ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLÓGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN: GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

MODELO PARA IDENTIFICAR LOS PATRONES DE COMPORTAMIENTO QUE INFLUYEN EN EL APROVECHAMIENTO ACADÉMICO DE LOS ESTUDIANTES EN LA ESCUELA DE FORMACIÓN DE TECNÓLOGOS DE LA ESCUELA POLITÉCNICA NACIONAL

AUTORA: PARRAGA VILLAMAR, VIVIANA CRISTINA

DIRECTOR: MSC. ZALDUMBIDE PROAÑO, JUAN PABLO

SANGOLQUÍ

2018

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y

TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación **"MODELO PARA IDENTIFICAR LOS PATRONES DE COMPORTAMIENTO QUE INFLUYEN EN EL APROVECHAMIENTO ACADÉMICO DE LOS ESTUDIANTES EN LA ESCUELA DE FORMACIÓN DE TECNÓLOGOS DE LA ESCUELA POLITÉCNICA NACIONAL"** fue realizado por la señora **Parraga Villamar, Viviana Cristina,** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 13 de agosto del 2018

Firma:

.....

Ing. Juan Pablo Zaldumbide MSc.
C.C.: 1715467948

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORÍA DE RESPONSABILIDAD

Yo, **Parraga Villamar, Viviana Cristina**, con cédula de identidad nº 1721903407, declaro que el contenido, ideas y criterios del trabajo de titulación: **"Modelo para identificar los patrones de comportamiento que influyen en el aprovechamiento académico de los estudiantes en la Escuela de Formación de Tecnólogos de la Escuela Politécnica Nacional"** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 13 de agosto del 2018

Firma:

.....
Ing. Viviana Cristina Parraga Villamar
C.C. 1721903407

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN

Yo, **Parraga Villamar, Viviana Cristina** autorizo a la Universidad de la Fuerzas Armadas ESPE publicar el trabajo de titulación **"Modelo para identificar los patrones de comportamiento que influyen en el aprovechamiento académico de los estudiantes en la Escuela de Formación de Tecnólogos de la Escuela Politécnica Nacional"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 13 de agosto del 2018

Firma:

.....

Ing. Viviana Cristina Parraga Villamar
C.C. 1721903407

DEDICATORIA

A Cristopher,

Mi pequeño inventor, por ser esa luz que alegra cada día de mi vida y por ser la inspiración para querer alcanzar metas más grandes. Hijo mío te dedico este trabajo como ejemplo para que nunca te des por vencido, para que veas que, a pesar de los obstáculos, siempre es posible llegar a la cima de la montaña, y que nunca dudes que estaré a tu lado apoyándote para construir el mundo contigo.

A Sebitas,

Mi pequeño futbolista, por ser siempre mi bebé cariñoso y por no permitirme olvidar que, en el juego de la vida, aunque se pierda hay que levantarse y continuar. Hijo mío este trabajo te lo dedico por todos esos partidos de fútbol que no pude acompañarte, y como compromiso que ahora no faltaré a ninguno. Nunca olvides que tú y tu hermano son el pilar de mi vida y que esta meta es una de las tantas que alcanzaremos juntos.

A mis padres,

Por ser los únicos responsables de este título, puesto que, sin ustedes, esto no hubiera sido posible. Papitos este trabajo se los dedico por siempre apoyarme, confiar en mí y sobre todo por querer tanto a mis pequeños retoños.

Vivi

AGRADECIMIENTO

A Dios,

Por darme las fuerzas necesarias para nunca darme por vencida, por permitirme demostrar una vez más que entre mayor sea el obstáculo, mayor será mi esfuerzo. Gracias Dios, por darme la oportunidad de vivir esta experiencia que me ayudo a crecer en lo personal y profesional, por estar conmigo y mis hijos en cada paso que doy, por fortalecer mi corazón e iluminar mi mente para ser una buena madre y profe, y sobre todo por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante todo este periodo de estudio.

A mis hijos,

Ustedes son el mejor regalo que Dios me ha dado, son mi mayor tesoro y la fuente de mi inspiración, es por esto que les quiero agradecer por ser la felicidad de mi vida, por ser el apoyo y compañía cuando mamá tenía que estudiar. Los amo con todo mi ser.

