ISSN 1870-4069

Evaluación de parámetros de encuesta de ingreso del CENEVAL para alumnos candidatos a ingresar al nivel superior, caso de estudio ITP

David Gonzalez-Marron,Angelica Enciso-Gonzalez, Ana Karen Hernandez-Gonzalez, David Gutierrez-Franco, Brenda Guizar-Barrera, Alejandro Marquez-Callejas

Instituto Tecnológico de Pachuca, Pachuca, Hidalgo, México

{dgonzalez, a_enciso}@itpachuca.edu.mx, {karen_heez,davidgf_1993}@hotmail.com, {b.soffgb.comalejandromarqueztec}@gmail.com

Resumen. Los estudiantes candidatos a ingresar a las universidades, requieren efectuar un examen de ingreso donde se verifiquen sus conocimientos adquiridos en el nivel académico anterior a fin de poder desempeñarse adecuadamente en sus estudios superiores, uno de los principales exámenes realizados es el EXANI-II, el cual es aplicado por CENEVAL, para la realización de este examen es necesario llenar una encuesta con datos generales que proporcionan datos socioeconómicos de cada uno de los candidatos, en este trabajo se presenta un análisis de las preguntas utilizadas en esta encuesta junto con los resultados obtenidos en los exámenes, con el fin de validar si hay algunos parámetros socioeconómicos que permitan predecir el factor de éxito en la realización de exámenes de ingreso al Instituto Tecnológico de Pachuca.

Palabras clave: minería de datos, ETL, educación.

Evaluation of CENEVAL Admission Surveyed Parameters for Students that are Candidates to Enter the Higher Education, ITP Study Case

Abstract. Students who are candidates to enter universities, require an entrance examination to verify their knowledge acquired at the previous academic level in order to be able to perform adequately in their higher studies, one of the main tests carried out is the EXANI-II, the Which is applied by CENEVAL, who demands to fill out a survey with general data that provide socioeconomic data of each of the candidates, in this work an analysis of the questions used in this survey together with the results obtained in the exams, are used to validate if there are some socioeconomic parameters that allow predicting the success factor in the entrance exams to the Instituto Tecnologico de Pachuca.

Keywords: data mining, ETL, education.

1. Introducción

En México el Centro Nacional de Evaluación para la Educación Superior (CENEVAL) es el órgano encargado de diseñar y aplicar los instrumentos de evaluación orientados al ingreso y egreso del nivel superior, realiza además el análisis y la difusión de los resultados obtenidos en estas evaluaciones. En este documento se analiza el cuestionario de contexto que se realiza al momento de la solicitud del examen de ingreso a educación superior EXANI-II, el cual es un examen estandarizado que está dirigido a sustentantes que han concluido los estudios de bachillerato y aspiran a ingresar al nivel superior en institutos, colegios y universidades que han contratado los servicios del CENEVAL. Este examen tiene el propósito de establecer una valoración global que permita determinar quiénes son los aspirantes con mayor probabilidad de éxito en el nivel superior. El examen se compone de dos pruebas; EXANI-II Admisión, que evalúa aptitudes y competencias disciplinares predictivas del desempeño, su propósito es detectar el potencial de los aspirantes para cursar con éxito el primer año del nivel educativo al que desean ingresar, apoyando a la toma de decisiones de las instituciones educativas sobre el ingreso a los programas académicos que ofertan, y el EXANI-II Diagnóstico, que mide las competencias disciplinares esenciales que deben dominar los estudiantes para ingresar al programa de educación superior que han elegido. Dado su carácter diagnóstico, la institución usuaria tiene la prerrogativa de incluir o no esta prueba en su proceso de selección.

Tabla 1. Histórico de aspirantes y aceptados.

Plan de Estudios	2013		2014		2015		2016	
	Aspirantes	Aceptados	Aspirantes	Aceptados	Aspirantes	Aceptados	Aspirantes	Aceptados
Arquitectura	373	80	356	80	357	80	342	80
Administración	84	40	63	40	72	36	64	40
Civil	344	80	328	80	396	80	390	80
Diseño Industrial	0	0	0	0	71	40	63	40
Eléctrica	92	40	84	70	96	80	91	80
Industrial	171	80	168	80	247	80	195	80
IGE	96	40	74	40	107	60	106	80
ITIC´s	29	25	52	40	51	35	38	28
Mecánica	160	80	143	80	141	76	127	77
Química	138	80	129	80	138	80	123	80
Sistemas Computacionales	215	80	167	80	168	80	183	80

Los resultados de este examen se pueden ubicar en 3 categorías, estas son: *elemental*, *satisfactorio* y *sobresaliente*, que permiten identificar los factores que inciden en el desempeño de los sustentantes del examen EXANI-II y que están basados en el índice CENEVAL, cuya escala abarca desde los 700 puntos (calificación más baja) hasta los

