su término. Por lo que con orgullo utilizo este espacio para agradecerlo.

A Dios, por bridarme salud, fortaleza y sabiduria necesaria para cumplir esta meta que parecia inalcanzable.

A mi familia les agradezco por su apoyo incondicional y porque creyeron en mi en todo momento dificil ...

- A mis hijos Pepe y Samy ya que son la fuente de motivación para superarme siempre desde antes de su nacimiento.
- A Altagracia, quien con su cariño y comprensión tuve valor para seguir adelante.
- A mis padres y mis hermanas, que desde siempre me han dado sus bendiciones.

Gracias a la Universidad Autónoma del Estado de Hidalgo (UAEH) y a cada sus docentes que participaron en la impartición del Programa de Posgrado en Ciencias Computacionales en la UTXJ y que con consejos motivó a la conclusión del presente proyecto.

Gracias a la UTXJ, a su Rector, el Ingeniero Jesús Santos Picazo y a la Secretaría Académica, la Contadora María Elena Hernández Briones, quienes dieron las facilidades para el uso de la información fuente de este proyecto.

Un agradecimiento muy especial al Mtro. Félix A. Castro Espinoza, quien no dudó en darme la confianza, los medios para terminar este proyecto, las asesorías, el tiempo, su positivismo, exigencia y su apoyo incondicional.

A mis estimados compañeros de grupo de maestría: Randolfo, Matilde, Raúl, Emma, con quienes crecí profesionalmente en la UTXJ y compartí vivencias inolvidables como compañero de maestría.

Para José Pepe, para Samy Aby y

Para Altagracia Catana,
con cariño y amor . . .

Resumen

La preocupación por el desempeño de los alumnos de primer año de carreras universitarias (ingresantes), que surge de numerosos y desfavorables indicadores de deserción y bajo rendimiento académico, ha llevado a las universidades del país a investigar las causas que subyacen en esta problemática [25, Porcel et al, 2010].

En este trabajo se analiza la relación de la admisión de aspirantes a la Universidad Tecnológica de Xicotepec de Juárez (UTXJ) contra tres metas importantes que predecir y caracterizar: el promedio de egreso, la conclusión de estudios y el índice de inscripción del Técnico Superior Universiario (TSU).

Se consideraron a 5547 aspirantes de diferentes generaciones desde 2009 a 2014 con un corte a octubre de 2014, todos provenientes de 5 áreas académicas existentes: Tecnologías de la Información y Comunicación, Salud, Económico Administratriva, Electromecánica Industrial y Agroalimentaria Biotecnológica.

La metodología de Razonamiento Inductivo Difuso (FIR) y el algoritmo LR - FIR se utilizaron en el presente proyecto, los cuales, fueron utilizados para predecir y caracterizar el promedio académico, el índice de inscripción y la conlusión de estudios tanto de estudiantes como de aspirantes en su caso, obteniendose así reglas lingüisticas que permiten entender dichos resultados.

Entre las variables más reelevantes se consideraron el Año de Admisión , la Selección de la Carrera, el Turno Solicitado, la Escuela de Educación Media Superior proveniente, la Especialidad del Bachillerato de procedencia, el Mes de Registro, el Promedio de Egreso del Bachiller, el Medio de Difusión por el que fué enterado el alumno de la oferta educativa, la Edad al ingreso, el Género.

Así mismo, se tomó en cuenta el promedio alcanzado en el examen de admisión que consta de tres áreas importantes: habilidad matemática, verbal e Inglés y por último la información de calificaciones finales de asignaturas, los cuales dieron significancia a la predictibilidad junto con la obtención de reglas LR - FIR.

Es importante que se haga referencia al tema de la minería de datos, que aunque no es un tópico nuevo, ha venido utilizándose considerablemente en los últimos años en la Educación, Redes Sociales y Educativas, Ambientes Virtuales de Aprendizaje, Gestión Educativa y en general al descubrimiento de conocimiento en bases de datos.

En la fase experimental se ejecutaron los tratamientos a las fuentes de información para acreditar su confiabilidad durante el uso de los datos, los cuales se utilizaron durante los experimentos, los cuales se acopiaron para decidir con que experimento servirían, se estandarizaron cuando su origen no fué el mismo, se estructuraron para su uso en las plataforma FIR yse dividieron para su entrenamiento y prueba.

Con respecto a los experimentos ejecutados, éstos fueron analizados tomando varios factores, desde la cantidad de variables a considerar, la calidad de la máscara para cada modelo, las magnitudes del error de predicción, la gráfica que muestra

los datos reales y los predichos por FIR, así como las reglas obtenidas con LR - FIR.

Algunas variables que se consideran en las reglas obtenidas durante los experimentos como en la predicción de la inscripción, la especialidad del Bachillerato de procedencia es reelevante, además que hay coincidencia con los programas educativos actualmente ofertados por la UTXJ, así mismo, los resultados del examen de admisión y el tiempo de ejecución de cada apartado.

Las reglas para la predicción de la conclusión de estudios dejan ver que la selección que tiene el alumno con respecto al área académica, así como que la UTXJ sea primer opción de estudios, son factores que influyen en el témino de sus estudios.

En la predicción del promedio académico, influye la lejania de procedencia del alumno respecto a la ubicación de la UTXJ, el promedio alcanzado en el bachillerato y el mes en que decidió solicitar su ingreso a la universidad.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Uno de los problemas a enfrentar en la universidad es la deserción en los primeros años de las carreras. La deserción universitaria afecta tanto en los ámbitos personales como en los institucionales, sociales y económicos [13, Fischer, 2012].

- En lo personal, implica una condición de fracaso que afecta emocionalmente por la discrepancia con las aspiraciones personales e incide en la trayectoria ocupacional de los individuos.
- En lo institucional, implica una disminución del rendimiento académico de la universidad.
- En lo social, la deserción contribuye a generar inequidad y desequilibrios sociales y desvirtúa los objetivos que la sociedad le ha entregado a la educación superior.
- En lo económico, el costo que esto implica para los sistemas educativos es considerable.

En los últimos años una preocupación ha surgido en muchos países ante el problema del fracaso escolar y un creciente interés por determinar los múltiples factores que pueden influir en él. La mayoría de los trabajos que intentan resolver este problema utilizan una gran cantidad de información, todos estos datos constituyen una auténtica mina de oro de valiosa información sobre los estudiantes. El problema es identificar y encontrar información útil y oculta. Una solución muy prometedora para alcanzar este objetivo es el uso de técnicas de extracción de conocimiento o minería de datos educativa (Educational Data Mining, EDM)[25, Porcel et al, 2010].

En una mirada contextualizada, el rendimiento académico es el producto de condiciones institucionales (diseño curricular, práctica docente, valores y concepciones institucionales, etc.), socioeconómicas (situación laboral, estado civil, nivel educativo del grupo familiar, entre otras) e individuales (formación previa, hábitos de estudios, etc.) de los estudiantes [26, Porcel, 2010(1)].

Las técnicas de la Mierí a de Datos Educativa (EDM, por sus siglas en Inglés) ya se han empleado con éxito para crear modelos de predicción del rendimiento de los estudiantes, obteniendo resultados prometedores que demuestran como determinadas características sociológicas, económicas y educativas de los alumnos pueden afectar en el rendimiento académico [25, Porcel et al, 2010].

En otro sentido si bien todos los jóvenes tienen derecho a asistir a la escuela, independientemente del grado escolar, existe una gran proporción de ellos que no lo hacen. Del total de los jóvenes que están en edad normativa para cursar el nivel medio superior, sólo 7 de cada 10 estudia y no todos ingresan a la educación superior [24, Ogarrido, 2012].

1.1. Planteamiento del Problema

La institución de enseñanza superior puede ser visualizada, para fines de análisis y planeación, como un sistema social en equilibrio dinámico y es importante analizar la consecuencia del crecimiento o decremento de la matrícula de primer ingreso a la institución, y sus efectos derivados sobre los demás componentes del sistema.

El aumento en el número de estudiantes usuarios del sistema, representa un cambio en las instalaciones y equipo, que suelen volverse insuficientes, por lo que se requieren más salones de clase, laboratorios, facilidades higiénicas, libros, salas de biblioteca, muebles, instalaciones deportivas, estacionamientos, lugares de esparcimiento, medios de comunicación, etc.

El pronóstico de la magnitud probable de la inscripción de primer ingreso a una institución de enseñanza superior, depende de numerosos factores. Algunos de ellos son identificables y susceptibles de cuantificación; otros son de naturaleza cualitativa y por ello escapan a las posibilidades de medición [19, Kleiman, 1975].

En la UTXJ se ha visto un cambio notorio de demanda en las áreas académicas y en especial aquellas con poca inscripción tienen la necesidad de asegurar a todo aspirante desde el momento en que solicita su ingreso.

En las figuras 1.1 y 1.2 se observa la evolución de matrícula de la UTXJ desde el 2002 al 2013 y la clasificación de matricula de TSU por área académica, la cual es importante porque indica las áreas con mayor demanda, también que a partir de 2008 existe un incremento anual de matrícula, sin embargo, este incremento se acentúa más en algunas carreras.

Por otro lado, la deserción, el rezago estudiantil y los bajos índices de eficiencia terminal se encuentran entre los problemas más complejos y frecuentes que enfrentan las Instituciones de Educación Superior del país (ANUIES, 2001).

La mayoría de las instituciones han hecho algún tipo de esfuerzos por disminuir estos índices realizando y estableciendo programas de tutorías, asesorías, congresos, talleres, eventos para que los alumnos se involucren directamente y aumente su compromiso y una serie de actividades.

HISTÓRICO DE MATRICULA TOTAL POR CICLO ESCOLAR			
CICLO ESCOLAR	MATRÍCULA A INICIO DE CICLO ESCOLAR	INCREMENTO DE MATRÍCULA RESPECTO AL CICLO ESCOLAR ANTERIOR	
2002-2003	304	0	0.0%
2003-2004	786	482	158.6%
2004-2005	725	-61	-7.8%
2005-2006	605	-120	-16.6%
2006-2007	602	-3	-0.5%
2007-2008	613	11	1.8%
2008-2009	684	71	11.6%
2009-2010	1193	509	74.4%
2010-2011	1346	153	12.8%
2011-2012	1361	15	1.1%
2012-2013	1954	593	43.6%
2013-2014	2473	519	26.6%

FIGURA 1.1: Histórico de matrícula total por ciclo escolar.

HISTÓRICO DE MATRÍCULA DE T.S.U.															
CICLO-ESCOLAR		AARH	QAB	MECAA	MAI	PA	TICASI	TICAMC	TFAR	MIAP	FOT	GAS	TOTAL	DISTRIBUCIÓN POR GÉNERO	
														Hombres	Mujeres
2002-2003		91			57	63	93						304	53.29%	46.71%
2003-2004		227			149	153	257						786	52.04%	47.96%
2004-2005		146			119	196	264						725	51.72%	48.28%
2005-2006		178		27	88	83	229						605	50.58%	49.42%
2006-2007		169		54	86	86	207						602	53.82%	46.18%
2007-2008		157		72	100	107	177						613	53.34%	46.66%
2008-2009		178	21	87	98	122	178						684	48.39%	51.61%
2009-2010		182	51	125	271	105	178						912	51.54%	48.46%
2010-2011		193	67	134	360	105	162						1021	53.28%	46.72%
2011-2012		246	52	123	386	117	169						1093	56.18%	43.82%
2012-2013		260	51	132	500	99	170	46	180				1438	52.85%	47.15%
2013-2014	Sep - Dic 2013	262	49	117	598	85	126	50	370	92	27	75	1851	51.86%	48.14%
	Ene - Abr 2014	235	42	99	497	77	112	38	317	87	19	51	1574	51.46%	48.54%
	FINES DE FEBRERO 2014	231	40	100	486	76	110	41	309	72	19	51	1535	51.14%	48.86%

FIGURA 1.2: Histórico de matrícula total de TSU por ciclo escolar.

Sin embargo, muchos de estos esfuerzos no han sido suficientes y el fenómeno se sigue repitiendo constantemente por ciclo escolar, así que el estudio de los factores e índices que afectan a la deserción ha cobrado mayor importancia en los últimos años. La necesidad de identificar y predecir la deserción de los estudiantes en los primeros cuatrimestres es indispensable para tomar las acciones pertinentes y poder disminuirla.[30, Valero, 2009] en las figuras 1.3 y 1.4 se muestra un concentrado de la evolución de indicadores sobre la deserción y el bajo aprovechamiento académico.

Por último, con la revolución industrial, el concepto de rendimiento adquirió un nuevo valor que se incorporaba para que fuera posible valorar al hombre: tanto "rindes", tanto vales. Desde esa época de la producción, la sociedad no se ha desprendido del concepto sino que lo ha extendido y refinado a todos los aspectos y

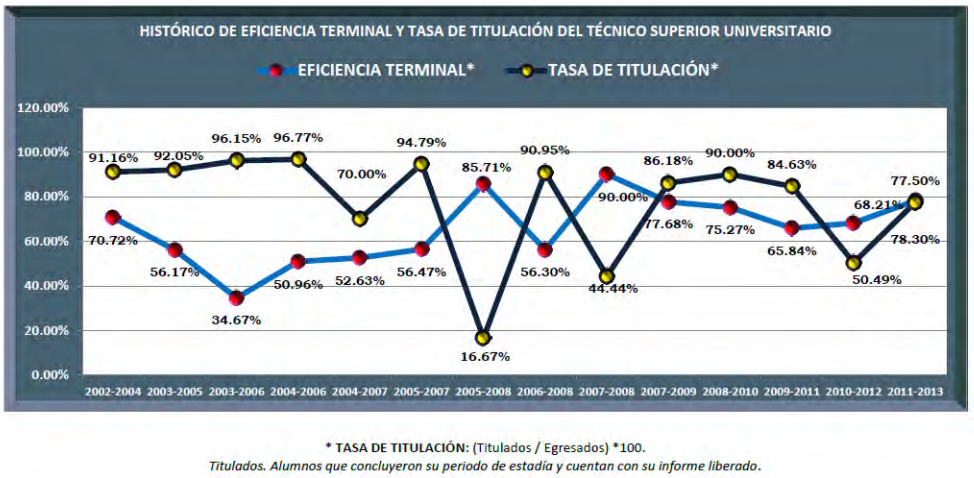


FIGURA 1.3: Histórico de eficiencia terminal y tasa de titulación.

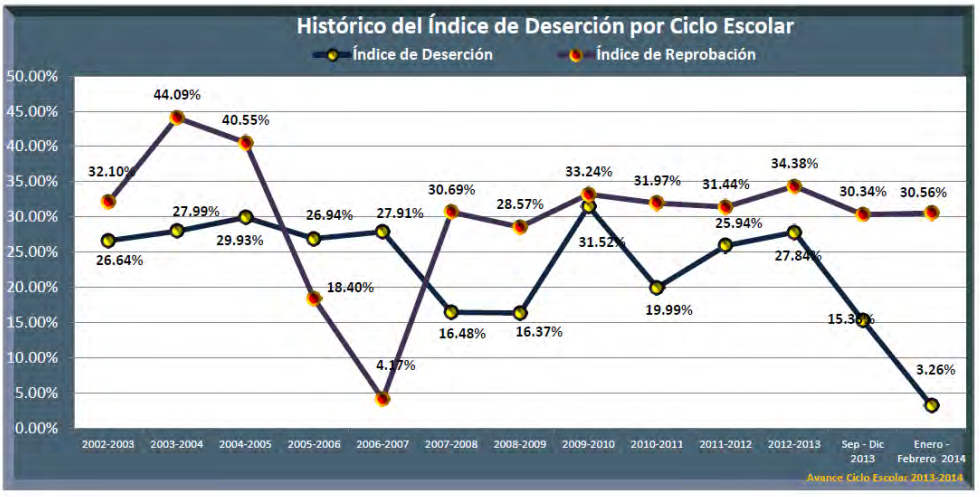


FIGURA 1.4: Histórico de índice de deserción por reprobación por ciclo escolar.

sectores de la vida productiva del hombre. Por lo tanto, la educación ha asumido este concepto sumándose así a la sociedad de producción, particularmente, el rendimiento académico se ha visto como un producto.

Al considerar a la educación como un factor más de producción, es lógico preocuparse por la calidad de ese producto, siendo los resultados educativos del propio sistema, éxitos o fracasos escolares, los indicadores de esa calidad.

Muchas de las encuestas realizadas en Europa explican el éxito escolar a partir de las características individuales, familiares y escolares de los estudiantes. En Norteamérica se han centrado en el establecimiento escolar, las características del medio universitario y la composición social que allí se establece, tratando de especificar así la tipología de los centros y la influencia que la institución como tal tiene en los resultados académicos. Estos estudios, de tipo descriptivo, aportan una valiosa información que permite establecer comparativamente algunas regularidades [7, Celis et al, 2010].

Considerando que las actividades cotidianas de índole académico y administrativo en la UTXJ generan grandes volumenes de información, se ha percatado que no son analizados con perspectivas que integren conocimiento predictivo sustentado en herramientas y metodológias, entre los datos acopiados en la comunidad estudiantil en su paso por la universidad se tienen aquellos de caracter personal durante el proceso de registro en la admisión, el examen de admisión Institucional, que involucra habilidades matemáticas, verbales, conocimiento del idioma Inglés y examinación psicométrica, los resultados de la evaluación por el CENEVAL a través del EXANI II, las calificaciones históricas durante el cuatrimestre, arrojadas por cada asignatura, además de la aplicación de encuestas especiales sobre la trayectoria educativa de cada estudiante en su paso por la carrera.

Por lo tanto, se observa comportamientos como el ingreso de matrícula para algunas áreas académicas específicas, el aprovechamiento académico y la deserción escolar que enfrentan las Instituciones de Educación Superior, para los cuales se ha hecho una esfuerzo institucional a través de un programa Institucional de Tutorías, observandose resultados que no alcanzan las metas idóneas.

1.2. Hipótesis

Con información básica de la admisión y calificaciones históricas de los alumnos de las generaciones de 2009 a 2014, es posible predecir y caracterizar el promedio académico, el índice de de inscripción y la conclusión de estudios mediante la identificacion de un modelo FIR, utilizando variables que hacen referencia a información de admisión, algunos atributos personales y calificaciones alcanzadas al corte cuatrimestral.

La descripción de patrones de comportamiento para el grueso de los aspirantes es posible caracterizarla utilizando la extensión de la metodología FIR denominada LR - FIR, que permite extraer reglas del tipo IF - THEN que representan el comportamiento del sistema.

1.3. Objetivos

1.3.1. Objetivo General

Con el uso de una Plataforma de Minería de Datos Educativa basada en algoritmos predictivo difusos como FIR y LR - FIR, obtener evidencias de la caracterización y proponer una predicción del promedio académico, el índice de inscripción y la conclusión de estudios de los aspirante a la UTXJ

1.3.2. Objetivos Específicos

- Aplicar técnicas de Minería de Datos a información generada en el entorno educativo de la UTXJ, tomando información de los aspirantes y estudiantes de TSU de las generaciones de 2009 a 2014 como objetos de estudio que se encuentran en el modelo de enseñanza presencial.
- Realizar experimentos con datos de admisión de aspirantes a la UTXJ e información histórica de calificaciones de alumnos.
- Utilizar la metodología FIR para identificar las características que más influyen en la determinación del promedio académico, el índice de inscripción, y la conclusión de estudios tanto de los aspirantes como de los alumnos.
- Aplicar el mecanismo de predicicón de FIR para determinar el promedio académico, el índice de inscripción, y la conclusión de estudios tanto de los aspirantes como de los alumnos.
- Utilizar el modelo de reglas obtenido por LR - FIR para descubrir patrones de comportamiento de los aspirantes y alumnos.
- Comprobar la validez del modelo propuesto para el problema en cuestión y evaluar los resultados.

Capítulo 2

Estado del Arte

A principios de los años sesenta, la educación superior experimentó diversos cambios en todo el mundo. En Francia, el gobierno fué consciente de formar jóvenes como Técnico Superior Universitarios y se creó una modalidad de egreso como profesional, la cual tenía una corta duración, que debía cursarse inmediatamente después del bachillerato, lo que propició la educación del bachillerato tecnológico. En 1966 surgieron los primeros Institutes Universitaires de Technologie (IUT), con formaciones tecnológicas diseñadas alrededor de áreas del conocimiento aplicables a diversos campos profesionales, con programas de dos años y que otorgaban el Diploma Universitario de Tecnología (DUT).

Entre 1970 y 1973, la Secretaría de Educación Pública (SEP) realizó por primera vez estudios comparativos de los sistemas educativos de los paises con mayor desarrollo industrial, a través de los cuales percibió la modalidad de estudios de bachillerato más dos años de acreditación profesional, sin embargo, por falta de visión y de decisiones unilaterales que persistian en ese momento, no se consideró que en México existiera una demanda real del sector productivo respecto a la necesidad de contratar recursos humanos con ese nivel de formación tecnológica y fué hasta que el Programa de Desarrollo Educativo 1995 - 2000 previó la apertura de mayor número de opciones educativas a nivel superior.

La Universidad Tecnológica de Xicotepec de Juárez (UTXJ) fué creada en septiembre del 2002 como respuesta a la demanda de educación superior en la sierra norte de Puebla y como reacción propositiva ante la necesidad de las empresas de la región de contar con recursos humanos calificados, profesionales con sólida preparación científica, humanísitica y tecnológica.