A mis padres,

No me alcanzarían las palabras para escribirles lo tan agradecida que estoy con ustedes, gracias papitos por su ayuda con los tesoros más grandes y preciados que tengo en mi vida, por haber cuidado de mis pequeños durante esta etapa de estudio y por apoyarme siempre para no caer.

A mi director,

JuanPa en verdad, te estoy eternamente agradecida por el apoyo brindado durante el desarrollo de este proyecto, créeme que tus consejos y enseñanzas están reflejados en este trabajo.

A mi familia,

Hermanas mías, gracias por el apoyo que siempre me han dado y sobre todo por ser las tías más amorosas para mis pequeños. De igual manera como no agradecer a toda mi familia que siempre estuvo pendiente de mi para apoyarme cuando los necesite.

A mis amigos,

Amigos masters, gracias por las experiencias compartidas y sobre todos por su apoyo en la culminación de esta etapa de estudios, siempre los llevaré en mi corazón. Gracias a todos mis demás amigos y compañeros por siempre apoyarme con sus palabras y enseñanzas para conseguir esta meta. Sobre todo, te quiero agradecer Nathys por ser esa amiga incondicional que siempre estuvo ahí para no dejarme caer.

Vivi

ÍNDICE DE CONTENIDOS

CERTIFICADO DEL DIRECTOR ..... **i**

AUTORÍA DE RESPONSABILIDAD ..... **ii**

AUTORIZACIÓN ..... **iii**

DEDICATORIA ..... **iv**

AGRADECIMIENTO ..... **v**

ÍNDICE DE CONTENIDOS ..... **vi**

ÍNDICE DE TABLAS ..... **viii**

ÍNDICE DE FIGURAS ..... **ix**

RESUMEN ..... **xii**

ABSTRACT ..... **xiii**

CAPÍTULO I

INTRODUCCIÓN

1.1. Antecedentes 1

1.2. Justificación e Importancia 2

1.3. Planteamiento del problema 4

1.4. Objetivo general 5

1.5. Objetivos específicos 6

1.6. Formulación del problema 6

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

2.1. Marco teórico 8

2.1.1. Fundamentación de la variable Independiente 8

2.1.2. Fundamentación de la variable dependiente 15

2.2. Antecedentes del estado del arte 18

2.3. Marco conceptual 27

CAPÍTULO III

MEMORIA TÉCNICA METODOLÓGICA

3.1. Metodología de Investigación 31

3.2. Ejecución del proceso de investigación 32

CAPÍTULO IV

RESULTADOS

4.1 Informe de Resultados 37

4.1.1. Análisis y selección de datos 37

4.1.2. Preparación de los datos 39

4.1.3. Proceso ETL 42

4.1.4. Base de datos 49

4.1.5. Auto Model RapidMiner 52

4.1.6. Identificación de algoritmo de minería de datos 57

4.1.7. Modelo de minería de datos 61

4.1.8. Evaluación del modelo 74

4.2 Metodología para ejecutar la propuesta 79

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones 82

5.2. Recomendaciones 84

BIBLIOGRAFÍA 86

ÍNDICE DE TABLAS

Tabla 1 *Número de estudiantes ESFOT* *16*

Tabla 2 *Número de registros* *17*

Tabla 3 *Estudios por Grupo de Control* *20*

Tabla 4 *Construcción de cadenas de búsqueda* *21*

Tabla 5 *Datos entregados por la DGIP* *38*

Tabla 6 *Comparación de herramientas ETL* *42*

Tabla 7 *Datos de los atributos determinado por Auto Model* *57*

Tabla 8 *Casos de estudio* *59*

Tabla 9 *Comparación de modelos por precisión* *60*

Tabla 10 *Comparación de modelos por atributo probable* *60*

ÍNDICE DE FIGURAS

Figura 1 Relación de variables 8

Figura 2 Proceso ETL Fuente: (Latino BI, 2013) 9

Figura 3 Tarea metodología Kimball 10

Figura 4 Fase metodología CRISP-DM 13

Figura 5 Cadena de búsqueda ejecutada en SCOPUS 22

Figura 6 Investigación mediante ciclos 32

Figura 7 Diagrama Entidad – Relación 40

Figura 8 Diagrama modelo multidimensional 41