1300 puntos (calificación más alta). La categoría 1 o *elemental* corresponde a la obtención de 700 a 899 puntos del índice CENEVAL. La categoría 2 o *satisfactoria* va de 900 a 1099 puntos del índice CENEVAL y la categoría 3 o *sobresaliente*, se asigna a los sustentantes que obtienen entre 1100 y 1300 puntos del índice CENEVAL [1]

En este artículo se analizan los atributos socioeconómicos del cuestionario de contexto del EXANI-II, aplicados a 3266 aspirantes a ingresar al Instituto Tecnológico de Pachuca en dos años diferentes, y así tratar de determinar si algunos de estos datos tienen una influencia en el desempeño del examen, se busca identificar si existe un perfil social de los aspirantes aceptados que permita ser reforzado por tutores a través de una tutoría efectiva durante los primeros semestres de estadía de los estudiantes en la institución. El Instituto Tecnológico de Pachuca ofrece 11 programas educativos, los cuáles se muestran en la Tabla 1, pudiendo ser visto que existen programas que tienen más demanda que otros, en la tabla igualmente se muestra el número de aspirantes y aceptados en cuatro diferentes años [2].

Cabe mencionar que la política del Instituto Tecnológico de Pachuca para la selección del EXANI- II se conforma utilizando el siguiente criterio: Examen diagnóstico, al cual se le asigna un valor de 80% y de un 20% para el módulo de especialidad, el criterio para los aspirantes aceptados es que el ICE (Índice de desempeño del CENEVAL) sea de 1000 puntos o más, aunque hay excepciones para aceptar a aspirantes cuyo puntaje es de 920 o mayor, esto con el objeto de formar grupos más nutridos dependiendo del programa educativo seleccionado.

Por la diversidad de los programas educativos es necesario contar con información confiable que permita abatir el índice de deserción, de reprobación y aumentar la eficiencia terminal, como puede ser constatado en las investigaciones llevadas a cabo por investigadores de la Universidad Autónoma de Yucatán [10] y por investigadores del Centro Nacional de Evaluación de la Educación Superior [11].

La estructura del documento está conformada de la siguiente manera, en la sección 2 se describe la encuesta utilizada para este análisis, en la sección 3 se describe el proceso de ETL (extracción, transformación y cargado de los datos), las alternativas para la integración de datos, los comandos utilizados para hacer que las bases de datos de los diferentes exámenes aplicados sean equivalentes, en la sección 4 se explica el algoritmo de Minería de datos empleado en este trabajo y por último en la sección 5 se presentan las conclusiones a las que se llegaron con los resultados de este análisis.

2. Descripción de encuesta realizada

El cuestionario de contexto del EXANI-II incluye factores generales y educativos; el cual se realiza al momento del registro para la obtención de ficha. Éste recopila información socioeconómica, psicológica, personal y académica, estructurada en tres áreas, como se muestra en la Tabla 2.

La información previa del EXANI-II está constituida por 90 atributos originales, los cuales fueron normalizados a 52 para un tipo de exámenes y de 49 para otro tipo de exámenes como se muestra en la Tabla 3, una vez realizada esta operación, se procedió a hacer un filtrado adicional debido a que muchos datos proporcionados por el CENEVAL eran datos redundantes , datos resultados de los exámenes, o datos personales, por lo cual fueron retirados del análisis 13 campos adicionales, los atributos

relacionados con aspectos socioeconómicos fueron englobados en seis constructos como se propone en [12].

Tabla 2. Estructura del cuestionario de contexto del EXANI-II.

Área	Dominio Indicador	Descripción
Personal	Datos Generales	Recopila información general del sustentante, así como si alguno o ambos padres hablan una lengua indígena o dialecto.
	Recursos no cognitivos	Recopila información acerca de la motivación, autocontrol, manejo del tiempo y el compromiso académico.
	Recursos cognitivos	Recopila información sobre la habilidad para realizar actividades de ofimática, el uso de internet y el nivel de dominio del inglés.
	Situación laboral	Recopila información sobre los antecedentes laborales y el trabajo actual.
	Trayectoria académica	Recopila el promedio del bachillerato del sustentante así como si tuvo alguna beca previa.
Escolar	Características de la escuela de procedencia	Identifica el año de egreso, el régimen, modalidad y la institución de origen entre otros.
Social	Familiar	Evalúa a través de la estructura familiar, el capital cultural y económico, así como el involucramiento de los padres en sus estudios.

Tabla 3. Conformación de los constructos encontrados.