La UTXJ ofrece educación superior a nivel Técnico Superior Universitario e Ingeniería con una formación de 3150 horas, en 6 cuatrimestres a lo largo de 2 años, para el primer modelo educativo y 1 año 8 meses adicionales para el segundo ciclo, acumulando 1980 horas más. Con poco más de 12 años de labores, la UTXJ ofrece 11 Programas Educativos de TSU, 1 de Licencia Profesional y 5 de Ingeniería.

El sistema educativo mexicano ha experimentado grandes avances en materia de acceso y cobertura, así como en indicadores de trayectoria escolar. Sin embargo,

aun persisten rezagos en estos rubros cuando se compara el acceso que tienen los jóvenes acorde a variables básicas como el género, la condición indígena, el nivel de ingresos o el tamaño de la localidad, sólo por mencionar algunas. Los problemas se incrementan en el nivel superior y se visualizan en algunas situaciones como la inscripción escolar, el rendimiento académico y la deserción escolar [24, Ogarrido, 2012].

2.1. Pronósticos de demanda de matricula de primer ingreso

El pronóstico de la magnitud probable de la inscripción de primer ingreso a una institución de enseñanza superior, depende de numerosos factores. Algunos de ellos son identificables y susceptibles de cuantificación, otros son de naturaleza cualitativa y por ello escapan a las posibilidades de medición. Ayuda el suponer que la magnitud que se busca pronosticar está en función de numerosas variables independientes, que reflejan condiciones culturales, sociales, económicas o, aún, de la naturaleza y estructura del propio sistema educativo.

Algunos métodos simples de predicción relacionan los pronósticos de inscripción con factores puramente descriptivos de los procesos reales, sin penetrar en la esencia causal de los agentes que motivan el fenómeno bajo observación. Otros métodos son tan elaborados, que requieren una comprensión muy detallada de las interrelaciones básicas de los componentes del sistema, tanto entre sí como con numerosos elementos del medio, que constituyen el marco de referencia externo a los procesos bajo análisis.

Ciertos métodos de pronóstico son muy confiables, aun cuando manejen pocas variables y la metodología sea simple. Todo depende de la complejidad del sistema, dada por sus requisitos de admisión, la diversidad de las carreras profesionales que ofrecen, el calendario que maneja, el número de instituciones de enseñanza media de la que recibe egresados, la proximidad a otras instituciones de enseñanza superior, complementarias o competitivas, la vulnerabilidad a factores económicos de naturaleza general, la adaptabilidad a nuevas corrientes de mayor atractivo para el joven estudiante, y factores semejantes [19, Kleiman, 1975].

2.1.1. Métodos de Pronóstico de la demanda

Aunque la aplicación de modelos de pronóstico de la demanda de matrícula a instituciones de enseñanza está aún en sus albores, existen diversos métodos concretos que han sido utilizados con provecho en numerosos trabajos analíticos enfocados a la consideración de esta problemática [19, Kleiman, 1975]. Los más usuales son:

- El método de proporciones;

- El método de supervivencia de cohortes;
- El método de regresión, y
- El método de flujo escolar.

Además existen otras técnicas que han probado su eficacia en tipos generales de pronósticos, tales como:

- El método de los promedios móviles;
- El método de filtrado exponencial, y;
- El método de Box y Jenkins.

2.2. Rendimiento académico

El rendimiento académico es un claro indicador del avance en la carrera de estudios de algún alumno en un momento particular, y a su vez también es un pronosticador de la posibilidad de completar dicha carrera de estudios.

Por tal motivo el rendimiento académico ha sido representado de diferentes maneras en los diversos estudios que han abordado el tema. En algunos, está representado sólo por el número de materias aprobadas por un alumno en una carrera, en otros por los resultados de tests específicamente diseñados o el promedio de notas de las asignaturas cursadas [26, Porcel, 2010(1)].

Muchos factores influyen en el rendimiento académico, unos que pertenecen o se encuentran en el mismo estudiante (endógenos), y otros que pertenecen o se encuentran en el mundo circundante (exógenos). Estos factores no actúan aisladamente, el rendimiento académico es el resultado de la acción recíproca de lo interno y lo externo [25, Porcel et al, 2010].

Diversas características del alumnado han sido consideradas a la hora de relacionarlas con el rendimiento académico, desde las características aptitudinales, intelectuales y de la personalidad del alumno hasta los aspectos motivacionales y de percepción personal de los estudiantes durante el transcurso de la carrera, así como también razones de ingreso a la misma. También se ha estudiado cómo la pertenencia a un cierto sector socioeconómico o características personales del alumno, tales como, edad, género y lugar de procedencia, pueden relacionarse y a su vez explicar el rendimiento académico.

El rendimiento académico ha sido representado de diferentes maneras, en algunos de ellos, el rendimiento académico es representado por el número de materias aprobadas por un alumno en una carrera, en otros por el resultado de tests específicamente diseñados, así como también, por el promedio de notas de las materias

cursadas. Esta variedad de manifestaciones del rendimiento académico están referidas al nivel de estudios en el cual se analiza el desempeño de los alumnos, el tiempo de la investigación o el enfoque del investigador.

El término rendimiento tiene muchas implicaciones, principalmente si se considera a las calificaciones obtenidas por los alumnos como el referente casi exclusivo. Esta información puede generar, incluso, una lectura ingenua que centra sólo la responsabilidad académica en el alumno. Sin embargo, la responsabilidad institucional es clave para evaluar lo que se entiende por rendimiento. Más allá de las condiciones internas a las instituciones y de las prácticas docentes, resulta imprescindible también conocer las características que aportan quienes son los receptores de la labor docente [25, Porcel et al, 2010].

2.3. Deserción escolar

Se entiende por deserción estudiantil al hecho de que un número de estudiantes matriculados no siga la trayectoria normal del programa académico, bien sea por retirarse de ella, por repetir cursos o por retiros temporales. La separación de la matrícula académica, ya sea por parte del estudiante o de la universidad, tiene efectos de tipo financiero, académico y social que implican la pérdida de esfuerzos y recursos en un país. La deserción estudiantil en los programas de pregrado de la gran mayoría de instituciones de educación superior (IES) tanto de Colombia como de Latinoamérica es un problema que tiene un impacto multidimensional en el desarrollo social y económico de un país [28, Timaran, 2010].

El tema de la deserción es central al momento de definir políticas universitarias que promuevan principios de equidad social y oportunidades educativas. El objetivo es alcanzar condiciones igualitarias para todos los alumnos, contando con la oportunidad de acceso a las instituciones educativas y de permanencia en ellas. La deserción puede ser vista desde el punto de la calidad, como un indicador de falta de eficacia al medir la incapacidad del sistema para conservar a los estudiantes y permitirles cursar sin retrasos ni salidas del sistema educativo.

La deserción es un problema que se puede ver reflejado en los diferentes niveles educativos. El informe sobre la educación superior en América Latina y el Caribe 2000 - 2005, según UNESCO en su informe de: La educación Superior en el siglo XXI, Visión y Acción, 2005 indica [12, Fabiana, 2006]:

La deserción en las universidades está provocando afecciones en la salud física y mental de los jóvenes, y un elevado costo, ya que se estima que al año en América Latina y el Caribe se pierden entre 2 y 415 millones de dólares por país, producto del abandono de los estudios universitarios.

Durante mucho tiempo, la deserción universitaria fue considerada como un fenómeno normal e incluso, como una muestra de la exigencia de la universidad y de la carrera particular; sin embargo, hoy se ve como un signo de ineficiencia y como un

gran costo para el país, los estudiantes y para las instituciones de educación superior, lo que pasó a convertirse en un problema que hay que entender para poder combatirlo.

Existen dos modelos principales desde los cuales se puede estudiar el problema de los determinantes de la deserción:

- El paradigma funcionalista y el paradigma dialéctico. La perspectiva funcionalista tiene un enfoque individualista de la educación, donde lo más importante son los talentos, habilidades, dones y esfuerzo individuales.
- El paradigma dialéctico enfoca la deserción dentro de todo el sistema educativo, donde no solo importa el estudiante como individuo, sino como parte de todo el sistema educativo.

En los estudios sobre deserción asociados a los distintos segmentos de enseñanza, abarcando enseñanza básica, media y universitaria, se puede apreciar una falta de consenso respecto de un concepto unificado que permita una recolección de datos con una metodología igualmente unificada.

El fenómeno comprende a quienes no siguieron la trayectoria normal de la carrera, bien sea por cancelar su matrícula o por no matricularse. Cuantitativamente el fenómeno puede expresarse como el número de estudiantes que abandonan la universidad en un período determinado, antes de haber obtenido el título correspondiente, relativo al total de estudiantes asociado a la cohorte correspondiente.

La deserción también sería consecuencia de interacciones insuficientes con otros (estudiantes, profesores y personal administrativo) en la escuela y congruencia insuficiente con los modelos de valores predominantes en la colectividad escolar.

El fenómeno se puede observar desde tres ópticas diferentes [13, Fischer, 2012].

- En primer lugar, la individual, que se refiere al hecho de que la persona llega a la universidad buscando obtener un título que lo acredite ante la sociedad como alguien que tiene la idoneidad intelectual. En consecuencia, quien no logra esta meta individual es llamado desertor.
- En segundo lugar, se encuentra la óptica institucional, que se relaciona con el choque del estudiante contra los preceptos institucionales que lo repelen, llevándolo lentamente a comprender que debe retirarse, unas veces conscientemente, otras de manera irracional y dolorosa.
- En tercer y último lugar, se encuentra la óptica estatal para la cual, la deserción está en la base de la organización educativa del país.

2.4. La minería de datos

Las primeras investigaciones sobre Minería de Datos se remontan aproximadamente a finales de la década de los 80's. Se impulsó en gran parte por el desarrollo de áreas como la inteligencia artificial, el aprendizaje automático, las bases de datos relacionales y avances en la microelectrónica e informática.

Al hablar de Minería de Datos es necesario hacer referencia a las áreas con las cuales esta tiene relación; la estadística tradicional y el análisis de datos son algunos de estas. Los métodos estadísticos y el análisis sobre los datos, no proporcionan conocimiento como tal, debido a esto, fue necesario fomentar una práctica más profunda, para utilizar los datos y extraer beneficios de estos.

La respuesta a estas necesidades y a muchas otras, como el almacenamiento de gran cantidad de datos y la necesidad de herramientas adecuadas e innovadoras que apoyen la toma de decisiones, está reflejada en una de las áreas de investigación más recientes, la Minería de Datos. A continuación, se dan algunas definiciones de Minería de Datos [27, Rodriguez, 2012]:

- La Minería de Datos es la exploración de forma automática o semiautomática de grandes cantidades de datos para el descubrimiento de reglas y patrones.
- La Minería de Datos es la búsqueda para nueva y valiosa información no trivial en grades volúmenes de datos.
- La Minería de Datos puede definirse como un proceso iterativo de detección y extracción de patrones a partir de grandes bases de datos: esto es modelo-reconocimiento.
- La Minería de Datos es el análisis de un conjunto de datos para encontrar relaciones desconocidas y resumir los datos de nuevas formas entendibles para el minero.

Algunos sinónimos con los que se referencia a la minería de datos:

- Descubrimiento de conocimiento en bases de Datos.
- Minería de conocimiento de bases de datos.
- Extracción de conocimiento.
- Análisis de datos y patrones.
- Arqueología de datos.

2.4.1. Descubrimiento de conocimiento KDD

"Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos".

El descubrimiento de patrones validos es posible gracias a la Minería de Datos (Data Mining), que entre otras sofisticadas técnicas aplica la Inteligencia Artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el Descubrimiento de Conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

El procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil para un usuario y satisfacerle sus metas, es el objetivo principal del área de Descubrimiento de Conocimiento en Bases de Datos o KDD. Este es el campo que está evolucionando para proporcionar soluciones al análisis automatizado. El conocimiento es descubierto usando técnicas de aprendizaje inteligente que van examinando los datos a través de procesos automatizados. El descubrimiento del conocimiento debe ser interesante; es decir, debe tener un valor potencial para el usuario.

La cantidad de datos que requieren procesamiento y análisis en grandes bases de datos exceden las capacidades humanas y la dificultad de transformar los datos con precisión es un conocimiento que va más allá de los límites de las bases de datos tradicionales.

La utilización de los datos almacenados depende del uso de técnicas del descubrimiento del conocimiento. "KDD se puede utilizar como un medio de recuperación de información, de la misma manera que los agentes inteligentes realizan la recuperación de información en el Web.

También se puede utilizar el KDD como una base para las interfaces inteligentes del mañana, agregando un componente de Descubrimiento de Conocimiento a un motor de Bases de Datos o integrando KDD con las hojas de cálculo. El proceso de KDD usa algoritmos de Minería de Datos para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post procesamientos".

El Descubrimiento de Conocimiento en Bases de Datos (DCBD) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos y presentar resultados.

Para la selección de atributos y la extracción de parámetros de información es necesario preparar correctamente los datos para procesarlos, elegir un método adecuado para la extracción de los patrones deseados y finalmente determinar cómo evaluar los patrones encontrados, estas etapas han sido organizadas en un

esquema conocido como el proceso de Descubrimiento de Conocimiento en Bases de Datos donde una de las principales etapas utiliza la DM.

El proceso KDD es un proceso no-trivial de identificación de patrones válidos, novedosos y potencialmente útiles sobre un conjunto de datos, esto es, el objetivo es encontrar conocimiento útil, valido, relevante y nuevo sobre una determinada actividad.

En este contexto los datos hacen referencia a un conjunto de hechos o ejemplos en una base de datos y los patrones son resultados o expresiones en algún lenguaje que puedan describir de manera compacta los datos; el termino no-trivial comprende que alguna búsqueda o inferencia es llevada a cabo, esto es, implica la búsqueda de modelos, estructuras, patrones o parámetros.

De manera que el proceso KDD está dividido por una serie de etapas en el cual se estructura por tres grandes bloques los cuales son: el pre-procesamiento, búsqueda o identificación de patrones (La utilización de técnicas y métodos de DM) y la evaluación, esto es, está dividido por una serie de pasos desde la selección y limpieza de la base de datos hasta la evaluación e interpretación de los resultados tal como se ilustra en la Figura 2.1.

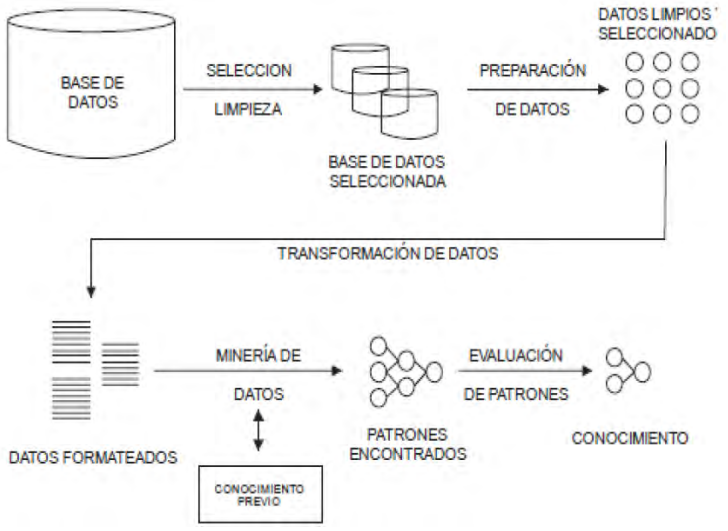


FIGURA 2.1: Proceso de descubrimiento de conocimiento con bases de datos.

.

Por lo tanto, de la figura 2.1 anterior se puede observar que el proceso KDD se estructura por una serie de pasos iniciando por la selección, preparación, limpieza y el formateo de los datos de acuerdo a los patrones analizar, a esta etapa es conocida como el pre-procesamiento, posteriormente interviene la etapa de la minería de datos en el cual tiene como tarea buscar y descubrir patrones ocultos en las bases de datos con base a la utilización de algún algoritmo de DM a implementar, pasando a la última etapa de evaluación, determinando la validez y confiabilidad del conocimiento adquirido, es decir, los patrones deben ser válidos y de alto impacto

para el usuario final. A continuación se describen las etapas del proceso KDD para enmarcar el objetivo general del funcionamiento de este tipo de sistemas.

2.4.1.1. Pre-procesamiento

Dentro de esta etapa se determina la preparación de los datos para implementar la siguiente etapa donde se implementaran las técnicas de la minería de datos, por consiguiente el pre-procesamiento se rige por tres pasos básicos, los cuales son la selección y limpieza, preparación y transformación de los datos.

Selección y limpieza de los datos, este proceso se encarga de determinar las fuentes y características de la información, por lo que, permite la navegación y visualización previa de los datos determinando que aspectos son de interés y puedan ser estudiados.

Así mismo, existen varias bases de datos que tienen diversas inconsistencias de tal manera que la limpieza y el procesamiento de datos involucra una estrategia para manejar adecuadamente el ruido que contengan algunos datos los cuales pueden ser valores faltantes, inconsistencias en los valores que no corresponden a los dominios de los atributos o que puedan ser contradictorios, esto es, valores incompletos o erróneos en la fuente de información por mencionar algunos. De tal forma que este tipo de problemas deben eliminarse antes que la etapa de DM, de tal forma que puedan afectar la precisión de los resultados o incluso el algoritmo, puede construir modelos ineficientes a partir de un conjunto de datos incorrectos.

Preparación de los datos, este proceso busca eliminar los datos que no serán relevantes para el procesamiento de la etapa de la minería de datos, no todas las bases de datos necesitaran la aplicación de todas las etapas del pre-procesamiento, por ejemplo una base de datos que tiene los registros de los estudiantes en el cual almacena el estatus de un curso. Si todos los atributos son significativos después de eliminar las inconsistencias el proceso omitirá dicha etapa pasando a la transformación de los datos, la tarea consiste en identificar características específicas de los estudiantes en la Transformación de Datos;

Transformación de los datos, en este proceso cada algoritmo que se implementará siempre establece el tipo y estructura de los datos con los que procesará, es decir, cada algoritmo de minería de datos a utilizar requieren un formato y la estructura para sus entradas, de tal forma que la tarea que está resolviendo los datos no tiene la entrada establecida por el algoritmo por lo que se procederá a transformarlos.

2.4.1.2. Minería de Datos (DM)

Esta etapa es considerada como la parte central del proceso KDD, con la finalidad de encontrar o descubrir los patrones de interés para el usuario final, estos patrones pueden ser grafos, reglas de asociación, clasificaciones, una red neuronal,

clustering, entre otros. La tarea que realiza la DM son: Seleccionar y aplicar el método de DM apropiado, es decir, la realización de una selección de la tarea para el descubrimiento del conocimiento, tales métodos son como la clasificación, agrupamiento (Clustering), reglas de asociación, regresión, por mencionar algunas.

Este sección lleva a cabo el proceso de la DM en busca de patrones para expresarlos en modelos o simplemente la expresión de dependencias de datos, este modelo depende de su función (clasificación) y de los métodos de representación por la elección de algún algoritmo como los Arboles de Decisión, Reglas de Asociación, el teorema Naive – Bayes, algunos métodos de Inteligencia Artificial como las Redes Neuronales, etc., así mismo se tiene que especificar un criterio de preferencia para la selección de un modelo dentro de un conjunto posible de modelos, también es necesario especificar la estrategia de búsqueda a utilizar que normalmente se encuentra dentro de algún algoritmo de DM.

2.4.1.3. Evaluación

Esta etapa se considera como la evaluación, interpretación, transformación y representación de los patrones extraídos del proceso KDD donde se establecen parámetros que permitan comparar la calidad de los resultados obtenidos y su validación por medio de modelos representativos de manera gráfica como curvas de aprendizaje, tasas de error, perfiles de rendimiento por mencionar algunos.

Es decir, repetir el proceso si los resultados obtenidos no fueron satisfactorios para un modelo cualitativo o se requiera implementar un nuevo algoritmo o quizá generar un análisis de nuevos datos y la implementación de nuevas estrategias que sean de utilidad para el usuario final.

Esta etapa final del proceso KDD implica que el conocimiento obtenido realice las acciones requeridas para el buen desempeño del sistema o para almacenarlo y reportarlo a los usuarios interesados.

Este proceso es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones [29, Timaran(1), 2010].

Algunos elementos básicos que se encuentran en el proceso son:

- Bases de datos
  - Hojas de cálculo, Datawarehouse u otra clase de repositorio.
  - Técnicas de limpieza e integración.
- Servidor de bases de datos
  - Utilizado para obtener la información relevante según el proceso de minería de datos.

- Base de conocimiento de minería de datos
  - Conocimiento del dominio para guiar la búsqueda, evaluar que tan interesantes son los patrones
  - Creencias de los datos (del usuario: lo que se espera de los datos para descubrir comportamientos inesperados)
  - Umbrales de evaluación
  - Conocimiento previo
  - Meta-datos

- Algoritmo de minería de datos
  - Modular para realizar distintos tipos de análisis
  - Caracterización
  - Asociación
  - Clasificación
  - Análisis de grupos
  - Evolución (en espacio o tiempo)
  - Análisis de desviaciones

- Módulo de Evaluación de Patrones de minería de Datos
  - Medidas de que tan interesante es un patrón
  - Interactúa con el algoritmo de M.D. para guiar la búsqueda hacia patrones interesantes

- Interfaz gráfica de Minería de Datos
  - Interacción con el usuario
  - Elección de la tarea de minería de datos
  - Proveer informacón para enfocar la búsqueda
  - Ayudar a evaluar los patrones
  - Explorar los patrones encontrados y la base de datos original
  - Visualizar los patrones en distintas formas

2.4.2. Modelos de extracción del conocimiento

En la práctica, los modelos para extraer patrones pueden ser de dos tipos: supervisados o predictivos y descriptivos o de descubrimiento del conocimiento.