Figura 9 Proceso ETL dimensión estudiante 44

Figura 10 Proceso ETL dimensión facultad 44

Figura 11 Proceso ETL dimensión factores 45

Figura 12 Proceso ETL dimensión materia 46

Figura 13 Proceso ETL dimensión periodo 47

Figura 14 Proceso ETL cargar datos 48

Figura 15 Proceso ETL tabla de hechos 48

Figura 16 Creación de la base de datos en MySQL 49

Figura 17 Conexión con la base de datos MySQL mediante *Data Integration* 50

Figura 18 Creación de tablas de dimensiones en base de datos MySQL 51

Figura 19 Datos cargados a FAC_DATOS de la base de datos Tesis_BDD_V1 51

Figura 20 Base de datos Tesis_BDD_v1 con todas las tablas creadas 52

Figura 21 Unión de tablas en RapidMiner 53

Figura 22 Paso 1 Auto Model RapidMiner 53

Figura 23 Paso 2 Auto Model RapidMiner 54

Figura 24 Paso 3 Auto Model RapidMiner 55

Figura 25 Paso 4 Auto Model RapidMiner 55

Figura 26 Paso 5 Auto Model RapidMiner 56

Figura 27 Resultado Auto Model 56

Figura 28 Proceso creado por Auto Model 57

Figura 29 Atributos de entrada y salida del modelo 59

Figura 30 Proceso para remplazar valores nulos 62

Figura 31 Selección de datos para el modelo 62

Figura 32 Selección de atributo a predecir 63

Figura 33 Opciones de la variable a predecir 63

Figura 34 Selección de atributos de entrada 64

Figura 35 Comparación de algoritmos 64

Figura 36 Resultado con algoritmo Naive Bayes 65

Figura 37 Resultado con algoritmo Lineal Generalizado 65

Figura 38 Resultado con algoritmo Decision Tree 66

Figura 39 Proceso del algoritmo Decision Tree 66

Figura 40 Proceso para eliminar opción “Ninguno” 67

Figura 41 Modelo utilizando árbol de decisión 67

Figura 42 Proceso de validación cruzada 74

Figura 43 División de la data 75

Figura 44 Matriz de confusión del modelo 75

Figura 45 Porcentaje de precisión y cobertura del modelo 77

Figura 46 Proceso de evaluación del modelo.....78

Figura 47 Matriz de precisión 78

Figura 48 Matriz de error de clasificación 79

Figura 49 Simulador del modelo con parámetros óptimos 80

Figura 50 Ejemplo de simulador de modelo 81

RESUMEN

La deserción académica dentro del sistema universitario constituye un desafío para las Instituciones de Educación Superior dentro y fuera del Ecuador. En la Escuela de Formación de Tecnólogos (ESFOT) de la Escuela Politécnica Nacional (EPN) continuamente se busca aumentar la retención de estudiantes, mediante alternativas que suban las tasas de deserción académica. Mediante este trabajo se analizó la información de los estudiantes de la ESFOT, armando una bodega de datos, que luego de ser sometido a un proceso ETL, se agrupó a factores del alumnado con base al rendimiento académico. Esta bodega de datos, permitió crear un modelo de minería de datos que determinó patrones y características que generan repitencia y deserción, para guiar adecuadamente al alumnado con mayor tendencia de abandono o fallo. El proyecto se enmarcó a la metodología de investigación científica basada en el diseño en conjunto con la metodología de minería de datos CRISP-DM, que proporcionó una guía de referencia normalizada del ciclo de vida de un proyecto de análisis de datos. Utilizando el algoritmo DECISSION TREE, y Auto Model de RapidMiner se obtuvo un modelo que identifica los patrones de comportamiento que afectan el rendimiento escolar de los estudiantes de la ESFOT con un nivel de confianza del 96,9%. Se detectaron algunos patrones de comportamiento que influyen en su aprovechamiento académico.