	Constructo	Número de variables
1	Habilidad para escribir	6
2	Trabajo bajo presión	7
3	Aspiraciones personales	4
4	Qué tanto lo describe	8
5	Planeación de actividades	6
6	Entorno social	21
	Total	52

A continuación, se describen estos constructos:

Aspiraciones personales: Se valoran aspectos como el máximo nivel de estudios que le gustaría estudiar al sustentante, a que sueldo aspira dentro de 10 años en caso de que se gradúe de una carrera universitaria, en caso de que no se gradúe de una carrera universitaria y en caso de que se gradúe de un posgrado.

Habilidad para escribir: Se conforma de un grupo de valoraciones relacionadas con la habilidad que tiene el sustentante para expresar sus ideas de manera escrita. Involucra su percepción con respecto a sus compañeros, la redacción de su opinión sobre un tema,

la escritura de reportes de una lectura, la redacción de una historia, la escritura de una carta a un familiar y la redacción de una solicitud a una autoridad.

Trabajo bajo presión: Incluye variables que evalúan la forma de respuesta del sustentante a distintas condiciones de evaluación. Por ejemplo, evalúa que tan familiarizado está el estudiante con los exámenes de opción múltiple, si los nervios afectan su desempeño, si el nivel de importancia del examen afecta su desempeño, si el nerviosismo repercute en su velocidad de respuesta, si la presión de tiempo incrementa su nerviosismo, su tipo de respuesta ante preguntas muy difíciles y si ha tenido una experiencia previa al examen en cuestión.

Que tanto lo describe: Se pretende evaluar la actitud y perseverancia del sustentante con valoraciones sobre si se desamina al encontrarse con problemas inesperados, si hay dificultad para mantener la atención en metas que requieren varios meses para ser alcanzadas, si se considera una persona que se esmera, si nuevas metas lo distraen de otras previamente establecidas, entre otras.

Planeación de actividades: Evalúa la organización y cumplimiento de prioridades mediante la valoración de aspectos como la elaboración de una lista de actividades, la planeación de actividades del día, claridad de logros para la semana siguiente, establecimiento de prioridades, cumplimiento de prioridades, realización de cosas que intervienen con sus prioridades.

Entorno social: Evalúa la situación socioeconómica del sustentante mediante la evaluación del nivel de estudios de los padres, la existencia de una figura que oriente académicamente al sustentante, si el sustentante cuenta con servicios como teléfono, lavadora, internet, tabletas electrónicas, horno de microondas, televisión de paga, número de televisores, número de reproductores de DVD, número de computadoras, número de autos, número de baños completos de la casa, entre otras.

Tabla 4. Áreas clasificadas.

Arquitectura	Administración	Ingenierías
Índice de pensamiento matemático	Índice de pensamiento matemático	Índice de pensamiento matemático
Índice de pensamiento analítico	Índice de pensamiento analítico	Índice de pensamiento analítico
Índice de escritura de la lengua	Índice de escritura de la lengua	Índice de escritura de la lengua
Índice de comprensión lectora	Índice de comprensión lectora	Índice de comprensión lectora
	Dictamen de Estadística	Dictamen de Física
		Dictamen de Matemáticas
Dictamen de Ingles	Dictamen de Ingles	Dictamen de Ingles

Con el fin de identificar qué factores socioeconómicos impactan en el desempeño de los sustentantes del EXANI-II, se realizó un análisis con datos provistos por el departamento de Desarrollo Académico del Instituto Tecnológico de Pachuca,

correspondientes a dos años diferentes y aplicada a 3266 aspirantes, de los cuales solo 3105 de éstos contaban con información completa para poder realizar el análisis.

Los datos proporcionados por los 11 programas educativos se clasificaron para su análisis en tres áreas diferentes, ya que por su naturaleza presentaban diferencias fuertes que impactan en el análisis, en la Tabla 4 se muestran los índices de relevancia para cada programa.

3. Realización del proceso de ETL

Para la realización del proceso ETL, se clasificaron los datos en 3 diferentes rubros, ya que son tres áreas que se evalúan con diferentes parámetros, como: (Arquitectura, Administración e Ingenierías) buscándose datos comunes para cada una de éstas áreas, El análisis de cada uno de los campos y de cada uno de los archivos consistió en indagar qué significado tiene cada uno para los objetivos planificados en el proyecto, solo se consideraron los campos de mayor relevancia y que con la aplicación de minería de datos podrían ofrecernos los resultados relevantes. Los campos que contienen datos completamente equivalentes en las 3 áreas seleccionadas son los siguientes y se muestran en la Tabla 5. [3]

Tabla 5. Datos comunes en encuestas.