Los modelos **predictivos** pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de la base de datos, llamadas variables independientes o predictivas. Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos).

A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases:

- Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y
- Prueba (prueba del modelo sobre el resto de los datos).

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, hay que recurrir a los métodos **descriptivos** o de descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos).

El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas. Los modelos descriptivos identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos.

Por lo tanto, en la minería de datos se ha dado lugar a una paulatina sustitución del análisis de datos dirigido a la verificación por un enfoque de análisis de datos dirigido al descubrimiento del conocimiento. La principal diferencia entre ambos se encuentra en que en el último se descubre información sin necesidad de formular previamente una hipótesis.

2.4.3. Aplicaciones de Minería de Datos

Las empresas y organizaciones actuales buscan la extracción fácil y oportuna de conocimiento útil para la toma de decisiones, a partir de los grandes volúmenes de datos que se van recopilando a través de los años, a través del registro de transacciones de las bases de datos construidas y alimentadas regularmente, y de las encuestas realizadas.

Con la tecnología actual, resulta más o menos sencillo coleccionar grandes volúmenes de información. Con el uso de lectura óptica y código de barras, las cadenas de supermercados pueden fácilmente coleccionar la información de cada canasta de compra, es decir, cual es el conjunto de artículos que el cliente compra.

Un concepto similar es el estado de cuenta mensual de una tarjeta de crédito en el que se describe un conjunto de artículos que el cliente adquirió ese mes. De igual manera, gobiernos, instituciones públicas y privadas, están en la posibilidad de juntar millones y millones de datos de actividades individuales que contienen información altamente detallada sobre montos, fechas, horas, lugares, productos y servicios [14, Giraldo y Jimenez, 2013].

- Astronomía
  - Clasificación de estrellas y galaxias
- Análisis de Mercado y Administración
  - Perfil de clientes.
    - Qué tipos de clientes compran que productos? (Clasificación o Agrupamiento (clustering))
    - ¿Qué productor se compran normalmente juntos? (Reglas de asociación)
  - Descubrir las relaciones entre características personales y el tipo de productos que se compran.
  - Descubrir correlaciones entre compras.
- Finanzas
  - Compañías de inversión hacen transacciones en la bolsa de valores basándose en resultados de Minería de Datos
  - Predicción de flujo de efectivo
- Detección de fraude
  - Utilizan bases de datos históricas para crear modelos de comportamiento fraudoliento y utilizar Minería de Datos para identificar nuevos fraudes.
    - Seguros de autos
    - Seguros médicos
    - Lavado de dinero
    - Telefónicos
    - Tratamiento médico inapropiado

- Deportes
  - Para interpretar las estadísticas
- Web
  - Analizar logs en general
  - Analizar el comportamiento de los usuarios de un sitio
- E-mail
  - Clasificar e-mail y repartirlo al departamento adecuado
- Personalización
  - Hacer recomendaciones de acuerdo a características conocidas del usuario
- Recursos humanos
  - Ayudar a seleccionar empleados
- Bancos
  - Analizar clientes para otorgar crédito
- Medicina
  - Aplicaciones que buscan nuevos medicamentos
  - Análisis de secuencias de genes
  - Predecir si un compuesto causa cáncer
  - Análisis de secuencias de proteínas

2.4.4. Minería de Datos Educativa

Se ha incrementado el interés en utilizar la minería de datos en el estudio educacional, centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educativas y en el uso de esos métodos para comprender mejor a los estudiantes y el entorno en el que aprenden.

La minería de datos educativa ofrece numerosas ventajas con los paradigmas más tradicionales de investigación relativa a la educación, como experimentos de laboratorio, estudios sociológicos o investigación de diseño. La minería de datos

aplicada al ambiente educativo posee el potencial de extender un conjunto de herramientas mucho más amplio para el análisis de cuestiones importantes sobre diferencias individuales.

Los métodos empleados en la minería de datos en la educación suelen diferir de los métodos más generalistas, explotando explícitamente los múltiples niveles de jerarquía presentes en los datos. Métodos psicométricos suelen ser integrados con métodos de aprendizaje, máquina y textos de minería de datos para lograr los objetivos.

Por ejemplo, obteniendo datos sobre cómo los estudiantes eligen utilizar el software educacional, puede ser realmente útil considerar datos a distintos niveles sobre las pulsaciones de teclas, nivel de respuestas del alumno, de la clase o de la escuela entera. Otros temas como el tiempo, secuencia o incluso el contexto juegan papeles importantes en el estudio de datos educativos [18, Jimenez y Alvarez, 2010].

Para comprender cómo y por qué la minería de datos funciona en la educación superior, es importante conocer algunos conceptos fundamentales. La minería de datos se basa en cuatro métodos esenciales [17, Jing, 2010]: Clasificación, categorización, estimación y visualización.

- La clasificación identifica las asociaciones y los conglomerados, y separa los sujetos que son objeto de estudio.
- La categorización utiliza algoritmos de inducción de reglas para tratar los resultados categóricos, como “repetir” o “abandonar” y “trasladar” o “permanecer”.
- La estimación incluye funciones predictivas o probabilidades y trata con variables continuas de resultados como, por ejemplo, la nota media o el nivel salarial.
- La visualización utiliza gráficos interactivos para demostrar de manera matemática las reglas y las puntuaciones inducidas, y es mucho más sofisticada que los gráficos de barra o de sectores. La visualización se utiliza principalmente para representar ubicaciones geográficas tridimensionales o coordinadas matemáticas.

Capítulo 3

FIR y LR - FIR

3.1. FIR - Razonamiento Inductivo Difuso

La conceptualización de la metodología Razonamiento Inductivo Difuso (FIR por sus siglas en Inglés), proviene de una especialización de la Teorá General de Sistemas (GSPS, Resolvedor de Problemas de Sistemas Generales) desarrollada por G. Klir[KLIR (1985)], y es una herramienta para el análisis de sistemas generales que permite estudiar los modos de comportamiento de los sistemas dinámicos [15, Gomez, 2009].

FIR es una metodologá de modelado y simulación cualitativa basada en el análisis del comportamiento del sistema en lugar del conocimiento de su estructura interna [NEBOT (1998)]. FIR realiza dos tareas principales:

- La primera es identificar las relaciones causales y temporales entre las variables del sistema para construir el modelo cualitativo del sistema observado;
- La segunda es predecir el comportamiento futuro del sistema a partir de las observaciones pasadas y del modelo previamente identificado.

Para cumplir con estas tareas, la metodología FIR cuenta con cuatro funciones básicas (Ver figura 3.1):

- Fusificación (Recodificación).
- Modelado cualitativo (Búsqueda de la máscara óptima).
- Simulación cualitativa (Predicción).
- Defusificación (Regeneración).

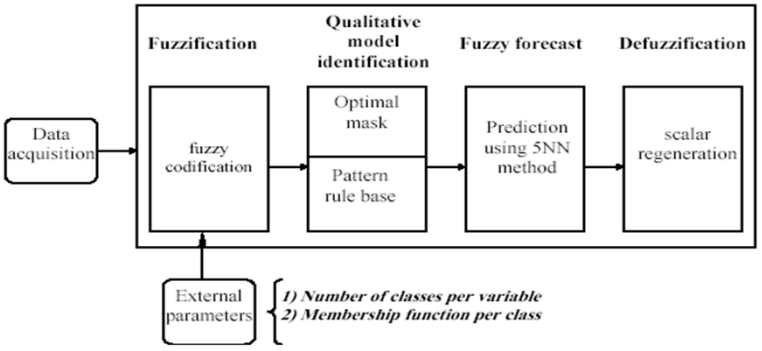


FIGURA 3.1: Etapas de la metodología FIR.

3.1.1. Fusificación

El proceso de **Fusificación** convierte las entradas cuantitativas del sistema en datos cualitativos difusos. FIR se alimenta de datos reales registrados del sistema. En esta etapa un valor cuantitativo se convierte en una tripleta cualitativa, compuesta por un primer elemento que corresponde a la clase, el segundo al valor de pertenencia difusa y el tercero corresponde al valor del lado de la función de pertenencia.

El valor del lado es la estrategia que utiliza FIR para garantizar que no se perderá información durante el proceso de fusificación, logrando saber en todo momento donde se encuentra el valor cualitativo, izquierda, centro o derecha del máximo de la función de pertenencia.

Para ejemplificar este proceso se considera que se tiene un valor cuantitativo de 7.33 que indica la habilidad obtenida por un evaluado. La discretización de esta variable se hace en tres clases cuyos valores cualitativos son: Bajo, Regular y Bueno.

Cada una de estas etiquetas lingúisticas se representa por medio de una función de pertenencia, como se muestra en Figura 3.2. Así, 7.33 corresponde a la clase Regular, con un valor de pertenencia de 0.86 y un valor de lado de izquierdo, debido a que se encuentra en el lado izquierdo respecto al valor máximo de la función de pertenencia de la clase Regular.

Con el fin de convertir los valores cuantitativos en dichas tripletas, es necesario determinar el número de clases, así como sus respectivos landmarks, mismos que separarán las clases entre ellas. Una vez definidos los parámetros que describen el número de clases y los límites entre clases (landmarks) para cada variable, se lleva a cabo el proceso de fusificación de FIR sobre los datos registrados del sistema. Al finalizar este proceso se obtiene una tripleta de matrices con el mismo número de datos, la primera contiene los valores de clase, la segunda los valores de pertenencia y la tercera los valores de lado.

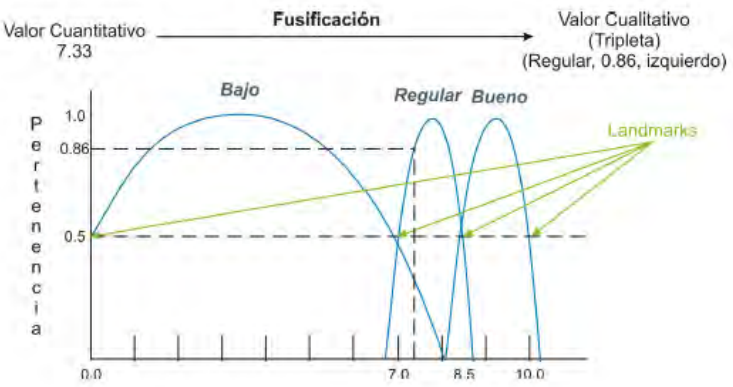


FIGURA 3.2: Proceso de fusificación de FIR.

3.1.2. Modelado cualitativo

La etapa de **Modelado Cualitativo** es la encargada de encontrar relaciones causales y temporales entre las variables del sistema, con la finalidad de obtener el modelo que mejor represente al sistema en análisis. En la notación de FIR, un modelo está compuesto por la máscara, que corresponde a la estructura del sistema, y la base de reglas patrón, que almacena el comportamiento del sistema.

La estructura del modelo cualitativo está definida por la máscara. Los valores negativos representan a un intervalo de muestreo. Al número de filas se le conoce como profundidad de la máscara. Los valores negativos representan las m-entradas o antecedentes e indican que existe relación causal de éstas con el consecuente y los valores 0 significan que no existe ninguna relación causal de éstos con el consecuente. Un ejemplo de máscara se muestra en la figura 3.3.

La secuencia en que las m-entradas y la m-salida son numeradas no tiene ningún significado especial, aunque normalmente suelen ser numeradas de izquierda a derecha y de arriba a abajo. Los términos m-entrada (entrada de la máscara) y m-salida (salida de la máscara) son utilizados para evitar confusión con las entradas y las salidas del sistema.

El número de elementos diferentes de cero en la máscara, indica el nivel de complejidad de ésta. Por ejemplo en la figura 3.3 la primera m-entrada i3(t-2t) correponde a la variable de entrada i3 muestreada dos intervalos de tiempo anteriores al actual, la segunda m-entrada i2(t-t) corresponde a la entrada del sistema i2 muestreada en el instante de tiempo anterior.

t\x	i_1	i_2	i_3	o
t-2δt	0	0	-1	0
t-δt	0	-2	0	0
t	-3	0	0	+1

FIGURA 3.3: Ejemplo de máscara.

La figura 3.4 muestra el proceso de conversión de relaciones dinámicas de los datos de entrada a las reglas patrón. En la parte izquierda se muestra un fragmento de la matríz con los valores de clase, la primera de las tres matrices que forman el conjunto de datos cualitativos. En el ejemplo, se muestran las variables de entrada del sistema: u1, u2 y u3 que se discretizaron en dos clases, mientras que la variable de salida y1 fué discretizada en tres. El rectángulo punteado simboliza que la máscara va desplazandose hacia abajo, a través de la matríz de valores de clase. Los círculos de la máscara denotan las posiciones de las m-entradas mientras que el cuadrado indica la posición de la m-salida.

Los valores de clases se leen a través de los hoyos de la máscara, y se colocan en forma consecutiva en la matriz entrada / salida. Como se puede apreciar en el lado derecho de la figura 3.4 cada renglón representa una posición de la máscara a través del recorrido que hace por la matríz de valores de clase. Estos valores se almacenan en la matríz de entrada / salida y se alinenan con el último renglón de la máscara cada renglón de esta matríz representa un estado cualitativo pseudo-estático o regla cualitativa basada en patrones.

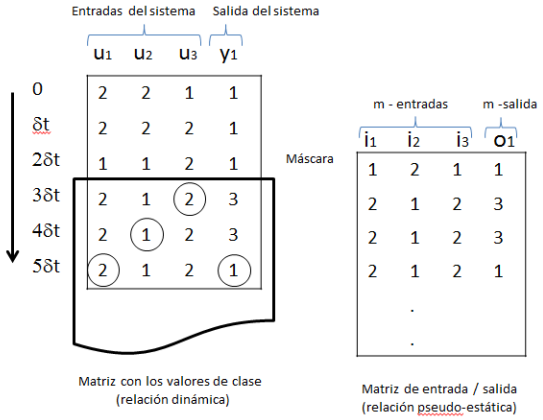


FIGURA 3.4: Proceso de obtención de reglas patrón.

3.1.3. Simulación cualitativa

Una vez identificado el modelo que mejor representa al sistema es posible realizar el proceso de predicción del comportamiento futuro del sistema, utilizando el motor de inferencia de FIR, este proceso es denominado **Simulación Cualitativa**. El objetivo principal de esta etapa de simulación es predecir los valores de la tripleta cualitativa. Basándose en el modelo obtenido por FIR (la máscara óptima y la base de reglas patrón) se realiza el proceso de predicción de valores de clase, pertenencia y lado.

Para la predicción se utiliza el motor de inferencia de FIR, que está basado en el agoritmo de los cinco vencinos más cercanos. Para el proceso de predicción la máscara es desplazada por los datos cualitativos, se extraen los valores correspondientes a las m-entradas, formando un patrón de estrada estático. Se buscan en la base de reglas patrón que coincidan con el patrón de entrada actual. Para todos

ellos se calcula su distancia respecto al mismo y se conservan aquellos 5 que esten más cercanos.

Los valores de las predicciones se obtienen como una combinación ponderada de la salida de los 5 vecinos que estén más cercanos.

Los valores cualitativos predichos (clase, pertenencia y lado) se almacena en las matrices cualitativas de clase, pertenencia y lado. Una vez predicho un valor, la mácara se desplaza una posición hacia abajo para predecir el siguiente valor.

3.1.4. Defusificación

La última etapa, denominada **Defusificación**, es la operación inversa a la fusificación, en esta etapa se convierte la salida cualitativa, predicha en la etapa anterior, en una variable cuantitativa. Aqú los valores cualitativos predichos en la fase previa, se convierten en datos cuantitativos.Para validar el modelo se comparan los resultados de predicción con los datos reales del sistema. Para calcular el error de las predicciones se puede utilizar RMS (Root Mean Square) o el MSE2 (Mean Square Error).

3.2. Plataforma Visual - FIR

Actualmente FIR está implementada como una aplicación de Matlab y recientemente se ha desarrollado la plataforma Visual - FIR que permite trabajar con FIR a través de un entorno amigable y de muy fácil manejo. Esta herramienta está pensada para el uso generalizado de FIR [9, Escobet y Nebot, 2008].

A continuación se realiza una descripción de la aplicación VisualFIR, para una una mejor comprensión de los experimentos realizados [8, Escobet, 2004].

De esta forma se desea explicar el proceso para obtener el conjunto de reglas que identifican el comportamiento y relación existente entre las variables de entrada y la variable de salida, y predecir de esta forma la habilidad inicial buscada.

Visual - FIR fué implementada como un toolbox de Matlab para su utilización, su pantalla principal muestra las cuatro fases o etapas de la metodologá FIR, (Ver Figura 3.5), así como el algoritmo para la extracción de reglas que identifican y describen patrones de comportamiento del sistema en análisis, y es invocado desde Matlab con la instrucción: VisualFIR [8, Escobet, 2004] .

Para poder identificar el mejor modelo con base en las variables determinadas, los siguientes pasos deben llevarse a cabo de manera secuencial, estos pasos corresponden a botones específicos de la pantalla principal:

- Configuración de los parámetros, cargar los datos de entrenamiento (training),

- Recodificar los datos (recoding) e identificar la máscara óptima.

Como se puede observar en Figura 3.5, la parte superior muestra la fase de Identificación del Modelo, mientras que la parte inferior corresponde a la fase de Predicción.

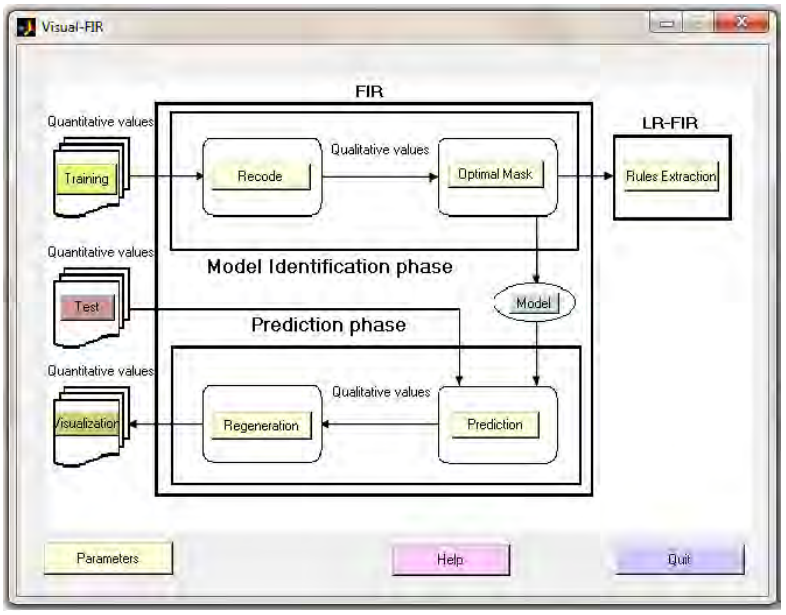


FIGURA 3.5: Pantalla principal de la aplicación Visual FIR.

3.2.1. Fase de identificación del modelo

En esta etapa se realiza la configuración de los parámetros: cargando los datos de **Training**, configurando los datos de **Recodificación** e identificando la **Máscara Óptima**.

El botón **Training**. Antes de iniciar con la fase de identificación del modelo es necesario ingresar el conjunto de datos, como lo muestra Figura 3.6. El conjunto de datos debe estar contenido en un archivo con extensión .mat, que debe incluir en variables de Matlab cada variable del sistema de manera separada.

El botón de **Recode** activa la pantalla para discretizar al conjunto de variables (convertidas de valores cuantitativos a valores cualitativos difusos). Es necesario entonces cargar los datos que fueron previamente especificados en el paso anterior.

La lista de todas las variables (de entrada y salida) es desplegada en la parte izquierda de esta misma pantalla. Estas variables son discretizadas por default en tres clases con el algoritmo de Equal Frecuency Partition, asignando un valor mínimo y un valor maximo para cada clase (landmarks), de igual manera es posible aplicar algún otro algoritmo o bien modificar estos parámetros individualmente de forma manual.

FIGURA 3.6: Pantalla de training.

En la figura 3.7 se observa la discretización realizada a las variables, mostrando del lado izquierdo de la pantalla el nombre de las variables, el algoritmo empleado para la discretización de las mismas y el número de clusters, mientras que del lado derecho se observa, entre otras opciones, los parámetros de cada una de ellas.

Por ejemplo: la variable Promedio ha sido discretizada de forma manual en tres clusters, estableciendo landmarks en cada uno de ellos, y es entonces cuando al presionar el botón recode, se realiza la conversión de los datos cuantitativos en tripletas cualitativas, generando las clases, pertenencia y el valor de lado para cada uno de los datos cuantitativos del conjunto de datos de entrenamiento.