PALABRAS CLAVES:

- **MODELO DE MINERÍA DE DATOS**
- **RENDIMIENTO ESCOLAR**
- **PATRONES DE COMPORTAMIENTO**
- **BODEGA DE DATOS**

ABSTRACT

The academic desertion within the university system constitutes a challenge for Higher Education Institutions inside and outside of Ecuador. In the School of Training of Technologists (ESFOT) of the National Polytechnic School (EPN) is continuously seeking to increase the retention of students, through alternatives that increase the rates of academic desertion. Through this work the information of ESFOT students was analyzed, setting up a Data Warehouse, which after being subjected to an ETL process, groups student factors based on academic performance. This Data Warehouse allowed the creation of a data mining model that determined characteristics and patterns that cause repetition and desertion, to adequately guide students with a higher risk of failure or abandonment. The project was framed to the scientific research methodology based on the design in conjunction with the data mining methodology CRISP-DM, which provided a standard reference guide for the life cycle of a data analysis project. Using the DECISION TREE algorithm and RapidMiner's Auto Model, a model was obtained that identifies behavior patterns that affect the school performance of ESFOT students with a confidence level of 96.9%. Some patterns of behavior that influence their academic performance were detected.

KEYWORDS:

- **DATA MINING MODEL**
- **SCHOOL PERFORMANCE**
- **BEHAVIOR PATTERNS**
- **DATA WAREHOUSE**

CAPÍTULO I

INTRODUCCIÓN

1.1. Antecedentes

Uno de los mayores desafíos de la Educación Superior a nivel mundial es la integración, permanencia y egreso de los estudiantes. “Según la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco), el abandono de ese nivel de formación llega a 40 %(Acosta, 2016) debido a factores como: no les gustan las carreras que eligieron, confusión, trabajo, mala orientación vocacional en los colegios, exigencia académica, razones personales, situación socioeconómica y hasta un contexto familiar en que se pide a los chicos seguir las profesiones de sus padres o hermanos por tradición (Bravo, 2016).

De acuerdo a cifras presentadas por la Secretaría Nacional de Educación Superior (Senescyt), en Ecuador, ocho de cada diez estudiantes, que ingresaron a una universidad o a una escuela politécnica pública, permanecieron estudiando en primer año. El porcentaje de deserción era del 52% antes de la implementación del ENES ahora llega al 20% (Bravo, 2016).

Mientras que en la EPN se tiene tasas de deserción del 50% según datos del 2015, por ejemplo, de 5200 estudiantes que se inscribieron al curso de Nivelación en dicho año el 33% aprobó, el 58% reprobó y el 10% se retiró (Acosta, 2016).

Esta problemática repercute negativamente en el avance económico y social de los países, especialmente, en los que se encuentran en vías de desarrollo tal como Ecuador.

Es por esto que uno de los principales intereses de profesores, autoridades educativas y estudiantes es detectar posibles patrones de comportamiento que influyan en el abandono para gestionar y reducir este comportamiento.

Para contribuir con la solución al problema de fracaso escolar aplicando minería de datos para encontrar conocimiento oculto de la información, y darle una aplicación educativa, es una gran solución.

1.2. Justificación e Importancia

Este proyecto está orientado a la minería de datos educativos aplicada al Ecuador, teniendo como caso de estudio, la Escuela de Formación de Tecnólogos de la Escuela Politécnica Nacional que al pertenecer al Sistema de Educación Superior, donde se tienen problemas de deserción y repitencia académica, con en el resto de instituciones de educación superior en Ecuador (Camana, 2016).

La EPN contaba con una gran cantidad de información de los estudiantes, pero sin algún sistema de minería de datos, realizado a través de bodegas de datos que entreguen datos relevantes de los alumnos y las forma en que aprenden (Ordoñez, 2013).

Se habían determinado algunas características que influyen en el desempeño académico del alumnado de la Escuela Politécnica Nacional como: incumplimiento de expectativas, problemas económicos, motivos personales, falta de comprensión al profesor, metodología de enseñanza inconsistentes, entre otros, pero no se daba el seguimiento adecuado para obtener conocimiento predictivo de la misma y poder conocer el impacto o contribución que tiene cada uno de ellos en el desempeño académico del alumno.