Campo	Representación
Carrera	identifica la carrera de los aspirantes
Sexo	sexo del aspirante
LI_MAD	lengua indígena de la madre
LI_PAD	lengua indígena del padre
CVE_PROC	clave de la institución de procedencia
PROM_BAC	promedio de bachillerato
BECA_NAC	beca de necesidad económica
HRS_TRAB	horas a la semana de trabajo
ESCO_MAD	escolaridad de la madre
ESCO_PAD	escolaridad del padre
SER_INTE	servicio de internet
SER_CABL	servicio de televisión por cable
BIEN_PC	número de computadoras que hay en casa
ICNE	índice de desempeño CENEVAL
IM	índice de desempeño en matemáticas
ILM	índice de desempeño en lógica matemática
IE	índice de desempeño en español
ICL	índice de desempeño en comprensión lectora
DFIS*	dictamen de desempeño en física
DMAT	dictamen de desempeño en matemáticas
DING	dictamen de desempeño en ingles
DEST**	dictamen de desempeño en estadística
Aceptado	muestra si el aspirante fue aceptado o no

*Solo aplica para las bases de datos de ingenierías
**Solo aplica para las bases de datos de administración

Existen 23 atributos comunes en las encuestas analizadas, siendo innecesario realizar alguna transformación en ellos, sin embargo la importancia de estos datos no se pudo

validar hasta el momento de hacer su análisis. En la Tabla 6 se muestran las diferencias existentes entre dos tipos de encuestas realizadas en diferentes años. Con el fin de poder realizar un análisis equivalente, se estandarizan los campos utilizados y los dictámenes obtenidos, utilizando el número 1 para resultados satisfactorios y el número 2 para resultados no satisfactorios.

Tabla 6. Análisis de atributos equivalentes de dos años diferentes.

Encuesta 20**			Encuesta 20**		
Arquitectura	IRLM	índice de razonamiento lógico matemático	Arquitectura	IPMA	índice de pensamiento matemático
	IMAT	índice de matemáticas		IPAN	índice de pensamiento analítico
	IRV	índice de razonamiento verbal		IELE	índice de escritura de la lengua
	IESP	índice de español		ICLE	índice de comprensión lectora
	DDD_MF_ING*	dictamen de ingles		DDD_ML_ING*	dictamen de ingles
Administración	IRLM	índice razonamiento lógico matemático	Administración	IPMA	índice de pensamiento matemático
	IMAT	índice de matemáticas		IPAN	índice de pensamiento analítico
	IRV	índice de razonamiento verbal		IELE	índice de escritura de la lengua
	IESP	índice de español		ICLE	índice de comprensión lectora
	DDD_MF_EST*	índice de estadística		DDD_ML_EST	dictamen de estadística
	DDD_MF_ING*	dictamen de ingles		DDD_ML_ING*	dictamen de ingles
Ingenierías	IRLM	índice razonamiento lógico matemático	Ingenierías	IPMA	índice de pensamiento matemático
	IMAT	índice de matemáticas		IPAN	índice de pensamiento analítico
	IRV	índice de razonamiento verbal		IELE	índice de escritura de la lengua
	IESP	índice de español		ICLE	índice de comprensión lectora
	DDD_MF_ING*	dictamen de ingles		DDD_ML_ING*	dictamen de ingles
	DDD_MG_FIS*	dictamen de física		DDD_ML_FIS*	dictamen de física
	DDD_MG_MAT	dictamen de matemáticas		DDD_ML_MAT*	dictamen de matemáticas

En la Tabla 7 se muestra la equivalencia de campos para diferentes años y el nombre estandarizado utilizado para uniformizar los atributos [3].

Tabla 7. Estandarización de Atributos equivalentes.

Carrera	Campo _año1	Campo_año2	Nombre Estandarizado
Todas	IRML	IPAN	ILM: Índice lógico matemático
Todas	IMAT	IPMA	IM: Índice de matemáticas
Todas	IRV	ICLE	ICL: Índice de comprensión lectora
Todas	IESP	IELE	IE: Índice de español
Todas	DDD_MF_ING	DDD_ML_ING	DING: Índice de inglés
Administración	DDD_MF_EST	DDD_ML_EST	DEST: Índice de Estadística
Ingeniería	DDD_MG_FIS	DDD_ML_FIS	DFI: Índice de Física
Ingeniería	DDD_MG_MAT	DDD_ML_MAT	DMAT: Índice de Matemáticas

Es necesario realizar diferentes transformaciones a otros campos, a fin de generar archivos compatibles que sea posible utilizar para hacer un análisis con diferentes algoritmos de minería, algunos de estos cambios realizados para estandarización se describen en la Tabla 8.

Tabla 8. Estandarización de Atributos equivalentes.