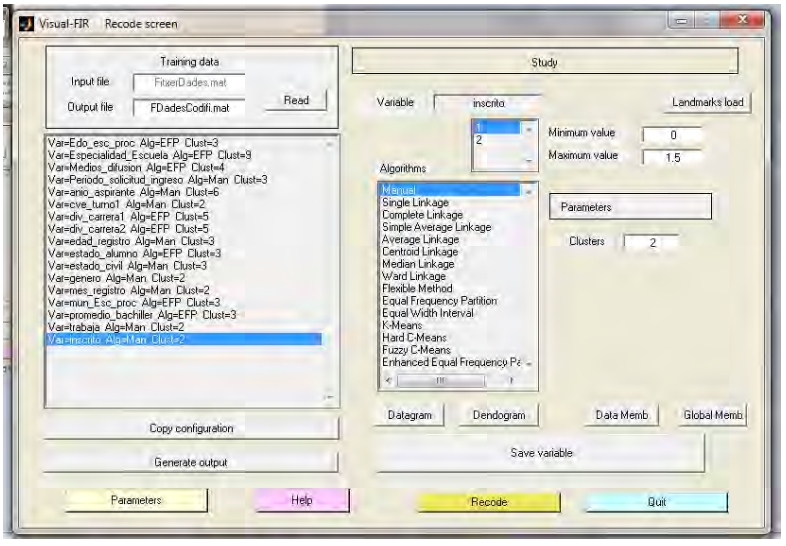


FIGURA 3.7: Pantalla de recodificación.

Una vez que se han establecido los parámetros de las variables involucradas, se procede a identificar la máscara que mejor describa la relación que existe entre dichas variables con fin de predecir de la mejor forma posible el comportamiento que tendrá el sistema (una máscara representa una relación entre las variables). Esta acción se realiza con el botón **Optimal Mask**, para lo cual es necesario establecer:

- La complejidad (número de elementos diferentes de cero en la máscara), y
- La profundidad (número de filas de la máscara), la obtención de dicha máscara se realiza por medio de un mecanismo de búsqueda exhaustiva.

Como resultado de mencionada búsqueda, FIR muestra aquellas máscaras con la profundidad especificada desde complejidad igual a uno hasta n-1, (donde n es la complejidad introducida), además de listar la máscara óptima y máscaras subóptimas para cada complejidad.

De entre éstas, se considera máscara óptima de FIR aquella cuya Calidad (q) sea mayor en relación con las demás máscaras.

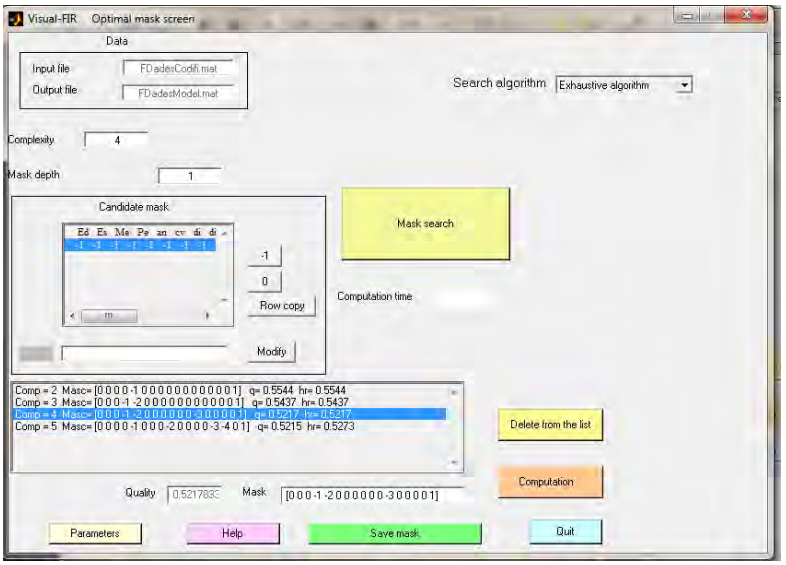


FIGURA 3.8: Pantalla de máscara óptima.

La figura 3.8 muestra la ventana que identifica la máscara óptima, la cual tiene la siguiente forma: [0 0 0 -1 -2 0 0 0 0 0 0 -3 0 0 0 0 0 1]

Los valores negativos se refieren a aquellas variables de entrada inputs que tienen relación causal, los valores positivos se refieren a las variables de salida outputs, mientras que los valores en cero, representan que no hay relación entre las variables y la salida. Continuando con el ejemplo, se observa por tanto que las variables de entrada que tienen relación con la variable Promedio son:

- División de Carrera.

- Estado Escuela de Procedencia.
- Medios de Difusión

3.2.2. Fase de predicción

Una vez identificado el modelo FIR (reglas, patrón y máscara), se procede a usar dicho modelo para predecir el comportamiento futuro del sistema, figura 3.9. Para ello se procede a la fase de predicción que se compone de cuatro pasos principales:

- **Test data.** Esta etapa ofrece las mismas opciones que Training, especificando variables de entrada y salida de acuerdo a los parámetros introducidos en la fase de training, pero para los datos de test, es decir aquellos utilizados para validar la generalidad del modelo identificado.
- **Predicción.** El algoritmo de búsqueda exhaustiva es ejecutado utilizando diversas opciones contenidas en el botón de parámetros.
- **Regeneración.** Finalmente, la predicción cualitativa difusa se convierte a datos cuantitativos mediante este proceso.
- **Visualización.** Muestra el resultado final del proceso de predicción, observando el error obtenido con el modelo seleccionado. Este error determina la relación entre los valores reales y los valores obtenidos con la predicción.

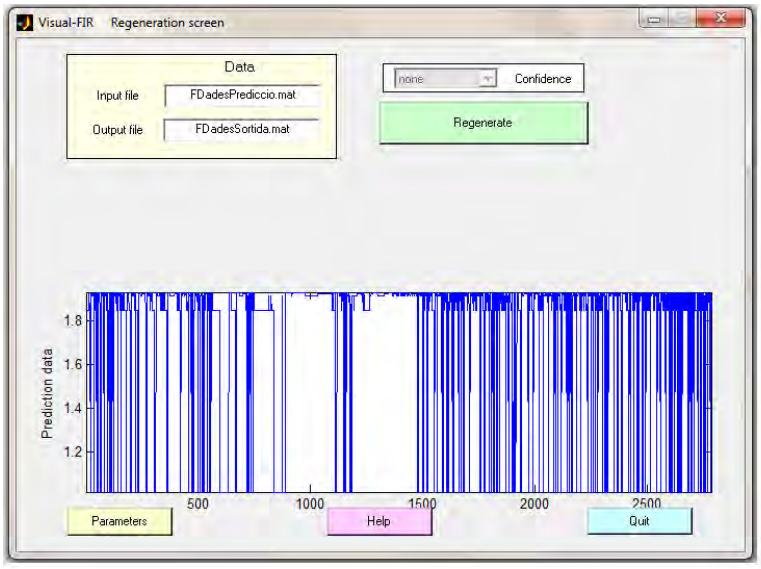


FIGURA 3.9: Pantalla de regeneración.

.

Finalmente después de obtener el valor cuantitativo de la predicción sólo es necesario compararlo con los datos originales, permitiendo validar el modelo cualitativo, lo cual puede observarse de forma gráfica en la pantalla de Visualización, figura 3.10.

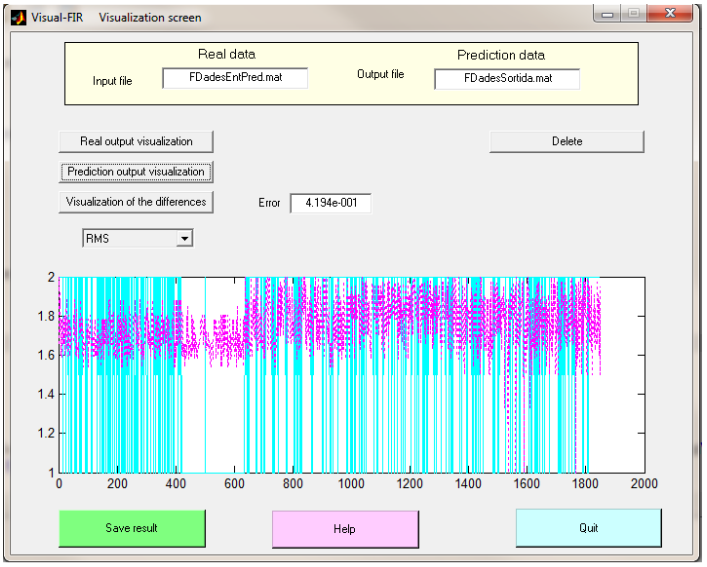


FIGURA 3.10: Pantalla de visualización.

.

3.3. LR - FIR, Algoritmo para extracción de reglas lingüsticas

LR - FIR (Algoritmo de extracción de reglas lingüisticas) es una extensión de FIR, que toma como entrada las reglas patrón generadas por FIR, tal como lo muestra Figura 3.11.

Este algoritmo fue desarrollado por Castro y sus colegas [2, Castro et al.(1), 2007] como una alternativa a los métodos de compactación de reglas utilizados en álgebra booleana, con la ventaja que puede compactar reglas con atributos multi-valuados [5, Castro and Nebot, 2007].

Es importante aclarar que el algoritmo es dependiente del número de variables y reglas patrón, por lo que, cuanto mayor sean éstos, mayor será el tiempo que el algoritmo tardará en compactar las reglas originales. Sin embargo, las reglas ya compactadas son mucho más intuitivas y entendibles que las originales. Las reglas compactadas describen patrones de comportamiento del sistema analizado, identificando las variables que son realmente importantes en la predicción de los datos de salida que correspondan a esos patrones.

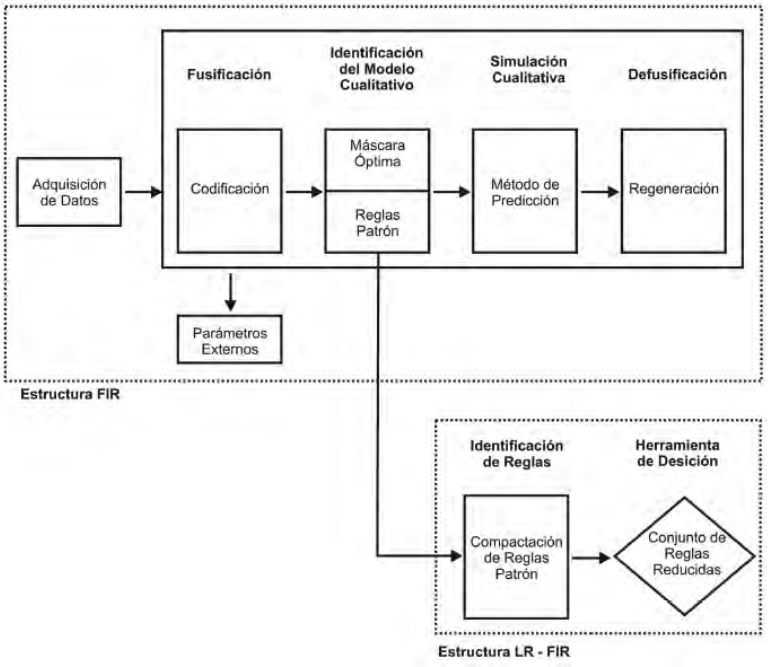


Figura 3.11: Estructura FIR LR - FIR.

A continuación se describen los pasos del algoritmo LR - FIR (ver figura 3.12):

1. **Compactación básica.** Es un proceso iterativo que evalúa una a la vez todas las reglas patrón generadas por FIR. En esta etapa se compactan las reglas patrón basándose en el conocimiento existente en las reglas patrón generadas por FIR. Un subconjunto de reglas, puede ser compactado en una regla simple, cuando todas las premisas, excepto una, así como el consecuente, tienen los mismos valores. En

FIGURA 3.12: Pasos principales del algoritomo LR - FIR

el contexto de este proyecto, las premisas se corresponden con las m-entradas y el consecuente con la m-salida (ver figura 3.13).

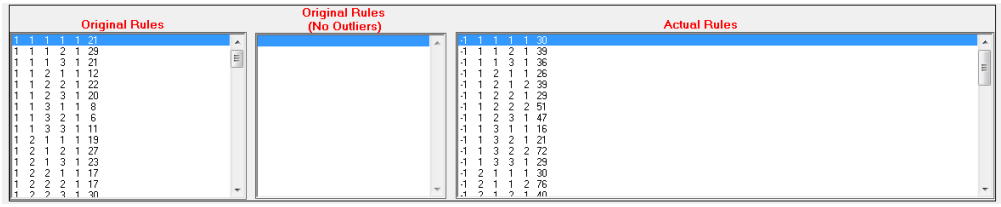
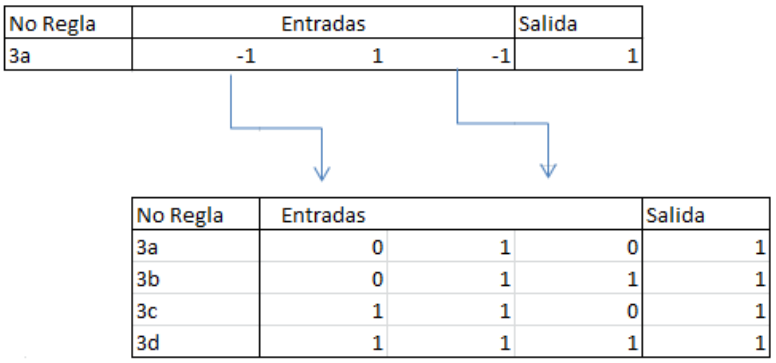


FIGURA 3.13: Compactación básica de reglas.

2. **Compactación mejorada.** En la compactación básica, sólo se estructura y representa el conocimiento disponible en una forma más compacta. En la compactación mejorada se asumen creencias para la generación de reglas compactas, a diferencia del algoritmo básico, que para compactar un subconjunto de reglas, debe existir todo el conocimiento suficiente en las reglas patrón originales. En el paso de compactación mejorada, se extiende la base de conocimiento, a casos que no se utilizaron para construir el modelo.

Para realizar la compactación mejorada se tienen 2 opciones: En la primera opción, utilizando las reglas obtenidas en la compactación básica, en un proceso iterativo, todas las premisas, P, que tienen valores no negativos (premisas no compactadas), en todas las reglas son reemplazados por valores -1. Se expande la regla a todos sus valores legales, basandose en la discretización utilizada para la variables, ver figura 3.14 y se comparan las reglas resultantes con las reglas originales.

Si no existen conflictos la regla se acepta y contrariamente se rechaza (un conflicto ocurre cuando una o más reglas expandidas, tienen los mismos valores en todas las premisas, pero diferente valor en el consecuente). La opción 2 es una extensión de la compactación básica, donde para compactar un subconjunto de reglas, debe existir un porcentaje mínimo, pero razonable de los valores legales, en las reglas candidatas.



No Regla	Entradas			Salida
3a	-1	1	-1	1

No Regla	Entradas			Salida
3a	0	1	0	1
3b	0	1	1	1
3c	1	1	0	1
3d	1	1	1	1

FIGURA 3.14: Ejemplo de expansión de una regla a sus valores legales.

3. **Eliminación de reglas duplicadas.** En este paso se eliminan reglas duplicadas que se pudieran generar durante la etapa de compactación mejorada. Esta situación puede presentarse debido a que al aplicar el algoritmo de compactación mejorado (paso anterior) sólo se evalúan conflictos con las reglas originales.

4. **Eliminación de reglas en conflicto.** Durante los pasos previos del algoritmo es posible encontrar reglas ambiguas o en conflicto, es decir, que tienen el mismo valor en todas las premisas (relaciones causales que permiten determinar el valor del consecuente, y se denotan con valores negativos) pero valor diferente en el consecuente (corresponde a la variable de salida que deseamos predecir y se denota con un valor positivo). Por ejemplo, las reglas: 1 2 1 3 y 1 2 1 1, tienen el mismo valor en las premisas (1 2 1) y diferente valor en el consecuente (3 y 1).

Estas reglas son ambiguas, por lo que se debe eliminar una de ellas y no causar conflicto y confusión al momento de evaluar el sistema en análisis. Las métricas que se utilizan para eliminar reglas en conflicto son la especificidad y la sensitividad.

Es importante aclarar que cuando se tienen reglas con valores del consecuente consecutivos, no se debe considerar como un conflicto, ya que comparten espacios de entrada contiguos por lo que se deben unificar en el paso posterior del algoritmo.

Si se tienen las reglas: 1 2 1 1 y 1 2 1 2, aparentemente parece ser un conflicto, sin embargo, en el algoritmo se consideran como 2 reglas que se deben unificar en una única regla: 1 2 1 (1-2).

5. **Unificación de reglas.** El paso de unificar reglas se lleva a cabo en 2 etapas. En una primera fase se unifican reglas que comparten espacios de entrada contiguos en alguna variable (premisas y consecuente) y que tengan el mismo valor en el resto de variables. En la segunda etapa se evalúan las reglas unificadas en la primera fase y las reglas no unificadas, en busca de nuevas unificaciones. Por ejemplo, las

reglas 1 2 1 2 y 1 2 2 2, comparten en la tercera variable espacios de entrada contiguos (1 y 2), por lo que se pueden unificar en la regla: 1 2 (1-2) 2.

6. **Evaluación de las reglas obtenidas.** En este paso se evalúan las reglas obtenidas en la etapa anterior, utilizando métricas estándares que permiten obtener una evaluación objetiva y realista, independientemente de la distribución de los datos de cada clase. Estas métricas son las siguientes:

Sensitividad = TP / (TP + FN)

Especificidad = TN / (TN + FP)

Donde:

TP ("True Positive") = Número de datos que la regla predice que pertenece a la clase x, y que realmente pertenecen a la clase x.

FN ("False Negative") = Número de datos que la regla predice que no pertenece a la clase x, y que realmente pertenecen a la clase x.

FP ("False Positive") = Número de datos que la regla predice que pertenece a la clase x, y que realmente no pertenecen a la clase x.

TN ("True Negative") = Número de datos que la regla predice que no pertenece a la clase x, y que realmente no pertenecen a la clase x.

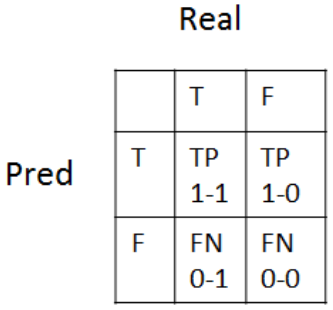


FIGURA 3.15: Matríz de confusión

7. **Eliminación de reglas de baja calidad.** Como última etapa del algoritmo de extracción de reglas se eliminan aquellas reglas de baja calidad, es decir, con valores bajos de las métricas de evaluación, especificidad y sensitividad.

Para la extracción de reglas, es necesario ejecutar la aplicación Rules Extraction, desde VisualFIR, misma que se basa en el algoritmo LF-FIR.

Figura 3.16 permite visualizar las reglas obtenidas tal como las muestra la interfaz del algoritmo.

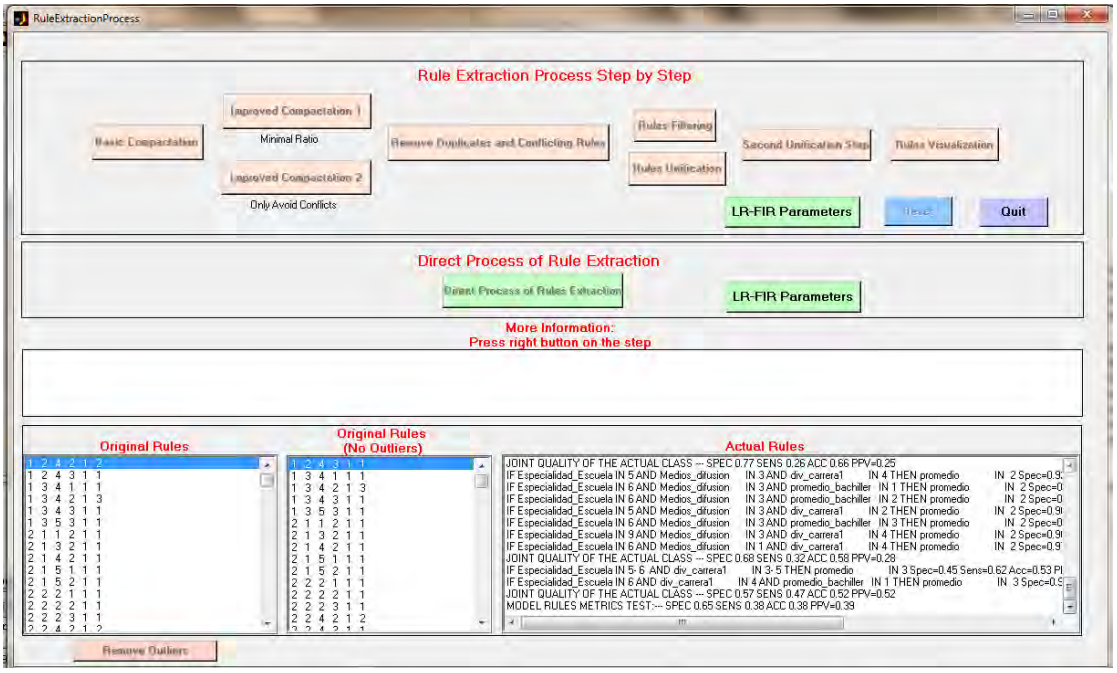


FIGURA 3.16: Extracción de reglas.

Capítulo 4

Evaluación Experimental

En este capítulo se describirá como se llevó a cabo la fase experimental a partir de la información provista por la UTXJ, la cual tuvo que someterse a una serie de tratamientos comparativos entre fuentes de información para acreditar que los datos manejados fueran confiables al aplicarlos durante los experimentos, así mismo se utilizó la metodología FIR junto con LR - FIR, a través de la realizacíon de diversos experimentos que son explicados en el presente capítulo.