La predicción del rendimiento de los estudiantes fue desafiante debido al gran volumen de datos, la falta de un sistema para analizarlos y monitorear el progreso del rendimiento de los mismos. Había dos razones principales de por qué esto estaba sucediendo. En primer lugar, el estudio sobre los métodos de predicción existentes seguía siendo insuficiente para identificar los métodos más adecuados para predecir el rendimiento de los estudiantes en las instituciones. En segundo lugar, la falta de investigaciones sobre los factores que afectan los logros de los estudiantes utilizando técnicas de minería de datos educativas para traer los beneficios e impactos a estudiantes, educadores e instituciones académicas.

Al aplicar este proyecto, se busca reducir la tasa de deserción con respecto a los años académicos anteriores en los que no se aplicó ningún mecanismo de prevención de deserción escolar. La principal contribución es una herramienta y un plan de tutoría que puede ser utilizado por la institución educativa (y otros) para reducir la tasa de abandono que afectan en una buena parte a las instituciones de educación superior del Ecuador.

La presente investigación realiza un análisis de la información de los estudiantes de la Escuela de Formación de Tecnólogos, almacenadas en las bases de datos del Sistema SAEW de la Escuela Politécnica Nacional correspondiente a los módulos de tutorías académicas, calificaciones y datos socio económicos, entregada por la Dirección de Gestión de Información y Procesos de los semestres que van desde el 2010.

Dicha información fue procesada, para luego ser manipulada por la herramienta de análisis de datos de Pentaho, para hacer la limpieza y transformación de la misma y conseguir la bodega de datos. La bodega de datos fue utilizada en el diseño y creación de la bodega de datos, a partir de las fuentes de información de la Escuela Politécnica Nacional.

A continuación, se identificó los algoritmos de minería de datos más acorde a los datos de la bodega de datos para el diseño del modelo utilizando la herramienta de minería de datos RapidMiner. A partir del modelo creado se determinó los patrones que afectan el desempeño de los alumnos.

Por último, se evaluó el modelo analítico-predictivo a través del uso de la técnica de validación de matriz de confusión que muestra el número de predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba, determinándose así el nivel de confianza en un resultado predictivo.

1.3.Planteamiento del problema

El aumento de la retención de estudiantes ha sido un objetivo común de muchas instituciones académicas, especialmente a nivel universitario. Los efectos negativos de la deserción de los estudiantes son evidentes para los estudiantes, los padres, la universidad y la sociedad en general. La alta deserción académicas, es una problemática que también enfrenta la Escuela de Formación de Tecnólogos (ESFOT) de la Escuela Politécnica Nacional (EPN), debido a diversos factores.

En consecuencia, el estudio de abandono o cese prematuro ha sido ampliamente reconocido como un problema serio, especialmente en el nivel universitario. Un gran número de institutos de educación superior enfrentan la dificultad común con baja tasa de graduaciones en comparación con el número de inscripciones. Por lo cual, la capacidad de predecir el rendimiento de un alumno podría ser útil para la ESFOT.

Con esta idea, se utilizó minería de datos para la detección temprana de estudiantes en riesgo de deserción, analizando información de los mismos sobre: tutorías académicas, calificaciones, datos socio económicos, entre otros, almacenadas en bases de datos de la EPN, enfocándola a encontrar patrones de comportamiento de los ellos.

Anteriormente en la ESFOT lo único que se hacía para reducir la deserción era mantener reuniones personalizadas con cada estudiante en las que un profesor tutor revisa el desempeño académico del mismo y almacenaba los posibles factores que afectaban el desempeño, de acuerdo a la reunión realizada, más no se consideraban datos históricos del rendimiento, problemas económicos, factores que le afectaron en semestres anteriores entre otros, a pesar que dicha información se encuentra almacenada en las bases de datos.

Además, la planificación del área administrativa, docente y psicopedagógica, no poseía porcentajes de probabilidad de que un alumno pueda desertar basada en datos históricos de otros estudiantes, desde que inicia su vida estudiantil, sin permitir que las autoridades de la ESFOT se anticipen con estrategias para disminuir el índice de deserción y tengan asistencia en los procesos de acreditación, gestión de recursos y contratación docente.