CAMPO	Actividades de estandarización realizadas
Promedio	Uso de 9 rangos diferentes
Licenciatura de los padres	Uso de 2 valores
Horas trabajadas	Uso de un rango de valores
Escolaridad de los padres	Uso de un rango de valores
Becas por desempeño académico, financieras y deportivas	Uso de 2 valores

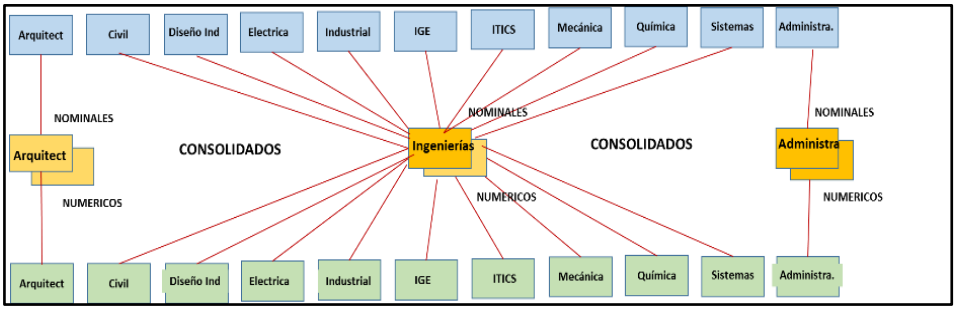

Fig. 1. Proceso de integración de datos utilizando ETL.

Tabla 9. Vista parcial del archivo de salida ARFF.

Viewer

Relation: ADM

PROC	10: PROM_BAC	11: BEC_DAC	12: BEC_NEC	13: BEC_HDA	14: HRS_TRAB	15: ESCO_MAD	16: ESCO_PAD	17: CUAN_LIB	18: SER_TELE	19: SER_LAV	20: SER_REF	21: SER_HOR	22: SER_INTE	23: SER_CABL
nal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
0	5.0	NO	NO	NO	1.0	CARRERA T...	CARRERA T...	3.0	SI	NO	SI	NO	NO	NO
0	4.0	NO	NO	NO	1.0	CARRERA T...	BACHILLER...	4.0	SI	SI	SI	NO	SI	NO
0	7.0	NO	NO	NO	1.0	BACHILLER...	SECUNDARIA	2.0	SI	NO	SI	NO	NO	NO
0	5.0	SI	NO	NO	2.0	BACHILLER...	LICENCIATU...	6.0	SI	NO	SI	NO	NO	SI
0	6.0	NO	NO	NO	1.0	BACHILLER...	LICENCIATU...	6.0	SI	SI	SI	SI	SI	SI
0	3.0	NO	NO	NO	1.0	BACHILLER...	SECUNDARIA	2.0	NO	SI	SI	NO	NO	NO
0	6.0	SI	SI	NO	1.0	BACHILLER...	NO LO SE	7.0	SI	SI	SI	SI	SI	SI
0	3.0	NO	NO	NO	1.0	SECUNDARIA	SECUNDARIA	4.0	NO	SI	SI	NO	NO	NO
0	5.0	NO	SI	NO	1.0	SECUNDARIA	SECUNDARIA	4.0	NO	NO	SI	NO	SI	NO
0	7.0	SI	NO	NO	1.0	SECUNDARIA	NO LO SE	3.0	NO	NO	SI	NO	NO	NO
0	5.0	NO	NO	NO	2.0	SECUNDARIA	SECUNDARIA	2.0	NO	SI	SI	NO	SI	SI
0	5.0	NO	NO	NO	4.0	CARRERA T...	LICENCIATU...	5.0	SI	SI	SI	SI	SI	SI
0	5.0	SI	SI	NO	1.0	LICENCIATU...	LICENCIATU...	4.0	NO	SI	SI	NO	NO	NO
0	5.0	NO	NO	NO	3.0	PRIMARIA	SECUNDARIA	5.0	SI	SI	SI	SI	SI	NO
0	4.0	NO	NO	NO	2.0	SECUNDARIA	SECUNDARIA	2.0	NO	NO	SI	NO	NO	NO
0	7.0	NO	NO	NO	1.0	SECUNDARIA	SECUNDARIA	7.0	NO	NO	SI	NO	NO	NO
0	5.0	NO	NO	NO	2.0	BACHILLER...	BACHILLER...	3.0	SI	SI	SI	SI	SI	SI
0	5.0	NO	SI	NO	4.0	PRIMARIA	NO LO SE	2.0	NO	NO	SI	NO	NO	SI
0	3.0	SI	NO	NO	6.0	SECUNDARIA	PRIMARIA	6.0	SI	SI	NO	NO	SI	NO
0	4.0	SI	NO	SI	6.0	CARRERA T...	PRIMARIA	4.0	SI	NO	SI	NO	NO	SI
0	5.0	NO	NO	NO	2.0	CARRERA T...	CARRERA T...	4.0	SI	NO	SI	NO	NO	NO
0	6.0	NO	NO	NO	2.0	POSGRADO	LICENCIATU...	4.0	SI	SI	SI	NO	SI	SI
0	7.0	NO	SI	NO	1.0	POSGRADO	CARRERA T...	6.0	SI	NO	SI	NO	SI	NO
0	5.0	SI	NO	NO	1.0	LICENCIATU...	CARRERA T...	3.0	SI	SI	SI	NO	NO	NO
0	8.0	SI	NO	NO	1.0	LICENCIATU...	LICENCIATU...	5.0	NO	SI	SI	SI	SI	SI
0	4.0	NO	NO	NO	2.0	LICENCIATU...	POSGRADO	6.0	SI	NO	SI	NO	SI	NO
0	7.0	NO	SI	NO	1.0	CARRERA T...	PRIMARIA	3.0	NO	NO	SI	NO	NO	SI
0	8.0	NO	NO	NO	6.0	CARRERA T...	CARRERA T...	5.0	NO	NO	SI	NO	SI	SI
0	5.0	NO	NO	NO	1.0	BACHILLER...	SECUNDARIA	2.0	NO	SI	SI	SI	NO	NO