La primer fuente de información proviene de la recepción del aspirante dentro del área de admisión escolar y que se registra en una hoja de cálculo electrónica.

La segunda fuente de información se referencia a la Base de Datos que controla la información del Sistema Informático SAIIUT, que es donde se registra toda la información académica Institucional de la UTXJ, y a través de la ejecución de búsquedas específicas a las tablas relacionadas a la información de interés (admisión, calificaciones, inscripción, municipios, estados, escuelas de procedencia, entre otras) de las cuales se obtuvieron archivos específicos en formato de hoja electrónica.

La tercer fuente de información fué obtenida de un servidor Institucional que contiene varias Aulas Virtuales basados en Moodle con versión para la plataforma operativa Debian Squeeze versión 6, epecíficamente se utilizó el Aula Virtual dedicada a aplicar el examen de Admisión Institucional, el cual tiene 3 áreas de conocimiento: Habilidad Matemática, Habilidad Verbal y Habilidad en Inglés.

Cabe señalar que algunas fuentes de información como resultados del EXANI II de CENEVAL, de Trayectorias Educativas así como estadística institucional, no fueron incorporadas a los ejercicios de prueba, porque su deteccción y consideración fueron después del tiempo planeado, así que valorando el tiempo de su acopio, tratamiento y manejo, se planteó su uso en trabajos futuros al presente.

4.1. Preparación de la información

Para llevar a cabo cada experimento se tomó un universo de información que consideró a 6 generaciones de aspirantes, del año 2009 al 2014, con un universo de 5547 estudiantes, todos de la UTXJ

Los datos básicos que se consideran en todos los experimentos fueron los siguientes:

Año Aspirante, Opción Carrera 1, Turno Carrera 1, Estado Escuela Procedencia, Especialidad Escuela de Procedencia, Mes Registro del Aspirante, Trabaja, Promedio del Bachiller, Periodo Solicitud Ingreso, Medios Difusión, Estado Civil, Edad al momento del Registro del Aspirante, Género, Estado de procedencia del Alumno, Mes de aplicación del examen de Admisión Institucional, Calificación del examen de Admisión apartado Habilidad Matemática, Habilidad Verbal e Inglés, Tiempo dedicado a contestar el examen de Admisión en el apartado de Habilidad Matemática, Habilidad Verbal e Inglés.

Los datos específicos según el experimento se muestran a continuación:

Primer experimento, Predecir que el aspirante se inscriba:
Sólo se agregó Inscrito, como variable de salida.

Segundo experimento, Predecir la conclusión de estudios de cualquier plan:
Sólo se agregó, Promedio, Término de Estudios como variable de salida. Se excluye: Opción Carrrera 2

Tercer Experimento, Predecir que el promedio final del aspirante:
Se utilizaron los mismos datos que en el experimento dos y se configuró Promedio como variable de salida. Se agregaron las calificaciones de todas las materias.

4.1.1. Datos simples de clasificar

Los datos que no requirieron una clasificación exaustiva fueron:

- Año de solicitud del aspirante
- Turno de la Carrera
- Estado Civil
- Mes de registro del aspirante
- Trabaja
- Promedio bachiller
- Periodo solicitud ingreso
- Edad al momento del registro

- Género
- Mes de aplicación del exámen de Admisión
- Tiempo dedicado a aplicar el examen de admisión en la sección del exámen de Habilidad Matemática, Verbal e Inglés
- Calificación del exámen de admisión en la sección de Habilidad Matemática, Verbal e Inglés
- Inscrito

4.1.2. Áreas académicas y su clasificación por área del conocimiento

Cada uno de los aspirantes tuvieron la oportunidad de seleccionar dos opciones de carrera, las cuales están clasificadas en 5 áreas académicas:

- Agroalimentaria
- Económico - Administrativa
- Electromecánica Industrial
- Salud
- Tecnologías de la Información y Comunicación

Así mismo se hizo la clasificación de carreras por división del conocimiento como se muestra en el cuadro 4.1, donde se tiene asignada la división 1 al área de la salud representada por Terapia Física, la divisón 2 al área Agroalimentaria Biotecnológica, la división 3 al área Económico Administrativa, la división 4 al área Electromecánica Industrial y la división 5 al área de Tecnologías de la Información.

4.1.3. Especialidad de escuelas de procedencia

Se hizo una clasificación de 9 áreas del conocimiento para Especialidad de las Escuelas de Procedencia consideremos entonces el cuadro 4.2, donde la clasificación 1 se refiere a las áreas relacionadas a Mantenimiento Industrial, la clasificación 2 al área Agroalimentaria Biotecnológica, la clasificación 3 al área de la Salud, la clasificación 4 al área de la Educación, la clasificación 5 a las Tecnologías de Información, la clasificación 6 a las Económico Administrativas, la 7 a las de Civíl y Arquitectura, la 8 a Turismo y la 9 a Mecatrónica.

Clave Carrera	Nombre	División
47	Terapia Física, Área Rehabilitación	1
6	Biotecnología	2
22	Procesos Agroindustriales	2
25	Procesos Alimentarios	2
27	Química Área Biotecnología	2
42	Procesos Alimentarios	2
49	Gastronomía	2
50	Biotecnología	2
1	Administración	3
37	Administración Área Recursos Humanos	3
43	Desarrollo e Innovación Empresarial	3
53	Desarrollo de Negocios área Mercadotecnia	3
11	Electrónica y Automatización	4
14	Mantenimiento Industrial	4
38	Mecatrónica Área Automatización	4
39	Mantenimiento Área Industrial	4
41	Mantenimiento Industrial	4
44	Mecatrónica	4
46	Fotónica, Área Telecomunicaciones	4
48	Robótica Industrial	4
51	Mantenimiento Área Petróleo	4
52	Mecatrónica Área Sistemas de Manufactura Felxible	4
13	Informática	5
34	Tecnologías de la Información y Comunicación Área Sistemas Informáticos	5
40	Tecnologías de la Información y Comunicación	5
45	Tecnologías de la Información y Comunicación, Área Multimedia y Comercio Electrónico	5

Cuadro 4.1: Clasificación de carreras por división del conocimiento.

4.1.4. Medios de difusión

Fué importante clasificar los distintos medios de difusión por los cuales el aspirante se enteró de la UTXJ como se muestra en el cuadro 4.3, donde se clasificó de manera general la difusión hecha por familiares, amigos, terceros y la UTXJ, considerando el origen de impacto que cada uno tiene.

4.2. Predicción del Índice de Inscripción

4.2.1. Consideración de variables

Las variables consideradas para el primer experimento fueron 21, las cuales se describen en el cuadro 4.4. Una vez reunidos los datos, se divieron para ser utilizados en el experimento en dos terceras partes para el entrenamiento y una tercera parte para la prueba, así mismo se utilizó la metodología FIR para identificar el mejor modelo que represente el comportamiento del sistema. Para lo cual se fusificaron las primeras variables, resultando la discretización de las mismas como se muestra en el cuadro 4.20.

Clave especialidad	Descripción	Área	Grupo
1	Instalaciones Hidraulicas Y Eléctricas	Mantenimiento Industrial	1
3	Mecánica	Mantenimiento Industrial	1
20	Técnico Electromecánico	Mantenimiento Industrial	1
27	Técnico En Máquinas Y Herramientas	Mantenimiento Industrial	1
45	Mantenimiento Industrial	Mantenimiento Industrial	1
2	Laboratorio Químico	Área Agroalmentaria Biotecnológica	2
9	Químico Biológico con Turismo	Área Agroalmentaria Biotecnológica	2
10	Químico Biológico e Informática	Área Agroalmentaria Biotecnológica	2
11	Químico Biologo	Área Agroalmentaria Biotecnológica	2
12	Químico Biologo en el Área De Alimentos	Área Agroalmentaria Biotecnológica	2
16	Técnica Agropecuaria	Área Agroalmentaria Biotecnológica	2
18	Técnica en Informática Agropecuaria	Área Agroalmentaria Biotecnológica	2
19	Técnico Agropecuario	Área Agroalmentaria Biotecnológica	2
22	Técnico en Alimentos	Área Agroalmentaria Biotecnológica	2
29	Técnico en Puericultura	Área Agroalmentaria Biotecnológica	2
32	Tecnología de Los Alimentos	Área Agroalmentaria Biotecnológica	2
33	Tecnología en Alimentos	Área Agroalmentaria Biotecnológica	2
36	Viveros	Área Agroalmentaria Biotecnológica	2
43	Procesos Agroindustriales	Área Agroalmentaria Biotecnológica	2
46	Procesos Agroindustriales	Área Agroalmentaria Biotecnológica	2
4	Nutrición	Salud	3
28	Técnico En Nutricion	Salud	3
31	Técnico Laboratorista Clinico	Salud	3
48	Enfermería General	Salud	3
5	Práctica Docente	Educación	4
7	Promotor Deportivo	Educación	4
8	Psicopedagógico	Educación	4
30	Técnico en Trabajo Social	Educación	4
34	Trabajadora Social	Educación	4
37	Físico Matemático	Educación	4
42	Ciencias y Humanidades	Educación	4
6	Programador Analista	Tecnologías de la Información y Comunicación	5
17	Técnica en Computación	Tecnologías de la Información y Comunicación	5
23	Técnico en Computación	Tecnologías de la Información y Comunicación	5
44	Tecnologías de la información y comunicación	Tecnologías de la Información y Comunicación	5
13	Recursos Humanos	Económico Administativa	6
14	Relaciones Humanas	Económico Administativa	6
15	Secretaria Ejecutiva En Español	Económico Administativa	6
21	Técnico en Administracion	Económico Administativa	6
24	Técnico en Computacion Fiscal Contable	Económico Administativa	6
26	Técnico en Contabilidad	Económico Administativa	6
38	Económico-Administrativo	Económico Administativa	6
40	Informática Administrativa	Económico Administativa	6
47	Administración	Económico Administativa	6
25	Técnico en Construcción Urbana	Civil Arquitectura	7
41	Dibujo Arquitectónico	Civil Arquitectura	7
35	Turismo	Turismo	8
39	Electrónica	Mecatrónica	9

Cuadro 4.2: Clasificación de especialidad de escuela de procedencia

4.2.2. Algunos aspectos a considerar para discretizar los valores

Concentrando los datos fuente y las características de todos los estudiantes, se obtuvo una matriz del total de aspirantes de cada una de las 6 generaciones por el número de variables, esto es: 5547 por 21.

Aquellos datos con valor igual a -1, representan datos perdidos, para los cuales se tiene un tratamiento especial con el fin de que no altere los resultados de predicción. Cada tabla refiere a un valor porcentaje, el cual significa la magnitud del porcentaje del atributo en todo el universo. Algunas de las características que describen al total de aspirantes se concentra en los cuadros 4.5 - 4.19.

Clave medio difusión	Descripción	Origen	Grupo
1	Padres	Familiar	1
2	Alumno de tu escuela	Amigos	2
3	Profesor de la Universidad	UT	3
4	Visita de la Universidad al plantel	UT	3
5	Volantes	UT	3
6	Módulo en centro comercial	UT	3
7	Televisión	Terceros	4
8	Familiares	Familiar	1
9	Directivos de tu escuela	Terceros	4
10	Visita a la Universidad	UT	3
11	Carteles	UT	3
12	Mantas	UT	3
13	Módulo en la Feria Municipal	UT	3
14	Periódicos	UT	3
15	Amigos	Amigos	2
16	Alumnos de la Universidad	Amigos	2
17	Vía telefónica	UT	3
18	Trípticos	UT	3
19	Automovil con alta voz	UT	3
20	Radio	UT	3
21	Internet	UT	3
22	Feria Profesiográfica	UT	3
23	Conocidos	Amigos	2
24	Folleto	UT	3
25	Espectacular	UT	3
26	Carta invitación	UT	3
27	Directorio Telefónico	UT	3
28	Un día en la UTH	UT	3
29	Visita individual a la UTH	UT	3
30	Grupos Culturales de la UTH	UT	3
31	Practicas en talleres de la UTH	UT	3
32	Presidencia Municipal	UT	3
33	Empresa	Terceros	4
34	Admisión BUAP o ITP	Terceros	4
35	Feria de Orientación Vocacional	UT	3
36	Otro	Terceros	4
37	Etiqueta de Agua	UT	3
38	Feria	UT	3
39	Ubicación de la UTH	UT	3
40	Trabajador de la UTH	UT	3
41	Feria profesiográfica del IPJ	UT	3

Cuadro 4.3: Clasificación de medios de difusión

En el cuadro 4.5 se observa el incremento poblacional conforme avanzan los años, siendo importante notar el incremento casi constante a partir del año 2012.

En el cuadro 4.6 se describe la proporción de matrícula por área académica la cual fué optada como primer opción por el aspirante, además que el área con mayor demanda histórica acumulada es Mantenimiento Industrial representada por la División Carrera 4.

En el cuadro 4.7 se observa que la mayoría poblacional ha estudiado en un esquema presencial seguido de un semipresencial o mixto.

En el cuadro 4.8 se muestra una clasificación de acuerdo a la ubicación por estadodel bachillerato, notándose en su mayoría que los bachilleratos que influyen pertenencen al estado de Puebla.

No.	Variable	Descripción	Rango
2	Carrera 1	Selección de la primer Carrera	1 – 5
3	Turno de la Carrera	Turno de la primer carrera seleccionada	1 –3
5	Estado escuela procedencia	Estado Bachillerato de procedencia del alumno	2 – 30
6	Especialidad Escuela	Especialidad del Bachillerato de procedencia	1 – 9
7	Mes de registro del aspirante	Mes en que se registro el aspirante	1 – 10
8	Trabaja	Estatus laboral	1 – 2
9	Promedio bachiller	Promedio final del Bachiller	6 – 10
10	Periodo solicitud ingreso	Anõs que transcurrieron después del egreso del bachiller	0 – 37
11	Medios difusion	Medio por el que se enteró de la UTXJ	1 – 4
12	Estado Civil	Estatus Civil	1 – 5
13	Edad aspirante	Años al momento del registro	17 – 58
14	Género	Género	1, 2
15	Estado del alumno	Estado de procedencia del aspirante	8 – 30
16	Mes examen	Mes de aplicación del examen de Admisión	1 – 2
17	Calificación 1	Calificación Habilidad Matemática n	0 – 10
18	Calificación 2	Calificación Habilidad Verbal	0 – 10
19	Calificación 3	Calificación Habilidad Inglés	0 – 10
20	Tiempo de ejecución 1	Tiempo de ejecución Habilidad Matemática	0 – 90
20	Tiempo de ejecución 2	Tiempo de ejecución Habilidad Verbal	0 – 120
20	Tiempo de ejecución 3	Tiempo de ejecución Habilidad Inglés	0 – 60
21	Inscrito	Estatus después del periodo de admisión	1 – 2

Cuadro 4.4: Variables a utilizar para el primer experimento

Año aspirante	Cantidad	Porcentaje
2009	591	10.65
2010	667	12.02
2011	646	11.65
2012	1001	18.05
2013	1121	20.21
2014	1521	27.42
	5547	100.00

Cuadro 4.5: Variable aspirante.

División Carrera 1	Cantidad	Porcentaje
1	704	12.69
2	801	14.44
3	896	16.15
4	2441	44.01
5	705	12.71
	5547	100.00

Cuadro 4.6: Variable división carrera.

Turno de Carrera 1	Cantidad	Porcentaje
1	4982	89.81
2	48	0.87
3	517	9.32
	5547	100.00

Cuadro 4.7: Variable turno de la carrera.

En el cuadro 4.9 se muestra una clasificación por especialidad de Bachilleratos en 9 áreas del conocimiento siendo el área Económico Administrativa quien tiene mayor presencia seguido del área de Tenologías de la Información, la Agroalimenta Biotecnológica, aunque este dato contrasta con la demanda vocacional de los aspirantes al área Electromecánica Industrial.

Estado Escuela Procedencia	Cantidad	Porcentaje
1	5087	91.71
2	430	7.75
3	22	0.40
-999	8	0.14
	5547	100.00

Cuadro 4.8: Variable estado del bachillerato de procedencia.

Especialidad Escuela	Cantidad	Porcentaje
1	203	3.66
2	586	10.56
3	54	0.97
4	429	7.73
5	1586	28.59
6	2200	39.66
7	75	1.35
8	49	0.88
9	365	6.58
	5547	100.00

Cuadro 4.9: Variable especialidad del bachillerato de procedencia.

El cuadro 4.10 muestra la distinción entre dos periodos de tiempo del proceso de admsión, uno de enero a mayo y el segundo de junio a octubre, siendo éste último en donde se capta la mayor parte de aspirantes.

Mes Registro del Aspirante	Cantidad	Porcentaje
1 – 5	1135	20.46
6 – 10	4412	79.54
	5547	100.00

Cuadro 4.10: Variable mes de registro del aspirante.

El cuadro 4.11 muestra que la gran mayoría de los aspirantes no ejerce alguna actividad económica.

Trabaja	Cantidad	Porcentaje
1	4688	84.51
2	859	15.49
	5547	100.00

Cuadro 4.11: Variable trabaja.

El cuadro 4.12 muestra que la mayoría de los aspirantes egresa del Bachiller con un promedio regular entre 7.5 y 8.5.

Promedio Bachiller	Cantidad	Porcentaje
P menor igual 7.5	1584	28.56
7.5 menor P menor o igual 8.5	2341	42.20
8.5 menor P menor 10	1622	29.24
	5547	100.00

Cuadro 4.12: Variable promedio alcanzado en Bachillerato.

El cuadro 4.13 muestra que la mayoría de los estudiantes solicitó su ingreso a la UTXJ inmediatamente que egreso del Bachillerato.

Periodo solicitud ingreso	Cantidad	Porcentaje
0 años	3887	70.07
1 -2 años	1003	18.08
2 - 4 años	273	4.92
4 - 37 años	384	6.92
	5547	100.00

Cuadro 4.13: Variable periodo que un aspirante esperó para registrarse en la UTXJ

El cuadro 4.14 describe que el medio que influyó más en los estudiantes fué promovido por la UTXJ a través de eventos detallados en el cuadro 4.3.

Medios difusion	Cantidad	Porcentaje
1	973	17.54
2	1550	27.94
3	2925	52.73
4	99	1.78
	5547	100.00

Cuadro 4.14: Variable medio de difusión por el que se enteró el aspirante de la UTXJ

El cuadro 4.15 muestra que la mayoría de los aspirantes tienen el estatus civíl de soltero, lo cual puede ser un punto de oportunidad para mantener cautivo en la academia a cualquier estudiante.

Estado civil	Cantidad	Porcentaje
0	137	2.47
1	5231	94.30
2	122	2.20
3	6	0.11
4	1	0.02
5	50	0.90
	5547	100.00

Cuadro 4.15: Variable estado civil.

En el cuadro 4.16 se observa que la mayoría de aspirantes solicita su ingreso a la UTXJ a una edad idónea correspondiente al egreso del Bachiller.

Edad registro	Cantidad	Porcentaje
17 - 18	2628	47.38
19 - 20	1854	33.42
21 - 58	1065	19.20
	5547	100.00

Cuadro 4.16: Variable edad al momento del registro.

El cuadro 4.17 describe que hay un mínimo de diferencia porcentual poblacional entre hombres y mujeres que son aspirantes al ingreso a la UTXJ

Género	Cantidad	Porcentaje
1	2691	48.51
2	2856	51.49
	5547	100.00

Cuadro 4.17: Variable género.

En el cuadro 4.18 se observa que la mayroría de aspirantes son originarios del estado de Puebla.

Estado alumno	Cantidad	Porcentaje
1	5155	92.93
2	376	6.78
3	4	0.07
desconocido	12	0.22
	5547	100.00

Cuadro 4.18: Variable estado de procedencia del alumno.

En el cuadro 4.19 se muestra que la mayoría de aspirantes completó el proceso de inscripción.

Inscrito	Cantidad	Porcentaje
1	1342	24.19
2	4205	75.81
	5547	100.00

Cuadro 4.19: Variable inscrito.

El cuadro 4.20 muestra la discretización propuesta para las 21 variables en el experimento de predecir y caracterizar el índice de inscripción, las cuales en su mayoría muestran valores que no son contínuos, por lo que en la columna de landmarks se muestran los límites de cada clase.