Es por esto que uno de los principales intereses de profesores, autoridades educativas y estudiantes fue determinar los múltiples patrones de comportamiento que pueden influir en el abandono para gestionar y reducir este comportamiento.

Para contribuir con la solución al problema de fracaso escolar se aplicó minería de datos para encontrar conocimiento oculto de grandes volúmenes de información, y darle una aplicación educativa.

1.4. Objetivo general

Construir un modelo de minería de datos analizando la información de los estudiantes de la Escuela de Formación de Tecnólogos de la Escuela Politécnica Nacional, para identificar los patrones de comportamiento del alumnado que influyen en el aprovechamiento académico.

1.5. Objetivos específicos

OE1: Analizar la información existente en las bases de datos de la Escuela Politécnica Nacional diseñando un modelo multidimensional, para determinar las posibles herramientas ETL en el proceso de limpieza y transformación de datos.

OE2: Seleccionar la herramienta ETL más factible para el diseño y creación de una bodega de datos, a partir de las fuentes de información de la Escuela Politécnica Nacional.

OE3: Identificar los algoritmos de minería de datos, para ser utilizados en el modelo a diseñar, utilizando una herramienta de minería de datos.

OE4: Crear el modelo de minería de datos mediante el algoritmo para determinar los patrones que afectan el desempeño de los alumnos.

OE5: Evaluar el modelo analítico-predictivo a través del uso de técnicas de validación implementadas en minería de datos, para determinar el nivel de confianza en un resultado predictivo.

1.6. Formulación del problema

Este proyecto analiza la información obtenida, y realizar un modelo minería de datos para identificar patrones que influyen en el desempeño académico y así conocer las causas generales de reprobación y repitencia de materias.

Las preguntas a resolver mediante este modelo son las siguientes:

OE1 – RQ1.1: ¿Qué herramientas ETL permitirán analizar la información existente en las bases de datos de la Escuela Politécnica Nacional para hacer la limpieza y transformación de la misma?

OE1 – RQ1.2: ¿Cuál es el modelo multidimensional que mejor se enmarca para facilitar la búsqueda de información en la bodega de datos?

OE2 – RQ2.1: ¿Cuál es la herramienta ETL más factible para el diseño y creación de una bodega de datos, a partir de las fuentes de información de la Escuela Politécnica Nacional?

OE2 – RQ2.2: ¿Cuál es el gestor de bases de datos que permitirá manipular la información de la bodega de datos de una manera eficaz y optima?

OE3 – RQ3.1: ¿Cuáles son los algoritmos de minería de datos que se deben utilizar en el modelo a diseñar?

OE3 – RQ3.2: ¿Cuál es la herramienta de minería de datos que mejor manipule el algoritmo de *data mining* seleccionado?

OE4 – RQ4.1: ¿Cuál es el mejor proceso para la búsqueda de patrones para aplicarlo en el algoritmo de minería de datos?

OE4 – RQ4.2: ¿Cuál es el modelo de minería de datos que determine las características que afectan el rendimiento de los estudiantes?

OE5 – RQ5.1: ¿Si es posible validar el modelo, se puede determinar un nivel de confianza en los resultados que permita demostrar una mejora en los mismos?

OE5 – RQ5.2: ¿Cuál es el margen de error que se debe considerar al implementar modelos de analítica predictiva?

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

2.1. Marco teórico

El marco teórico busca relacionar la congruencia teórica planteada con la hipótesis, por esta razón es importante organizar una red de categorías en donde se realiza un desarrollo teórico parte de categorías que incluyen las variables del problema, para ir descendiendo jerárquicamente hasta aquellas que comprenden y explican la esencia de las variables que intervienen en la explicación y entendimiento científico del tema de estudio, la red planteada se muestra a continuación:

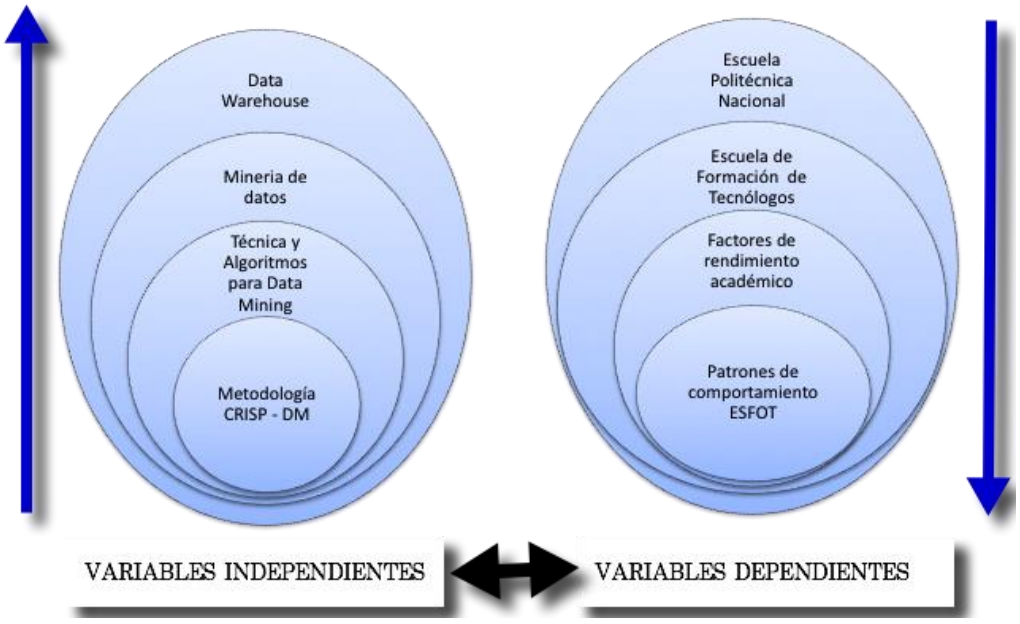


Figura 1. Relación de variables

2.1.1. Fundamentación de la variable Independiente

BODEGA DE DATOS (*DATA WAREHOUSE*): Según *Bill Inmon*, "una bodega de datos es una colección de datos, integrados, no volátiles, variante en el tiempo y orientados a temas, organizados para soportar necesidades empresariales", especialmente soporta los procesos de toma de decisiones gerenciales.

Una bodega de datos o data warehouse (DW) se encarga de integración de diferentes bases y fuentes de datos de los sistemas operacionales y transaccionales para obtener datos consolidados, almacenados en un dispositivo de memoria no volátil.

El proceso de integración de datos se lo realiza mediante el análisis de los datos extraídos sometiéndolos a transformaciones, para eliminar las inconsistencias y resumir la información, con el propósito de partir de estos para tomar decisiones en función de mejorar la gestión del negocio.

Las fases de transformación, extracción y carga de los datos de estas distintas fuentes que serán integradas en la bodega de datos se consolida mediante un proceso estandarizado denominado ETL (Extracción, Transformación y Carga de datos), por medio del cual se lleva a cabo un conjunto de procedimientos necesarios para la adecuada alimentación de los datos históricos de una bodega, que luego serán transformados para finalmente ser cargados a una nueva base de datos (Sarmiento, 2012).

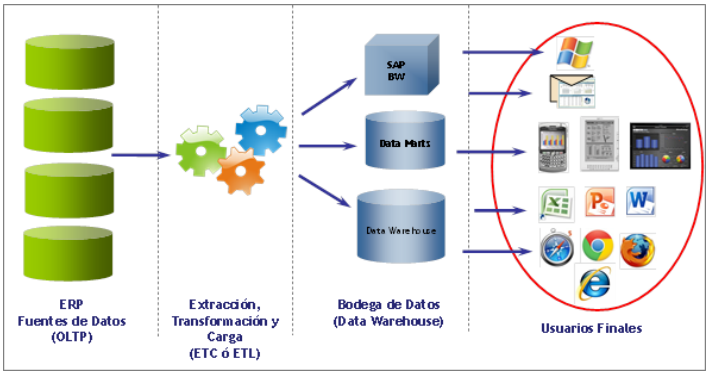


Figura 2. Proceso ETL
Fuente: (Latino BI, 2013)

Para el diseño y creación de un DW existen algunas metodologías, algunas impuestas por los fabricantes de software de inteligencia de negocios con sus productos y otras descritas por algunos autores, entre ellos las más utilizadas la Kimball e Inmon.