Para probar la efectividad de los algoritmos supervisados, se probaron en tipos de datos numéricos y nominales, cuidando en todo momento que los archivos fueran equivalentes a fin de tener una base común de comparación. El proceso de transformación de los datos se realizó utilizando Pentaho [5], permitiendo la transformación de archivos de Excel conformados por 22 archivos en formato CSV (comma-separated values), 11 por cada año analizado y convertidos a 6 archivos ARFF

(Attribute-Relation File Format) 3 nominales y 3 numéricos como puede ser visto en la Figura 1. Los archivos fueron conformados por los atributos que se muestran en la Tabla 9.

4. Realización del proceso de minería

Para la realización del proceso de minería se consideraron los resultados obtenidos de los candidatos en el examen de ingreso CENEVAL de 2 años, ya que se contaba con dicha información a detalle, primeramente se utilizaron algunos algoritmos de clasificación para determinar cuáles son los atributos más influyentes para la determinación de los resultados obtenidos por los candidatos (aceptado o rechazado), utilizando para esto el enfoque de análisis de componentes principales (PCA) [4][7], posteriormente se analizaron que métodos de clasificación proporcionaban mejores resultados. Para la realización del análisis se utilizaron métodos aplicables a datos nominales y métodos para datos numéricos, los tipos de algoritmos utilizados para la evaluación fueron los siguientes: Algoritmos de reglas, algoritmos de árboles, algoritmos de Bayes, algoritmos perezosos y metaalgoritmos [7]. El análisis de componentes principales (PCA) arrojó el siguiente resultado, como puede ser visto en la Tabla 10, donde se muestran los atributos con más relevancia en cada área.

Tabla 10. Análisis de componentes principales PCA para las tres áreas analizadas.

	Campo1	Campo2	Campo3	Campo4	Campo5	Campo6	Campo7	Campo8	Campo9	Campo10
INGENIERIA (2325 Registros)	Trabaja 22.31	Promedio 17.16	Año-Nac 7.01	Vacaciones 5.34	Lug-Proced 4.85	Internet 4.80	Esco-Madre 4.79	Num-libros 4.74	Television 3.85	Esco-Padre 3.74
ARQUITECT (633 Registros)	Trabaja 28.88	Esco-Padre 13.26	Promedio 7.66	Esco-Madre 7.24	Año-Nacim 6.0	Num-libros 5.46	Bec-DepArt 4.25	Automovil 4.12	Tam-Casa 3.89	LenInd-Pad 6.84
ADMON (145 Registros)	Trabaja 30.56	Num-Libros 10.02	Esco-Padre 9.2	Esco-Madre 8.98	Año-Nacim 7.03	LenInd-Mad 6.84	Bec-Acade 6.55	Bec-DepArt 6.26	Serv-Cable 4.35	Promedio 3.25

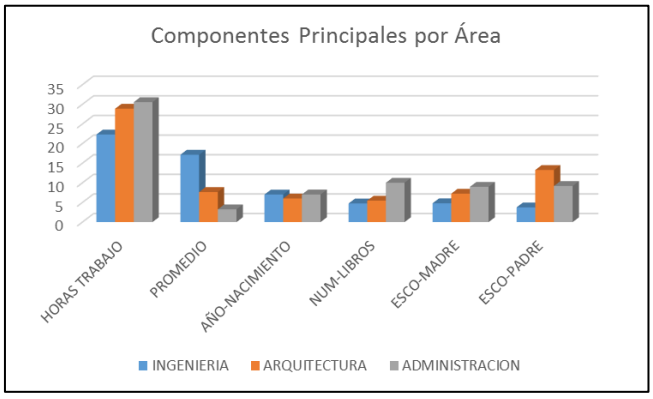

Fig. 2. Contraste de componentes principales por áreas analizadas.