No.	Variable	No Clases	Clase	Landmarks
1	División Carrera 1	5	1	0 – 1
			2	1 – 2
			3	2 – 3
			4	0 - 1
			5	0 - 1
2	Turno de la Carrera	2	1	0 - 1
			2	1 – 3
3	Estado escuela procedencia	3	1	0 – 1
			2	1 - 2
			3	2 - 3
4	Especialidad Escuela	9	1	0 – 1
			2	1 – 2
			3	2 – 3
			4	3 – 4
			5	4– 5
			6	5 –6
			7	6 –7
			8	7 – 8
			9	8 –9
5	Mes de registro del aspirante	2	1	1 – 5
			2	5 – 10
6	Trabaja	2	1	0 – 1
			2	1 – 2
7	Promedio bachiller	3	1	6 – 7.5
			2	7.5 – 8.5
			3	8.5 – 10
8	Periodo solicitud ingreso	4	1	0 – 0.5
			2	0.5 – 2.5
			3	2.5 – 4.5
			4	4.5 – 37
9	Medios difusion	4	1	0 – 1
			2	1 –2
			3	2 – 3
			4	3 – 4
10	Estado Civil	3	1	0 – 0.5
			2	0.5 – 1.5
			3	1.5 – 5
11	Edad aspirante	3	1	17 – 19
			2	19 – 21
			3	21 – 58
12	Género	2	1	Femenino
			2	Masculino
13	Estado del alumno	3	1	1 – 1.5
			2	1.5 – 2.5
			3	2.5 – 3
14	Mes examen de Admision	2	1	0 – 5.5
			2	5.5 – 10
15	Calificación Habilidad Matemática	3	1	0 – 3.64
			2	3.64 – 5.24
			3	5.24 – 10
16	Calificación Habilidad Verbal	3	1	6 – 7.5
			2	7.5 – 8.5
			3	8.5 – 10
17	Calificación Habilidad Inglés	3	1	6 – 7.5
			2	7.5 – 8.5
			3	8.5 – 10
18	Tiempo Habilidad Matemática	3	1	0 – 15
			2	15 – 27
			3	27 – 135
19	Tiempo Habilidad Verbal	3	1	0 – 24
			2	24 – 37
			3	37 – 174
20	Tiempo Habilidad Inglés	3	1	0 – 4
			2	4 – 49
			3	49 – 99
21	Inscrito	2	1	No Inscrito
			2	Inscrito

CUADRO 4.20: Discretizando los valores de las variables a utilizar en el experimento.

4.2.3. Identificación del modelo

Considerando la información de la UTXJ anteriormente detallada sobre los aspirantes con un cohorte de 6 periodos escolares (2009 - 2014) y con el uso de la herramienta Visual - FIR para identificar el mejor modelo que tiene la predicción del aspirante que se inscribirá y determinando las variables que más influen en su decisión y en la predicción buscada, se realizaron numerosos experimentos, pero sólo se describirá el que arrrojó mejores resultados en términos de predicción.

Tomando únicamente a 5525 aspirantes a ingreso a la UTXJ, de los cuales tenemos todos los datos por variable, se generaron dos grupos para llevar a cabo un entrenamiento (Training con 3948) y una verificación (Test con 1577) con Visual - FIR, la información obtenida se cita a continuación en el cuadro 4.21, del que se determina que la mejor calidad corresponde a la máscara de complejidad 3:

Complejidad	Máscara	Calidad q
2	[0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 1]	0.5736
3	[0 -1 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 1]	0.5786
4	[0 0 0 0 0 0 0 0 0 -1 0 0 0 -2 -3 0 0 0 0 0 1]	0.5592
5	[0 0 0 0 0 -1 0 -2 0 -3 0 0 0 0 0 -4 0 0 0 0 1]	0.5687
6	[0 0 0 0 0 -1 -2 0 0 0 0 0 -3 0 0 -4 0 0 -5 0 1]	0.5643
7	[0 0 0 0 0 -1 -2 0 0 0 0 0 -3 -4 0 0 0 -5 0 -6 1]	0.4989

Cuadro 4.21: Obtención de la máscara en referencia a su calidad.

Así mismo, se procedió a la etapa de predicción de inscripción de los aspirantes (incluido en el segundo conjunto llamado Test) mediante el motor de FIR y se obtuvieron los siguientes resultados mostrados en el cuadro 4.22, en la cual se observa que las máscaras con menor complejidad son las que presentan menor error cuadrático medio RMS:

Complejidad	Calidad q	Error RMS
2	0.5736	0.4663
3	0.5786	0.4724
4	0.5592	0.4704
5	0.5687	0.5008
6	0.5643	0.5084
7	0.4989	0.5043

Cuadro 4.22: Obtención de la máscara en referencia a su calidad.

Derivado de la tabla 4.22, se determinó considerar a la mácara con complejidad 3 como la óptima por tener la mejor calidad de todas las máscaras además de tener un índice de error por debajo de 0.5, lo que la pone en ventaja de 3 máscaras, por lo que a continuación se muestran las variables causales a la de salida según la máscara obtenida:

- Turno o modalidad de estudios
- Periodo de solicitud de ingreso

La figura 4.1 muestra la predicción generada con la mácaa de compledidad 3 con un error RMS = 0.4724, donde el trazo de los datos reales se muestra con aquellas líneas contínuas verticales, mientras el obtenido con la metodología de predicción se muestra con líneas punteadas verticales, los cuales hacen notar un acoplamiento con mayor acierto en la predicción del valor cercano a 2, es decir, de que alumno "si" se inscribirá.

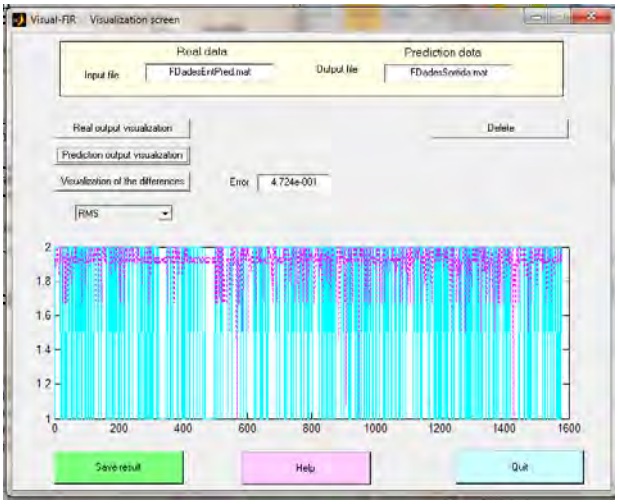


FIGURA 4.1: Predicción obtenida por FIR para el índice de inscripción, máscara de complejidad 3 y profundidad 1

4.2.4. Modelos sub-óptimos

Se realizaron experimentos adicionales para encontrar modelos distintos (sub-óptimos) y observar su comportamiento, los cuales se describen a continuación:

Experimento A: Para dicho ejercicio se seleccionó la máscara de complejidad 5 y profundidad 1, cuya calidad q:0.5687, siendo la segunda mejor:

Las variables identificadas de la máscara obtenida son:

- Trabaja
- Periodo solicitud de ingreso
- Estado civil
- Calificación Habilidad Verbal

Experimento B: En este experimento se generó una máscara de profundidad 6 y profundidad 1. Las variables identificadas de la máscara obtenida son:

- Trabaja

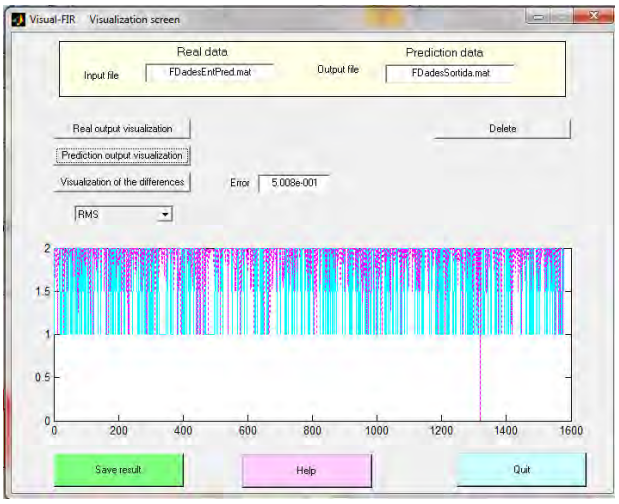


FIGURA 4.2: Predicción obtenida por FIR para el índice de inscripción, máscara de complejidad 5 y profundidad 1

- Promedio Bachillerato
- Estado del alumno
- calificación habilidad verbal
- Tiempo examen habilidad verbal

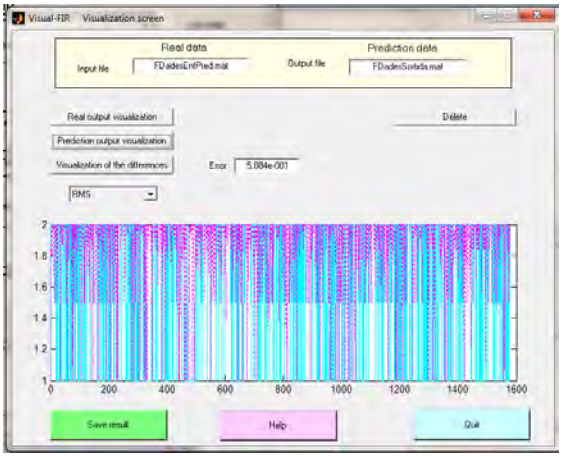


FIGURA 4.3: Predicción obtenida por FIR para el índice de inscripción, máscara de complejidad 6 y profundidad 1

El cuadro 4.23 muestra algunos datos que son comparados entre los reales y los predichos por FIR obtenidos con las diferentes máscaras y complejidad, los cuales se obtienen después de realizar la predicción, regeneración y visualización, y darle click al boton salvar, dentro de la plataforma Visual - FIR.

Dato	Máscara de Complejidad 3		Máscara de Complejidad 5		Máscara de Complejidad 6	
Real	Predicción	Diferencia	Predicción	Diferencia	Predicción	Diferencia
1	1.9259	-0.9259	1.9461	-0.9461	1	0
2	1.9103	0.0897	2	0	1.8333	0.1667
2	1.9324	0.0676	2	0	1.7895	0.2105
2	1.932	0.068	2	0	2	0
2	1.875	0.125	1.6422	0.3578	1.8393	0.1607
2	2	0	2	0	1.8047	0.1953
2	1.937	0.063	2	0	2	0
2	1.914	0.086	2	0	2	0
2	1.9087	0.0913	2	0	2	0
2	2	0	2	0	1.9594	0.0406
2	1.9259	0.0741	1	1	2	0
2	1.9259	0.0741	1.8337	0.1663	2	0
2	1.937	0.063	2	0	1.8333	0.1667
2	1.8333	0.1667	1.8726	0.1274	1.5	0.5
2	1.8571	0.1429	1.7987	0.2013	2	0
2	1.937	0.063	2	0	2	0
2	2	0	2	0	2	0
2	1.937	0.063	2	0	2	0
1	1.937	-0.937	2	-1	1.8333	-0.8333
2	1.9087	0.0913	1.88	0.12	1.947	0.053
2	2	0	1.9888	0.0112	2	0
1	1.937	-0.937	2	-1	1.5	-0.5
1	1.937	-0.937	2	-1	2	-1
1	1.937	-0.937	2	-1	2	-1
1	1.6745	-0.6745	2	-1	1.7208	-0.7208
2	1.9087	0.0913	2	0	2	0
2	2	0	2	0	2	0
2	1.937	0.063	1.9433	0.0567	2	0
2	1.9087	0.0913	2	0	2	0
1	1.937	-0.937	1	0	2	-1
1	1.937	-0.937	2	-1	2	-1
1	1.937	-0.937	2	-1	2	-1
2	1.6745	0.3255	2	0	2	0
2	1.937	0.063	1.6667	0.3333	1	1
1	1.9087	-0.9087	2	-1	2	-1
1	1.937	-0.937	2	-1	2	-1
1	1.6745	-0.6745	1.9016	-0.9016	1.9098	-0.9098
1	1.9087	-0.9087	2	-1	2	-1
2	1.9087	0.0913	1.9874	0.0126	2	0
2	2	0	2	0	1.8676	0.1324
2	1.6745	0.3255	1.5	0.5	2	0
2	1.6745	0.3255	1.7588	0.2412	2	0
2	1.971	0.029	2	0	1.8749	0.1251
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Cuadro 4.23: Datos reales, predichos y su diferencia según la máscara y su complejidad.

4.2.5. Reglas obtenidas con el algoritmo LR - FIR para el experimento de predicción de la inscripción

Para describir el comportamiento del sistema, se realizaron diversos experimentos con el modelo FIR identificado anteriormente, utilizando el algoritmo LR-FIR (Implementado en VisualFIR). En estos experimentos se una profundidad 1 y los datos deconsideró una profundidad 1y los demás datos de configuración se observan en el cuadro 4.24:

Experimento 1	Complejidad de la Máscara	Outliers	Filtro	Otherwise	Compactación mejorada
1	3	Si (3 %)	0.05	Si	Minimal Ratio
2	5	Si (3 %)	0.05	Si	Minimal Ratio
3	6	Si (3 %)	0.1	Si	Minimal Ratio

Cuadro 4.24: Valores de configuración de los experimentos.

Experimento 1, Máscara de Complejidad 3, Predicción de la Inscripción
Para el experimento de la máscara con complejidad 3 se obtuvieron las presentes

reglas del cuadro 4.25, donde se eliminaron outliers al 3 %, los cuales pueden corresponden a datos que pueden sesgar el proceso del modelado, y es posible en LR - FIR detectar y eliminar los outliers antes de iniciar el proceso de extracción de reglas. Las reglas obtenidas nos indican que los aspirantes, se inscribirán cuando provengan de áreas de TIC, Económico Administrativas o del área Agroalimentaria Biotecnológica.

Las presentes reglas concuerdan con el comportamiento a la hora en que los aspirantes se inscriben en cada ciclo escolar, según se ha corroborado con directivos de la UTXJ, es decir, que la inscripción depende de la vocación específica del Bachillerato en donde se formó al alumno y para el caso de estudio, tenemos que hay una gran cantidad de Bachilleratos que proporcionan los perfiles mostrados en las reglas.

En términos generales, con cualquier experiento de predicción realizado en este apartado, se observa en las reglas que se está caracterizando a partir de valores que consideran la especialidad del Bachillerato, lo cual promueve a la UTXJ a vincular con mayor énfasis a aquellos bachilleratos que tengan especialidades similares a las vigentes.

Reglas Training	Espec.	Sens.
IF Especialidad Escuela IN (Economico Administrativa) THEN Se inscribe	0.51	0.5
IF Especialidad Escuela IN (TIC) THEN Se inscribe	0.69	0.29
IF Especialidad Escuela IN (Agroalimentaria Biotecnológica) THEN Se inscribe	0.96	0.13
JOINT QUALITY OF THE ACTUAL CLASS	0.53	0.51
Otherwise: No se Inscribe		
Reglas Test	Espec.	Sens.
IF Especialidad Escuela IN (Economico Administrativa) THEN Se inscribe	0.63	0.4
IF Especialidad Escuela IN (TIC) THEN Se inscribe	0.68	0.27
IF Especialidad Escuela IN (Agroalimentaria Biotecnológica) THEN Se inscribe	0.92	0.12
JOINT QUALITY OF THE ACTUAL CLASS	0.54	0.47
Otherwise: No se Inscribe		

Cuadro 4.25: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

Experimento 2, Máscara de Complejidad 5, Predicción de la Inscripción
En el experimento de la máscara con complejidad 5 se obtuvieron las presentes reglas del cuadro 4.26, las cuales nos indican que si la calificación de Habilidad Matemática e Inglés son altas y además se apuntan al área Electromecánica Industrial y si el promedio de bachillerato es alto, entonces se inscribe el aspirante.

Así mismo notamos que tiene un papel importante la aplicación del examen Institucional de admisión, porque suelen referirse comunmente las reglas con la variable a predecir, que en este caso es la inscripción ya sea por los datos de calificación o de tiempo de ejecución de cada examen.

Reglas Training	Espec.	Sens.
IF Cal. Hab. Matemática IN (5.24 - 10) AND Cal. Hab. Inglés IN (4.3 - 9.32) AND division carrera MI AND promedio bachiller IN (8.4 - 10) THEN Se inscribe	0.74	1
JOINT QUALITY OF THE ACTUAL CLASS	0.74	1
Otherwise: No se Inscribe		
Reglas Test	Espec.	Sens.
IF Cal. Hab. Matemática IN (5.24 - 10) AND Cal. Hab. Inglés IN (4.3 - 9.32) AND division carrera MI AND promedio bachiller IN (8.4 - 10) THEN Se inscribe	0.99	0.024
JOINT QUALITY OF THE ACTUAL CLASS	0.99	0.024
Otherwise: No se Inscribe		

Cuadro 4.26: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

Experimento 3, Máscara de Complejidad 6, Predicción de la Inscripción Para el experimento de la máscara con complejidad 6, cuadro 4.27, las reglas obtenidas nos indican que si la calificación de Habilidad Matemática y Verbal son altas y además es mujer con promedio alto de Bachillerato y agotó la mayoría del tiempo del examen de Inglés, entonces se inscribe.

En la UTXJ se tiene un universo balanceado por género casi en proporción de 50 a 50 porciento, además hay comunidades en donde todavía existe el precedente de que las mujeres no se deben de conducirse al estudio superior, sin embargo, lo que obtenemos de los resultados del experimento, es que el género femenino asegura su inscripción cuando se dedica desde la educación media superior.

Reglas Training	Espec.	Sens.
IF Cal. Hab. Matemática IN (5.24 - 10) AND Cal. Hab. Verbal IN (5.24 - 10) AND Mujer AND promedio bachiller IN (8.4 - 10) AND Tiempo examen Hab. Inglés IN (49 - 99) THEN Se inscribe	1	1
JOINT QUALITY OF THE ACTUAL CLASS	1	1
Otherwise: No se Inscribe		
Reglas Test	Espec.	Sens.
IF Cal. Hab. Matemática IN (5.24 - 10) AND Cal. Hab. Verbal IN (5.24 - 10) AND Mujer AND promedio bachiller IN (8.4 - 10) AND Tiempo examen Hab. Inglés IN (49 - 99) THEN Se inscribe	1	0.12
JOINT QUALITY OF THE ACTUAL CLASS	1	0.12
Otherwise: No se Inscribe		

Cuadro 4.27: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

4.3. Predicción de la conclusión de estudios

4.3.1. Identificación del modelo

Para el experimento de predecir la conclusión de una carrera, se tomaron en cuenta varios factores, puesto que el modelo de estudios de TSU requiere dos años de estudio, se consideró realizar el experimento con las generaciones 2009 - 2012, porque son aquellas que tienen egresados al momento de realización del presente proyecto.

Considerando únicamente a 1860 aspirantes a ingreso a la UTXJ, de los cuales tenemos todos los datos por variable, se generaron dos grupos para llevar a cabo un entrenamiento (Training con 1240) y una verificación (Test con 620) con Visual - FIR.

Para el primer experimento se tomaron en cuenta las variables:

- División carrera
- Estado escuela
- Especialidad Escuela
- Mes registro
- Trabaja
- Promedio bachiller
- Periodo solicitud ingreso
- Medios difusión
- Estado civil
- Edad registro
- Género
- Estado alumno
- Promedio
- Terminó

En el cuadro 4.28 se listan las máscaras generadas, considerando que la mejor calidad corresponde a la máscara de complejidad 3:

Complejidad	Máscara	Calidad q
2	[0 0 0 0 0 -1 0 0 0 0 0 0 0 1]	0.385
3	[0 0 0 0 0 -1 -2 0 0 0 0 0 0 1]	0.3927
4	[0 0 0 0 0 -1 -2 0 0 -3 0 0 0 1]	0.3891
5	[0 0 0 0 0 -1 -2 0 0 -3 -4 0 0 1]	0.364
6	[0 0 0 0 0 -1 -2 0 0 -3 -4 0 -5 1]	0.3351
7	[0 0 -1 0 0 -2 -3 0 0 -4 0 -5 -6 1]	0.3044

Cuadro 4.28: Obtención de la máscara en referencia a su calidad.

Así mismo, se procedió a la etapa de predicción de conclusión de estudios de los estudiantes (incluido en el segundo conjunto llamado Test) mediante el motor de FIR y se obtuvieron los siguientes resultados mostrados en el cuadro 4.29, en la cual se observa que las máscaras con menor complejidad son las que presentan menor error cuadrático medio RMS:

Después de observar la tabla 4.28, se determinó considerar a la mácara con complejidad 3 como la óptima por tener la mejor calidad de todas las máscaras además de tener el segundo mejor índice de error, lo que la pone en ventaja de 3 máscaras, por lo que a continuación se muestran las variables causales a la de salida según la máscara obtenida:

Complejidad	Calidad q	Error RMS
2	0.385	0.6094
3	0.3927	0.6114
4	0.3891	0.6116
5	0.364	0.6208
6	0.3351	0.6309
7	0.3044	0.7218

CUADRO 4.29: Comparación de la calidad y el error entre las diferentes máscaras obtenidas.

- Promedio de Bachillerato
- Periodo de solicitud de ingreso

La figura 4.4 muestra la predicción obtenida con la máscara de complejidad 3 con un error RMS = 0.6114 y se observa de manera ligera que la tendencia de predicción se dirige más a ubicar a aquellos estudiantes que no concluirán sus estudios.