Para la selección de los componentes principales se utilizó la aplicación de BigML [8], seleccionando aquellos atributos que se presentaban en las 3 áreas analizadas, encontrándose coincidencia en 6 atributos, es importante considerar que los más representativos, fueron aquellos pertenecientes al área de ingeniería, ya que el número de registros analizados fueron 2325, contrastados con arquitectura con 635 registros y de únicamente 145 registros pertenecientes a el área de administración.

Aunque el número de registros varía considerablemente, se encontró coincidencia en 6 de los 10 campos más importantes, lo que nos indica que estos campos o atributos son de relevancia para determinar si un alumno es aceptado o rechazado en su examen de ingreso, como puede ser visto en la Figura 2.

El modelo obtenido es presentado en la Figura 3 en forma de árbol, donde puede ser visto que los atributos presentados en la Figura 2 son los que nos permiten hacer una correcta clasificación para el dato objetivo, el cual determina si un alumno aprueba el examen de admisión o no lo hace. A fin de validar la efectividad de la predicción, es generado un nuevo campo denominado "aceptado_prediccion", donde se compara el valor real con el valor calculado a fin de validar la efectividad del modelo utilizado, se seleccionó la métrica denominada como exactitud o accuracy, que está conformada por: (*total_aciertos / total_predicciones*).

Es importante mencionar que existen suficientes casos de aspirantes rechazados y aceptados para tener una muestra representativa de datos y poder realizar un buen análisis.

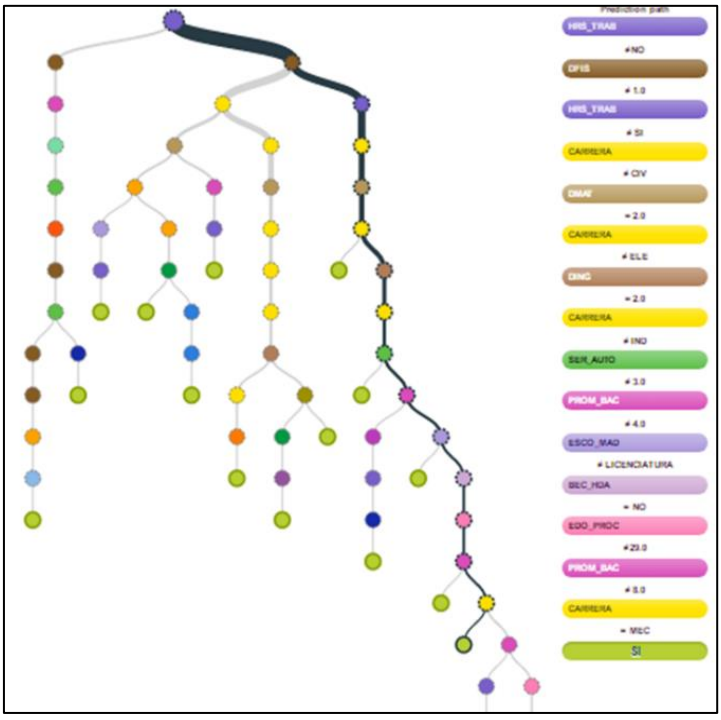

Fig. 3. Modelo de Clasificación obtenido para los datos analizados. [8]

La evaluación del modelo se realizó dividiendo los datos en 2 conjuntos de datos, utilizando 80% de los datos para entrenamiento y el 20% para prueba a fin de evitar un sobreajuste [9], los resultados obtenidos de esta evaluación se presentan en la Figura 4.

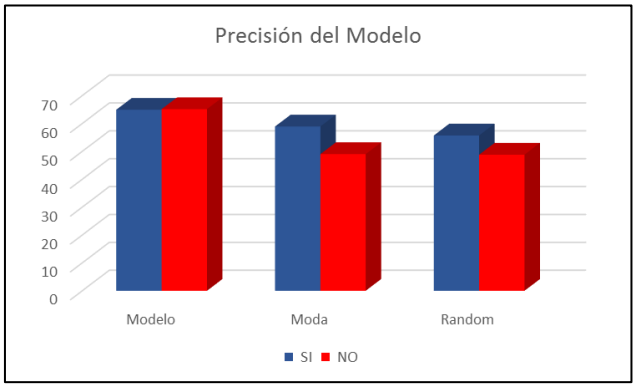

Fig. 4. Evaluación del modelo considerando la exactitud como métrica principal.

En la Figura 5 se muestra el detalle del campo predictor más importante que es el número de horas que trabaja el sustentante, pudiéndose ver que la mayor parte de los candidatos que desean ingresar al Instituto trabajan de 15 a 40 horas semanales.

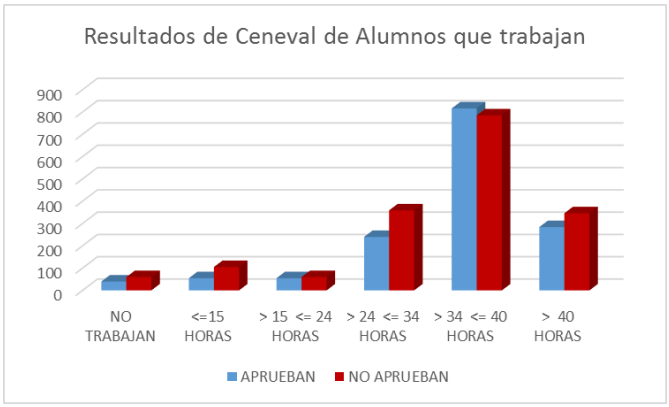

Fig. 5. Desempeño obtenido del campo predictor principal horas trabajadas.

5. Conclusiones

- En base a los resultados obtenidos para el análisis de datos realizado, se puede concluir lo siguiente, el proceso ETL requirió la realización de adecuaciones de los datos, su transformación y cargado en formatos comunes para los diferentes tipos de archivos proporcionados, se englobaron primeramente todas las ingenierías en un archivo común con tipos de datos numéricos y datos nominales, posteriormente se

reunió información de los archivos de las carreras de Arquitectura y Administración para generar un solo archivo global estandarizado.

- Los atributos más relevantes en orden de importancia fueron (horas trabajadas por semana, promedio del bachillerato, año de nacimiento del estudiante, número de libros en casa del estudiante, escolaridad de la madre y escolaridad del padre).
- Los algoritmos utilizados fueron de clasificación para datos nominales y de regresión para datos numéricos, en base a la información proporcionada por el modelo.
- Se pretende generar un plan que permita mejorar la acción de los tutores, donde sea factible, ya que el principal factor considerado fue el de las horas que dedica al trabajo el sustentante, sin embargo este factor seguramente requerirá un apoyo de parte de la institución y de los tutores a fin de no descuidar los estudios de los alumnos una vez que hayan sido aceptados, ya que los alumnos aceptados requieren de tiempo completo para el cumplimiento de sus actividades académicas y frecuentemente este es un factor de reprobación en los primeros semestres de los estudiantes.
- El promedio del bachillerato era algo esperado para poder aprobar el examen del CENEVAL, y efectivamente fue así, lo que valida que el modelo funciona correctamente.
- El número de libros en casa del estudiante fue algo que sorprendió gratamente, ya que indica que, si en la casa hay hábitos de lectura, esto incide en el desempeño de los alumnos.
- La escolaridad del padre y la madre fueron decisivos para un buen desempeño del CENEVAL, ya que en cierta manera refleja cierta presión o acompañamiento de parte de los padres para que los hijos estudien de manera regular.

Referencias

1. CENEVAL: Examenes Nacionales de Ingreso, Disponible en: http://www.ceneval.org.mx/web/guest/exani-ii (2017)
2. ITPACHUCA: Estadística de Ingresos de Alumnos 2013-2016. Departamento de Desarrollo Académico, Informe Interno (2016)
3. Chapa, M., Guizar, B., Franco, D., Hernández, A.: Minería de datos aplicada a encuesta de ingreso del ITP. Reporte de investigación interno (2016)
4. Pérez-López, C.: Minería de datos técnicas y herramientas. Madrid España: Thomson Ediciones Paraninfo S.A. (2008)
5. Rapidminercom.: RapidMiner. Disponible en https://rapidminer.com (2015)
6. Waikatoacnz: Waikatoacnz. Disponible en http://www.cs.waikato.ac.nz/ml/weka/requirements.html (2015)
7. Witten, I., Frank, E., Hall, M.: Data Mining Practical Machine Learning Tool and Techniques. Third Edition, Elsevier (2011)
8. BigML Inc.: Programmatic Machine Learning Application. Disponible en: https://bigml.com (2017)
9. Garcia, D.: Manual de Weka. Disponible en http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/.../weka.pdf (2014)

10. López, I., Echazarreta, C., Pech, S., Gómez, B.: Selección y Permanencia en la Educación Superior: el Caso de la Universidad Autónoma de Yucatán. Revista Iberoamericana de Evaluación Educativa, 3(2), pp. 91–93 (2016)
11. Martínez, J., Herrera, M.: Propiedades psicométricas de la escala de cómputo para el EXANI-II. Revista Electrónica de Investigación Educativa, 16(2), pp. 72–74 (2014)
12. Abreu, J.: Constructos, Variables, Dimensiones, Indicadores & Congruencia. Daena: International Journal of Good Conscience, 7(3), pp. 123–130 (2012)