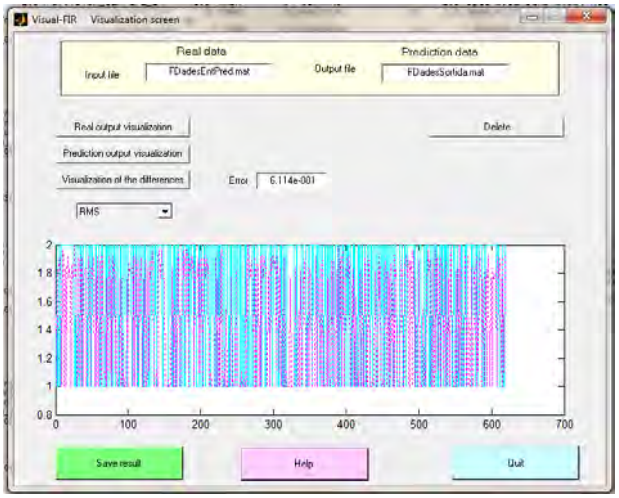


FIGURA 4.4: Predicción obtenida por FIR para la conclusión de estudios, máscara de complejidad 3 y profundidad 1

4.3.2. Modelos sub-óptimos

Se realizaron experimentos adicionales para encontrar modelos distintos (sub-óptimos) y observar su comportamiento, los cuales se describen a continuación:

Experimento A: Para dicho ejercicio se seleccionó la máscara de complejidad 4 y profundidad 1, cuya calidad q: 0.3891 como se muestra en la figura 4.5, siendo la segunda mejor:

Las variables identificadas de la máscara obtenida son:

- Promedio de Bachillerato
- Periodo de solicitud de ingreso
- Edad del aspirante

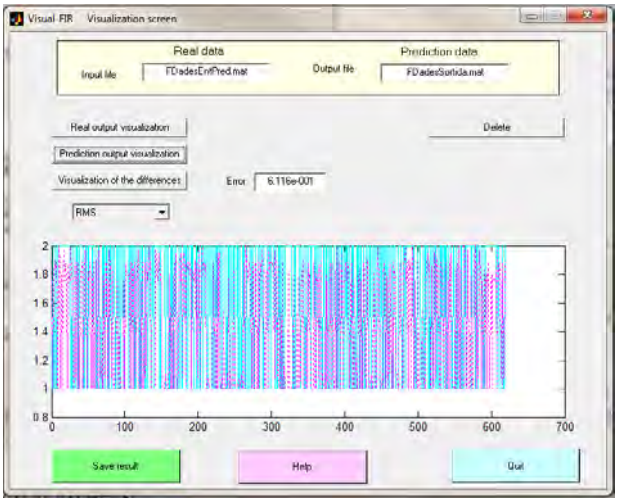


FIGURA 4.5: Predicción obtenida por FIR para la conclusión de estudios, máscara de complejidad 4 y profundidad 1

Experimento B: En este experimento se generó una máscara de profundidad 5 y profundidad 1, cuya calidad q: 0.3640 como se muestra en la figura 4.6.

Las variables identificadas de la máscara obtenida son:

- Promedio de Bachillerato
- Periodo de solicitud de ingreso
- Edad del aspirante
- Género

El cuadro 4.30 muestra algunos datos que son comparados entre los reales y los predichos por FIR obtenidos con las diferentes máscaras y complejidad, los cuales se obtienen después de realizar la predicción, regeneración y visualización dentro de la plataforma Visual - FIR, en cada columna se observa como la diferencia es notoria entre máscara y máscara.

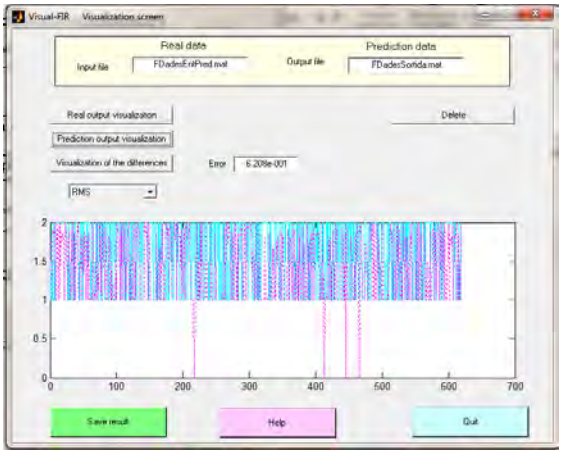


FIGURA 4.6: Predicción obtenida por FIR para la conclusión de estudios, máscara de complejidad 5 y profundidad 1

Dato	Máscara de Complejidad 3		Máscara de Complejidad 4		Máscara de Complejidad 5	
Real	Predicción	Diferencia	Predicción	Diferencia	Predicción	Diferencia
1	1.7622	-0.7622	1.7544	-0.7544	1.7607	-0.7607
2	1	1	1	1	1	1
1	1.814	-0.814	1.8182	-0.8182	1.6111	-0.6111
1	1.8868	-0.8868	1.8947	-0.8947	1.9048	-0.9048
2	1.8175	0.1825	1.828	0.172	1.8596	0.1404
1	1.9149	-0.9149	1.881	-0.881	1.8596	-0.8596
1	1.6111	-0.6111	1.5758	-0.5758	2	-1
2	1	1	1	1	1	1
2	1	1	1	1	1	1
2	1.9583	0.0417	2	0	2	0
2	1.9149	0.0851	1.881	0.119	1.8667	0.1333
2	1	1	1	1	1	1
2	1.8384	0.1616	1.7917	0.2083	1.8889	0.1111
2	1.6111	0.3889	1.5758	0.4242	2	0
2	1	1	1	1	1	1
2	1.814	0.186	1.8182	0.1818	1.9048	0.0952
2	1.9149	0.0851	1.9649	0.0351	1.8333	0.1667
2	1.7622	0.2378	1.7544	0.2456	1.7607	0.2393
2	1.8384	0.1616	1.8824	0.1176	1.8519	0.1481
2	1.7622	0.2378	1.7544	0.2456	1.7926	0.2074
2	1	1	1	1	1	1
2	1	1	1	1	1	1
2	1.8611	0.1389	1.8667	0.1333	2	0
2	1.7622	0.2378	1.7544	0.2456	1.8133	0.1867
2	1.7622	0.2378	1.7544	0.2456	1.7926	0.2074
1	2	-1	2	-1	1	0
1	1.7622	-0.7622	1.798	-0.798	1.8095	-0.8095
2	1.8384	0.1616	1.7917	0.2083	1.8889	0.1111
1	1.7622	-0.7622	1.7544	-0.7544	1.6667	-0.6667
1	1	0	1	0	1	0
1	1.101	-0.101	1.075	-0.075	1.0882	-0.0882
2	1.9149	0.0851	1.881	0.119	1.9048	0.0952
1	1.101	-0.101	1.075	-0.075	1.0882	-0.0882
2	1.9259	0.0741	1.8824	0.1176	2	0
2	1.7622	0.2378	1.7544	0.2456	1.7143	0.2857
1	1	0	1	0	1	0
1	1.9149	-0.9149	1.881	-0.881	1.8667	-0.8667
2	1	1	1	1	1	1
2	1.7622	0.2378	1.7544	0.2456	1.7607	0.2393
2	1.8175	0.1825	1.7879	0.2121	1.7143	0.2857
2	1.8384	0.1616	1.8824	0.1176	2	0
2	1.8384	0.1616	1.8824	0.1176	2	0
2	1	1	1	1	1	1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

CUADRO 4.30: Datos reales, predichos y su diferencia según la máscara y su complejidad.

4.3.3. Reglas obtenidas con el algoritmo LR - FIR para el experimento de predicción de la conclusión de estudios

Para describir el comportamiento del conjunto de datos, se realizaron diversos experimentos con Visual - FIR, los cuales utilizan información que ha sido utilizada por el algoritmo LR-FIR. De entre estos experimentos se describe las pruebas con las tres máscaras anteriores de las subsecciones 4.3.1 y 4.3.2, y que a continuación en el cuadro 4.31 se aporta la configuración utilizada para cada regla:

Experimento 1	Complejidad de la Máscara	Outliers	Filtro	Otherwise	Compactación mejorada
1	3	Si (2 %)	0.05	Si	Only Avoid Conflicts
2	4	Si (3 %)	0.05	Si	Minimal Ratio
3	5	Si (3 %)	0.05	Si	Only Avoid Conflicts

Cuadro 4.31: Valores de configuración de los experimentos.

Experimento 1, Máscara de Complejidad 3, Predicción de la conclusión de estudios

Para el experimento de la máscara con complejidad 3 se obtuvieron las presentes reglas del cuadro 4.32, donde se eliminaron outliers al 2 %, las reglas obtenidas nos indican que los estudiantes, siempre y cuando conserven un promedio entre 7 y 10 concluira sus estudios, además de que provengan de áreas de TIC, Económico Administrativas o del área Agroalimentaria Biotecnológica y Electromecánica Industrial.

Reglas Training	Espec.	Sens.
IF Promedio general IN (7 - 10) AND división carrera IN (Bio, Admon, TIC, MI) THEN Si Terminará	0.62	1
JOINT QUALITY OF THE ACTUAL CLASS	0.62	1
Otherwise: No concluye		
Reglas Test	Espec.	Sens.
IF Promedio general IN (7 - 10) AND división carrera IN (Bio, Admon, TIC, MI) THEN Si Terminará	0.64	0.56
JOINT QUALITY OF THE ACTUAL CLASS	0.64	0.56
Otherwise: No concluye		

Cuadro 4.32: Reglas obtenidas eliminando outliers al 2 % aplicando filtro a 0.05 con uso de Otherwise.

Experimento 2, Máscara de Complejidad 4, Predicción de la conclusión de estudios

En el experimento de la máscara con complejidad 4 se obtuvieron las presentes reglas del cuadro 4.33, las cuales nos indican que existen cuatro áreas académicas donde será más probable que un estudiante concluya.

Reglas Training	Espec.	Sens.
IF Promedio general IN(7 - 10) AND division carrera IN (Bio, Admon, TIC, MI) THEN Si Terminará	0.72	1
JOINT QUALITY OF THE ACTUAL CLASS	0.72	1
Otherwise: No concluye	0.68	0.78
Reglas Test	Espec.	Sens.
IF Promedio general IN(7 - 10) AND division carrera IN (Bio, Admon, TIC, MI) THEN Si Terminará	0.64	0.56
JOINT QUALITY OF THE ACTUAL CLASS	0.64	0.56
Otherwise: No concluye	0.45	0.7

Cuadro 4.33: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

Experimento 3, Máscara de Complejidad 5, Predicción de la conclusión de estudios

Para el experimento de la máscara con complejidad 5, cuadro 4.34, las reglas obtenidas nos indican situaciones obvias como cuando el promedio general está de regular a alto, entonces conluyen los estudios, así mismo, la selección del área académica influyen para la conclusión de estudios, y por último, que influye mucho que el estudiante haya tomado a la universidad desde un principio, como una institución de primera opción para realizar estudios superiores. Estos resultados han sido revisados y considerados por directivos de la UTXJ, quienes concuerdan con dichos valores y tendencias.

Reglas Training	Espec.	Sens.
IF Promedio general IN (7 -10) THEN Si Terminará	0.73	1
IF Promedio general IN (8 - 10) AND division carrera IN PAI THEN Si Terminará	1	0.13
IF mes registro IN (marzo - mayo) THEN Si Terminará	1	0.075
JOINT QUALITY OF THE ACTUAL CLASS	0.73	1
Otherwise No termina	0.69	0.79
Reglas Test	Espec.	Sens.
IF Promedio general IN (7 -10) THEN Si Terminará	0.63	0.58
IF Promedio general IN (8 - 10) AND division carrera IN PAI THEN Si Terminará	0.96	0.088
IF mes registro IN (marzo - mayo) THEN Si Terminará	0.85	0.22
JOINT QUALITY OF THE ACTUAL CLASS	0.56	0.69
Otherwise No termina	0.52	0.64

Cuadro 4.34: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

4.4. Predicción del Promedio Académico

4.4.1. Identificación del modelo

Para el experimento de predecir la conclusión de una carrera, se tomaron en cuenta varios factores, puesto que el modelo de estudios de TSU requiere dos años de estudio, se consideró realizar el experimento con todas las generaciones 2009 - 2013, porque son aquellas que tienen un promedio alcanzado al momento.

Considerando únicamente a 4196 estudiantes de la UTXJ, de los cuales tenemos todos los datos por variable, se generaron dos grupos para llevar a cabo un entrenamiento (Training con 2798) y una verificación (Test con 1398) con Visual - FIR.

Para el primer experimento se tomaron en cuenta las variables:

- División carrera
- Turno
- Estado escuela
- Especialidad Escuela
- Mes registro
- Trabaja
- Promedio bachiller
- Periodo solicitud ingreso
- Medios difusión
- Estado civil
- Edad registro
- Género
- Estado alumno
- Mes aplicación examen de Admisión
- Calificación Habilidad Matemática
- Calificación Habilidad Verbal
- Calificación Habilidad Inglés
- Tiempo de ejecución de Habilidad Matemática
- Tiempo de ejecución de Habilidad Verbal
- Tiempo de ejecución de Habilidad Inglés
- Terminó
- Promedio

En el cuadro 4.35 se listan las máscaras generadas, considerando que la mejor calidad corresponde a la máscara de complejidad 5.

Así mismo, se procedió a la etapa de predicción del promedio académico de los estudiantes (incluido en el segundo conjunto llamado Test) mediante el motor de FIR se obtuvieron los siguientes resultados mostrados en el cuadro 4.36, en la cual se observa que las máscaras con mayor complejidad son las que presentan menor error cuadrático medio RMS.

Complejidad	Máscara	Calidad q
2	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 1]	0.4793
3	[0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 -2 0 0 0 0 1]	0.4991
4	[0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 -2 0 0 0 -3 1]	0.5032
5	[-1 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 -3 0 0 0 -4 1]	0.5032
6	[-1 0 0 0 0 0 0 0 0 -2 0 0 0 0 -3 0 -4 0 0 0 -5 1]	0.5024
7	[-1 0 0 0 0 0 0 0 0 -2 0 0 -3 0 -4 -5 -6 0 0 0 0 1]	0.4899

Cuadro 4.35: Obtención de la máscara en referencia a su calidad.

Complejidad	Calidad q	Error RMS
2	0.4793	263
3	0.4991	5.935
4	0.5032	5.935
5	0.5032	5.866
6	0.5024	5.847
7	0.4899	5.8000

Cuadro 4.36: Comparación de la calidad y el error entre las diferentes máscaras obtenidas.

Después de observar la tabla 4.36, se determinó considerar a la mácara con complejidad 5 como la óptima por tener la mejor calidad de todas las máscaras además de tener el tercer mejor índice de error, lo que la pone en ventaja de 3 máscaras, por lo que a continuación se muestran las variables causales a la de salida según la máscara obtenida:

- Carrera
- Estado Civil
- Calificación de Habilidad en Inglés
- Conclusión de la carrera

La figura 4.7 muestra la predicción obtenida con la máscara de complejidad 5 con un error RMS = 5.866 y se observa que la predicción está orientada a los valores menores y cercanos a 8.

4.4.2. Modelos sub-óptimos

Se realizaron experimentos adicionales para encontrar modelos distintos (sub-óptimos) y observar su comportamiento, los cuales se describen a continuación:

Experimento A: Para dicho ejercicio se seleccionó la máscara de complejidad 6 y profundidad 1, cuya calidad q: 0.5024 como se muestra en la figura 4.8, siendo la segunda mejor:

Las variables identificadas de la máscara obtenida son:

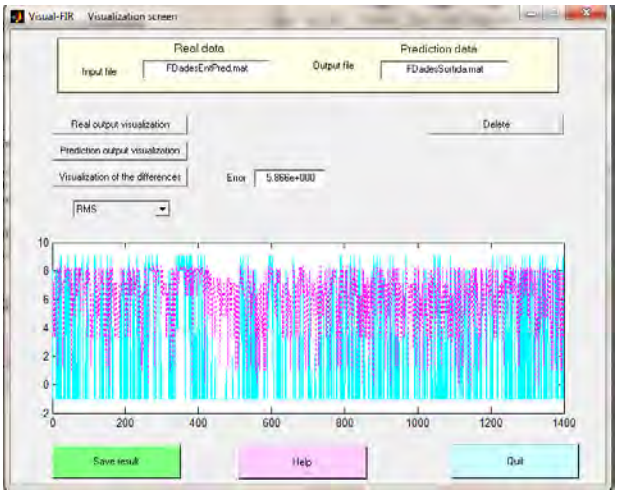


FIGURA 4.7: Predicción obtenida por FIR para el promedio académico, máscara de complejidad 5 y profundidad 1

- Carrera
- Estado Civil
- Calificación de Habilidad Matemática
- Calificación de Habilidad en Inglés
- Conclusión de la carrera

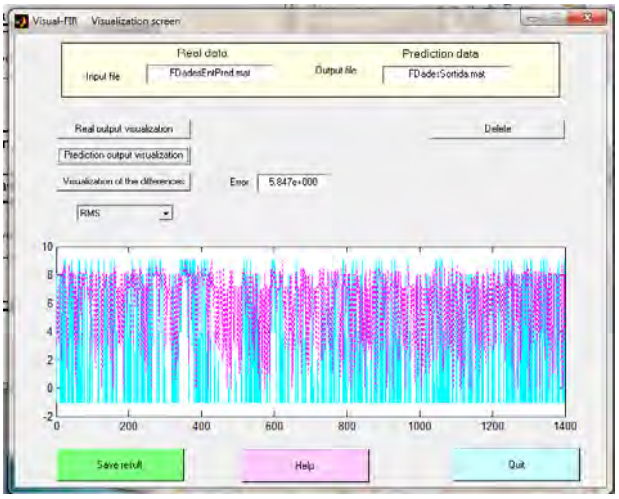


FIGURA 4.8: Predicción obtenida por FIR para el promedio académico, máscara de complejidad 6 y profundidad 1

Experimento B: En este experimento se generó una máscara de profundidad 7 y profundidad 1, cuya calidad q: 0.4899 como se muestra en la figura 4.9.

Las variables identificadas de la máscara obtenida son:

- Carrera

- Estado Civil
- Calificación de Habilidad Matemática
- Calificación de Habilidad en Inglés
- Conclusión de la carrera

FIGURA 4.9: Predicción obtenida por FIR para el promedio académico, máscara de complejidad 7 y profundidad 1

El cuadro 4.37 muestra algunos datos que son comparados entre los reales y los predichos por FIR obtenidos con las diferentes máscaras y complejidad, los cuales se obtienen después de realizar la predicción, regeneración y visualización dentro de la plataforma Visual - FIR, en cada columna se observa como la diferencia es notoria entre máscara y máscara.

Dato	Máscara de Complejidad 3		Máscara de Complejidad 4		Máscara de Complejidad 5	
Real	Predicción	Diferencia	Predicción	Diferencia	Predicción	Diferencia
8	7.2255	0.7745	5.9602	2.0398	7.074	0.926
-1	3.3158	-4.3158	3.125	-4.125	0	-1
-1	4	-5	4	-5	3.1508	-4.1508
-1	7.1057	-8.1057	7.1211	-8.1211	3	-4
-1	4.7596	-5.7596	4.7596	-5.7596	1.3333	-2.3333
-1	3.3158	-4.3158	3.3158	-4.3158	6.2668	-7.2668
-1	7.3793	-8.3793	7.1423	-8.1423	7.5	-8.5
-1	3.3158	-4.3158	3.125	-4.125	0.65211	-1.65211
-1	8	-9	7.3593	-8.3593	3.377	-4.377
-1	8	-9	8	-9	7.8007	-8.8007
8	8	0	8	0	4.6132	3.3868
8	8.1984	-0.1984	8.1984	-0.1984	7.8726	0.1274
-1	8	-9	8	-9	7.9258	-8.9258
-1	8	-9	8	-9	5.2939	-6.2939
-1	4.91	-5.91	2.6875	-3.6875	2.5585	-3.5585
-1	1.3333	-2.3333	2	-3	7.0491	-8.0491
8	8.0983	-0.0983	8.2308	-0.2308	8	0
-1	8.0983	-9.0983	7.9	-8.9	7.6746	-8.6746
8	8.029	-0.029	8.0526	-0.0526	8	0
-1	8.2791	-9.2791	8.6667	-9.6667	8.3884	-9.3884
-1	6.3906	-7.3906	5.8	-6.8	5.93	-6.93
-1	6.2453	-7.2453	5.3571	-6.3571	2.0168	-3.0168
9	8.2791	0.7209	8.6667	0.3333	8.5	0.5
9	8.0983	0.9017	8.1538	0.8462	8	1
-1	8.029	-9.029	8.2105	-9.2105	8.2428	-9.2428
-1	6.3906	-7.3906	5.8	-6.8	7.2343	-8.2343
-1	7.1057	-8.1057	7.1211	-8.1211	3.5	-4.5
-1	4.91	-5.91	1.5	-2.5	0.43938	-1.43938
-1	1.3333	-2.3333	5	-6	7.6681	-8.6681
-1	3.3158	-4.3158	2.4	-3.4	7.7638	-8.7638
-1	3.3158	-4.3158	3.3158	-4.3158	3.4682	-4.4682
-1	6.3906	-7.3906	4.4737	-5.4737	3	-4
8	8.2791	-0.2791	8.6667	-0.6667	9	-1
-1	7.5286	-8.5286	8.3333	-9.3333	3.7004	-4.7004
7	8.2791	-1.2791	8.6667	-1.6667	9	-2
-1	7.1057	-8.1057	7.1786	-8.1786	0	-1
-1	7.1057	-8.1057	6.25	-7.25	7.3426	-8.3426
8	8.0983	-0.0983	8.1538	-0.1538	9	-1
8	8.029	-0.029	7.9474	0.0526	8.6653	-0.6653
-1	8.0983	-9.0983	8.1538	-9.1538	8	-9
8	8.029	-0.029	7.3333	0.6667	8.75	-0.75
9	8.2791	0.7209	8.1724	0.8276	8.3333	0.6667
-1	6.2698	-7.2698	7	-8	3.1281	-4.1281
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

CUADRO 4.37: Datos reales, predichos y su diferencia según la máscara y su complejidad.

4.4.3. Reglas obtenidas con el algoritmo LR - FIR para el experimento de predicción del promedio académico

Para describir el comportamiento del conjunto de datos, se realizaron diversos experimentos con Visual - FIR, los cuales utilizan información que ha sido utilizada por el algoritmo LR-FIR. De entre estos experimentos se describe las pruebas con las tres máscaras anteriores de las subsecciones 4.4.1 y 4.4.2, y que a continuación en el cuadro 4.38 se aporta la configuración utilizada para cada regla:

Experimento 1	Complejidad de la Máscara	Outliers	Filtro	Otherwise	Compactación mejorada
1	5	Si (2 %)	0.05	Si	Only Avoid Conflicts
2	6	Si (3 %)	0.05	Si	Minimal Ratio
3	7	Si (3 %)	0.05	Si	Only Avoid Conflicts

Cuadro 4.38: Valores de configuración de los experimentos.

Experimento 1, Máscara de Complejidad 5, Predicción del promedio académico

Para el experimento de la máscara con complejidad 5 se obtuvieron las presentes reglas del cuadro 4.39, donde se eliminaron outliers al 3 %, las reglas obtenidas nos indican que los estudiantes de procedencia lejana tienden a tener un promedio bajo, sin embargo, en algunas carreras su promedio será regular siempre y cuando no trabajen.

Así mismo se observa una tendencia a tener buenos promedios en algunas carreras como PAI, Admon, MI y TIC, en ésta última siempre que la procedencia del alumno sea lejana y no trabaje.

Reglas Training	Espec.	Sens.
IF Edo Esc proc IN (Lejos de Puebla) AND No terminó AND No trabaja THEN Promedio general IN (0 - 7)	0.91	0.38
JOINT QUALITY OF THE ACTUAL CLASS	0.91	0.38
IF Edo Esc proc IN (Lejos de Puebla) AND division carrera 1 IN(PAI, Admon, MI) AND No trabaja THEN Promedio general IN (7 - 8)	0.59	0.42
JOINT QUALITY OF THE ACTUAL CLASS	0.59	0.42
IF division carrera 1 IN (PAI, Admon, MI) THEN Promedio general IN (8 - 10)	0.27	0.87
IF Edo Esc proc IN (Lejos de Puebla) AND division carrera 1 IN (TIC) AND terminó carrera AND No trabaja THEN Promedio general IN (8- 10)	0.97	0.13
JOINT QUALITY OF THE ACTUAL CLASS	0.69	0.46
Otherwise: El promedio (7 - 8)	0.62	0.43
Reglas Test	Espec.	Sens.
IF Edo Esc proc IN (Lejos de Puebla) AND No terminó AND No trabaja THEN Promedio general IN (0 - 7)	0.88	0.31
JOINT QUALITY OF THE ACTUAL CLASS	0.88	0.31
IF Edo Esc proc IN (Lejos de Puebla) AND division carrera 1 IN(PAI, Admon, MI) AND No trabaja THEN Promedio general IN (7 - 8)	0.63	0.43
JOINT QUALITY OF THE ACTUAL CLASS	0.63	0.43
IF division carrera 1 IN (PAI, Admon, MI) THEN Promedio general IN (8 - 10)	0.81	0.42
IF Edo Esc proc IN (Lejos de Puebla) AND division carrera 1 IN (TIC) AND terminó carrera AND No trabaja THEN Promedio general IN (8- 10)	0.97	0.13
JOINT QUALITY OF THE ACTUAL CLASS	0.65	0.44
Otherwise: El promedio (7 - 8)	0.64	0.41

Cuadro 4.39: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

Experimento 2, Máscara de Complejidad 6, Predicción del promedio académico

En el experimento de la máscara con complejidad 6 se obtuvieron las presentes reglas del cuadro 4.40, en él se observa que aquellos estudiantes que provienen de lejos y que se registraron entre junio y octubre, es decir, no necesariamente la UTXJ fué su primer opción de estudio y trabajan, su promedio académico será bajo.

Así mismo se indica que aquellos estudiantes que son de lejos, que no trabajan y que se encuentran en alguna de las siguientes áreas, es muy probable que su promedio vaya del nivel regular a lo más alto, aquí se hace notar que son estudiantes que se inscribieron en los meses a partir de junio a aagsto, y lo más probable es que esten hospedados en Xicotepec y se dediquen de tiempo completo.

Por último, por lo regular aquel estudiante que no conluyó sus estudios tiene un bajo promedio.

Reglas Training	Espec.	Sens.
IF Edo esc proc IN 3 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 1 THEN Promedio general IN 1	0.75	0.89
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 2 THEN Promedio general IN 1	1	0.055
IF Edo esc proc IN 3 AND division carrera 1 IN 1 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 2 THEN Promedio general IN 1	1	0.052
JOINT QUALITY OF THE ACTUAL CLASS	0.88	0.42
IF Edo esc proc IN 3 AND division carrera 1 IN 3- 5 AND mes registro IN 2 AND trabaja IN 1 THEN Promedio general IN 2	0.21	0.93
IF Edo esc proc IN 3 AND division carrera 1 IN 2 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 1 THEN Promedio general IN 2	0.96	0.072
JOINT QUALITY OF THE ACTUAL CLASS	0.62	0.43
IF Edo esc proc IN 3 AND division carrera 1 IN 3- 5 AND mes registro IN 2 AND trabaja IN 1 THEN Promedio general IN 3	0.3	0.87
IF Edo esc proc IN 3 AND division carrera 1 IN 2 AND mes registro IN 2 AND termino IN 2 AND trabaja IN 1 THEN Promedio general IN 3	1	0.13
JOINT QUALITY OF THE ACTUAL CLASS	0.71	0.48
Otherwise: Promedio (7 - 8)	0.45	0.7
Reglas Test	Espec.	Sens.
IF Edo esc proc IN 3 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 1 THEN Promedio general IN 1	0.75	0.63
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 2 THEN Promedio general IN 1	0.99	0.12
IF Edo esc proc IN 3 AND division carrera 1 IN 1 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 2 THEN Promedio general IN 1	0.99	0.042
JOINT QUALITY OF THE ACTUAL CLASS	0.87	0.39
IF Edo esc proc IN 3 AND division carrera 1 IN 3- 5 AND mes registro IN 2 AND trabaja IN 1 THEN Promedio general IN 2	0.42	0.83
IF Edo esc proc IN 3 AND division carrera 1 IN 2 AND mes registro IN 2 AND termino IN 1 AND trabaja IN 1 THEN Promedio general IN 2	0.93	0.056
JOINT QUALITY OF THE ACTUAL CLASS	0.64	0.41
IF Edo esc proc IN 3 AND division carrera 1 IN 3- 5 AND mes registro IN 2 AND trabaja IN 1 THEN Promedio general IN 3	0.48	0.81
IF Edo esc proc IN 3 AND division carrera 1 IN 2 AND mes registro IN 2 AND termino IN 2 AND trabaja IN 1 THEN Promedio general IN 3	0.99	0.11
JOINT QUALITY OF THE ACTUAL CLASS	0.73	0.45
Otherwise: Promedio (7 - 8)	0.45	0.7

Cuadro 4.40: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

Experimento 3, Máscara de Complejidad 7, Predicción del promedio académico

En el experimento de la máscara con complejidad 7, cuadro 4.41 se observa que aquellos estudiantes que provienen de lejos y que se registraron entre junio y octubre, que son del área Electromecánica Industrial, son hombres y su promedio de bachiller es de bajo a regular, aspirarán a lo más a un promedio.

Reglas Training	Espec.	Sens.
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND termino IN 1 THEN Promedio general IN 1	0.98	0.69
JOINT QUALITY OF THE ACTUAL CLASS	0.98	0.69
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND promedio bachiller IN 1 THEN Promedio general IN 2	0.52	0.72
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND promedio bachiller IN 2 AND termino IN 2 THEN Promedio general IN 2	0.76	0.28
JOINT QUALITY OF THE ACTUAL CLASS	0.58	0.5
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND promedio bachiller IN 1- 2 AND termino IN 2 THEN Promedio general IN 3	0.9	0.5
JOINT QUALITY OF THE ACTUAL CLASS	0.9	0.5
Otherwise Promedio (7 -8)	0.58	0.5
Reglas Test	Espec.	Sens.
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND termino IN 1 THEN Promedio general IN 1	0.9	0.23
JOINT QUALITY OF THE ACTUAL CLASS	0.9	0.23
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND promedio bachiller IN 1 THEN Promedio general IN 2	0.87	0.056
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND promedio bachiller IN 2 AND termino IN 2 THEN Promedio general IN 2	0.87	0.056
JOINT QUALITY OF THE ACTUAL CLASS	0.76	0.28
IF Edo esc proc IN 3 AND division carrera 1 IN 4 AND genero IN 2 AND mes registro IN 2 AND promedio bachiller IN 1- 2 AND termino IN 2 THEN Promedio general IN 3	0.97	0.13
JOINT QUALITY OF THE ACTUAL CLASS	0.97	0.13
Otherwise Promedio (7 - 8)	0.25	0.75

Cuadro 4.41: Reglas obtenidas eliminando outliers al 3 % aplicando filtro a 0.05 con uso de Otherwise.

Las presentes reglas muestran que la cercania a la UTXJ, el género, así como el mes en que se registraron y el área académica a la que se dirijen, es importante para predecir el promedio general. Estos resultados han sido revisados y considerados por directivos de la UTXJ, quienes concuerdan con dichos valores y tendencias.

Capítulo 5

Conclusiones y Trabajo Futuro

5.1. Conclusiones

A través del presente proyecto de investigación se ha fortalecido la percepción sobre el ingreso de matrícula nueva, el rendimiento y conclusión de estudios de la matrícula existente, mediante información objetiva que sin el empleo de metodologías como FIR y el algoritmo LR -FIR sería complicado emitir un juicio utilizando información cualitativa y cuantitativa.

Se utilizaron herramientas que permitieron extraer conocimiento de tres bancos de información, base de datos de SAIIUT, admisión escolar y Aula Virtual TIC, con la intensión de retroalimentar el resultado obtenido al predecir el ingreso de matrícula, su conclusión y rendimiento académico.

De los resultados obtenidos utilizando los datos de admisión y de servicios escolares, se demostró que es posible estimar, con un porcentaje de error aceptable, la predicción del tres conceptos: promedio académico, índice de inscripción y la conclusión de estudios de aspirantes y estudiantes de la UTXJ, por esta razón los resultados obtenidos con el enfoque propuesto en este proyecto de investigación se consideran importantes.

Durante el desarrollo del proyecto, se aplicaron las 4 etapas que marca el proceso de descubrimiento del conocimiento para identificar patrones, lo cual no fué trivial, como se describe a continuación:

- Preprocesamiento
    - Selección y limpieza de datos. Se determinaron las fuentes de información:
        - Sistema de Control Escolar que opera en plataforma operativa de Windows y con SQLServer,
        - Información de admisión recabada manualmete en Control Escolar en hojas electrónicas de Cálculo y

    - Aula Virtual de Examen de Admisión, que se opera bajo la plataforma Moodle.
  - Preparación de datos. Se analizó la información y se tomó la decisión de conservar aquellos atributos descritos en el capítulo 4 para continuar con los experimentos.
  - Transformación de datos. Fueron preparados los datos en el formato .csv que Matlab requiere para cargar la información a la plataforma Visual-FIR.
- Minería de datos. Afortunadamente se utilizó la herramienta Visual - FIR, que ocupa el algoritmo de los 5 vecinos más cercanos y así encontrar los patrones de mayor interés.
- Evaluación. Fueron comparados en tablas aquellos resultados reales contra los predecidos, resultando muy interesante para cada uno de los tres experimentos cada resultado, sin embargo, uno de los importantes fué derivado de las variables de salida con formato no contínuo (discreto) a la que se pretendía llegar, pues se demostró que la metodología resuelve mejor las pedicciones si son usados datos contínuos.

Durante el desenvolvimiento del proyecto se reconoció a través de la información investigada, que la Minería de Datos Educativa ya tiene dentro de sus intereses de estudio, el rendimiento académico, la deserción y la nueva admisión matriculada, sin embargo, el uso de herramientas de vanguardia como Visual - FIR aportan la automatización del procesamiento de datos acumulados, con el que se obtienen resultados objetivos y claros al momento de describir.

Se descubrió y confirmó que hay datos importantes como el lugar de procedencia de un estudiante, la especialidad del Bachillerato de procedencia, el mes en que solicitó su ingreso a la universidad, las calificaciones del examen de admisión y el tiempo que se le dedica a su aplicación, los cuales son decisivos en los comportamientos que influyen para el rendiemiento académico, la conclusión y el ingreso de matrícula.

Recomendaciones.
Realizar experimentos con datos de admisión de aspirantes a la UTXJ e información histórica de calificaciones de alumnos, ha dado pauta para fortalecer la idea de centralizar en un sólo sistema de software, el proceso de registro de información en la admisión y la aplicación de la evaluación de la admisión, el registro de calificaciones cuatrimestrales.

Es decir, utilizar sólo una base de datos y un modelado de sistema orientado a soportar toda las solicitudes de desarrollo bajo la perspectiva de uso financiero, de planeación, de evaluación, de control escolar, de academia, pues durante el desarrollo de la selección y limpieza de datos, hubo demora en estandarizar las correspondientes bases de datos.

Trabajos futuros.
El gran volumen de información tratada en el presente proyecto conllevó a proponer un conjunto de actividades futuras que fortalezcan la línea de investigación como a continuación se describre:

- Con el fin de obtener predicción y reglas de otras herramientas se pretende comparar los primeros resultados del presente proyecto, y se propone ocupar WEKA u otra herramienta metodológica bajo el mismo objetivo.
- Serán agregados los campos de información de septiembre - diciembre 2014, para enriquecer la base de información.
- Realizar un ejercicio de predectibilidad con uso de información al corte de la futura generación de ingreso 2015.

Bibliografía

[1] Castro, F., Vellido, A., Nebot, A., Minguillón, J., 2005, Detecting a typical Student Behaviour on an e-Learning System. In: VI Congreso Nacional de Informática Educativa, Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educación, SINTICE'2005. September 14-16, Granada, Spain 153-160.

[2] Castro, F., Nebot A., Mugica F., 2007, Extraction of Logical Rules to Describe Students' Learning Behavior. The Sixth IASTED International Conference on Web-Based Education, WBE 2007. Chamonix, Francia, aceptado para publicación.

[3] Castro, F., Vellido, A., Nebot A., Mugica F., 2007, Applying Data Mining Techniques to e-Learning Problems: a Survey and State of the Art. Evolution of Technology and Pedagogy. Springer-Verlag, Germany, aceptado para publicación.

[4] Castro, F., Vellido, A., Nebot, A., Minguillón, J., 2005, Finding Relevant Features to Characterize Student Behavior on an e-Learning System. In: Hamid, R.A. (ed.): Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering, FECS'05. Las Vegas, USA, 210-216.

[5] Castro, F. and Nebot A., 2007, Un Algoritmo para la Extracción Automática de Reglas Lógicas a partir de Modelos FIR., Research Report. Dept. Llenguatges i Sistemes Informátics, (LSI). Universitat Politécnica de Catalunya. LSI-07-7-R, Pp. 1-19.

[6] Castro F., Nebot A. and Mugica F., 2009 On The Extraction of Decision Support Rules from Fuzzy Predictive Models. Enviado a International Journal of Approximate Reasoning.

[7] Celis Ochoa B., Osorno Munguía J.R., Vallejo Casarín A., Mazadiego Infante T., 2010, Predictores del rendimiento académico universitario en el examen de ingreso a la Universidad Veracruzana en ciencias de la salud, Poza Rica, Veracruz, Estudios recientes en educación superior, Una mirada desde Veracruz, Universidad Veracruzana, Instituto de Investigaciones en Educación, Red de Investigadores en Educación de Veracruz.

[8] Antony Escobet, Angela Nebot and F. E. Cellier., 2004, Visual-Fir a New Platform for Modeling and Prediction of Dynamical Systems. Fecha Febrero, 2004. ISBN: 1-5655- 283-0., http://people.inf.ethz.ch/fcellier/Pubs/FIR/scsc04.pdf

[9] Escobet A., Nebot A., Cellier, 2008, F.E. Visual Fir: A Tool for Model Identification and Prediction of Dynamical Complex Systems, Simul. Model. 76-92.

[10] Etchells, T.A., Lisboa, P.J.G., 2006, Orthogonal Search-based Rule Extraction (OSRE) Method for Trained Neural Networks: A Practical and Efficient Approach. IEEE Transactions on Neural Networks 17(2) 374-384.

[11] Etchells, T.A., Nebot, A., Vellido, A., Lisboa, P.J.G., Mugica, F., 2006, Learning What is Important: Feature Selection and Rule Extraction in a Virtual Course. In: The 14th European Symposium on Artificial Neural Networks, ESANN 2006. Bruges, Belgium 401-406.

[12] Fabiana Haderne Fabiana , 2006, Uso de Tecnologías de la Información para Detectar Posibles Deserciones Universitarias, Marisa Universidad Nacional de Cuyo Universidad del Aconcagua, Mendoza, Argentina.

[13] Fischer Angulo Erwin Sergio, Modelo para la Automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios, 2012, Universidad De Chile, Facultad De Ciencias Físicas Y Matemáticas, Departamento De Ciencias De La Computación.

[14] Giraldo Mejia J. C., Jiménez Builes J. A., 2013 , Caracterización del Proceso de Obtención de Conocimiento y Algunas Metodologías para Crear Proyectos de Minería de Datos, Facultad de Minas, Universidad Nacional de Colombia, Medellin, Antioquia

[15] Gómez Miranda P. y Vázquez T. F., 2009, Metodología de Razonamiento Inductivo Difuso Para la Predicción de Concentraciones de Ozono. Unidad ProfesionalInterdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas (UPIICSA) del Instituto Politécnico Nacional (IPN). IV Simpósio Internacional de Medio Ambiente. Río de Janeiro, 6 a 10 de Julio de 2009., http://www.sepi.upiicsa.ipn.mx/papers/FVT 2009 2.pdf

[16] Hwang, G.J., 1999, A Knowledge-Based System as an Intelligent Learning Advisor on Computer Networks. J. Systems, Man, and Cybernetics 2 153-158.

[17] Jing Luan, 2010, Aplicaciones de minería de datos en la educación superior, 2010, Planificación y servicios educativos, San Mateo Community College District Fundador, Jiluan Knowledge, Discovery Laboratory, IBM Software, Business Analytics

[18] Jiménez Galindo Á., Álvarez García H., 2010, Minería de Datos en la Educación, Universidad Carlos III de Madrid

[19] Kleiman Ariel, 1975, La Prevision de la Demanda de Educacion Superior y los recursos necesarios para satisfacerla. Presentado a la XVI Asamblea General Ordinaria de la ANUIES celebrada en marzo de 1975.

[20] Minaei-Bidgoli, B., Punch, W.F, 2003, Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. In: Cantu, P.E., et al. (eds.): Genetic and Evolutionary Computation Conference, GECCO 2003, 2252-2263.

[21] Nebot, A., Castro, F., Vellido, A., Mugica, F., 2006, Identification of Fuzzy Models to Predict Students Performance in an e-Learning Environment. In: Uskov, V. (ed.): The Fifth IASTED International Conference on Web-Based Education, WBE 2006. Puerto Vallarta, México 74-79.

[22] Nebot, Ángela, Mugica, Francisco and Castro, Félix, 2009, Causal relevance to improve the prediction accuracy of dynamical systems using inductive reasoning',International Journal of General Systems, 331 — 358

[23] Nebot A., Castro F., Vellido A. and Mugica F., 2006, Identification of Fuzzy Models to Predict Students Performance in an E-Learning Environment: The Fifth Iasted International Conference on Web-Based Education. Puerto Vallarta, México, 2006, 74-79.

[24] Ogarrio Rojas Pascual, 2012, La deserción escolar en jóvenes en pobreza patrimonial: Programa de Becas de Educación Media Superior y los factores de deserción, Facultad Latinoamericana de Ciencias Sociales, Sede Académica de México.

[25] Porcel E., Dapozo G. y López M. (2010). Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. Revista Electrónica de Investigación Educativa, 12(2). Consultado el día de mes de año en: http://redie.uabc.mx/vol12no2/contenido-porcel dapozo.html

[26] Porcel, Eduardo, Dapozo, Gladys, López, María V., 2010, Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios, Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste. 9 de Julio Nº 1449. CP 3400. Corrientes. Argentina.

[27] Rodríguez Rodríguez Jorge Enrique, Desarrollo de herramientas para mineria de datos "UDMiner", 2012, Magíster en Ingeniería de Sistemas. Especialista en Ingeniería de Software. Especialista en Diseño y Construcción de Soluciones Telemáticas. Ingeniero de Sistemas. Docente investigador de la Universidad Distrital Francisco José de Caldas. Director del Grupo de Investigación en Inteligencia Artificial de la misma Universidad

[28] Timarán Pereira Ricardo, 2010, Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos. Revista

Científi ca Guillermo de Ockham. Vol. 8, No. 1. Enero - junio de 2010 - ISSN: 1794-192X - pp. 121-130

[29] Timarán Pereira Ricardo, 2010, Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos, Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño, San Juan de Pasto, Nariño, Colombia

[30] Valero Sergiol, 2009, Aplicación de técnicas de Minería de Datos para predecir deserción http://www.utim.edu.mx/ svalero/docs/MineriaDesercion.pdf

[31] Ian H. Witten and Eibe Frank, 2005, Data Mining, Practical MAchine Learning Tools and Techniques, Second Edition, 4 - 60.