La diferencia entre las metodologías planteadas por estos autores es el sentido de la construcción del DW, donde Kimball propone comenzar por los *Data marts*, que son pequeños DW orientados a un área en específico del negocio, o ascendente (*Bottom-up*) e Immon indica que se trabaje con el DW desde el principio, o descendente (Top- Down).

En el presente trabajo se trabajó con la metodología Kimball que propone como tareas para el diseño e implementación de una bodega de datos los mostrados en la figura 3 (Rivadera, 2012).

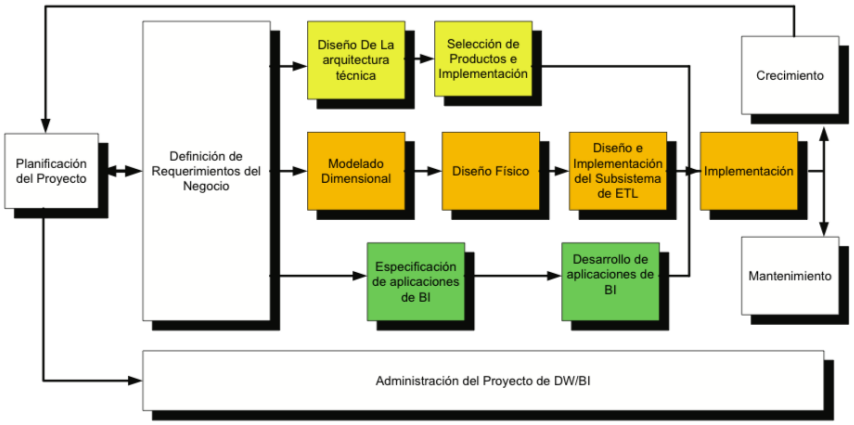


Figura 3. Tarea metodología Kimball
Fuente: (Rivadera, 2012)

MINERÍA DE DATOS: La minería de datos (DM) son métodos estadísticos que permiten determinar patrones de acuerdo a datos dados. Un DM tiene como estructura de conocimiento:

$$\text{Datos} + \text{Estadística} \rightarrow \text{Información}$$

→: Implica que los datos con la estadística bien aplicada dan como resultado información.

Mediante la minería de datos se analiza datos desde diversos factores para resumir los datos en segmentos de información útiles. Las técnicas de minería de datos permiten hacer evidentes relaciones ocultas entre sucesos para explorar grandes bases de datos, y encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos (Stark, 2016).

TÉCNICA Y ALGORITMOS PARA *DATA MINING*:

Técnicas para *Data Mining*

Redes neuronales artificiales. - Es una técnica implementada mediante aprendizaje secuencial, realizando transformaciones de datos originales para predecir un modelo. Se lo realiza analizando estadísticamente los datos para permitir la construcción de un modelo de comportamiento desde una determinada cantidad de ejemplos en función de variables descriptivas.

Árboles de decisión. - Técnica que obtiene visualmente reglas de decisión bajo las cuales se comportan los datos, desde datos históricos. Tiene una estructura en forma de árbol que representan conjuntos de decisiones.

Agrupamiento (*Clustering*). - Técnica que agrupa un conjunto de observaciones en un número dado de clusters o grupos, está basado en la idea de similitud de los grupos. El clustering no está supervisado y no requiere un set de datos de aprendizaje. Comparte un conjunto de metodologías con la clasificación. Es decir, muchos de los modelos matemáticos utilizados en la clasificación pueden ser aplicados al análisis cluster también. Usan los algoritmos de *clustering* y de *sequence clustering*.

Algoritmo Jerárquico. - Se debe calcular la distancia entre los pares de objetos o clusters, se busca los dos clusters más cercanos éstos se juntan y constituyen uno solo, se repite los pasos hasta que no quedan pares de comparación.

Regla de Inducción. – Deriva un conjunto de reglas utilizadas para clasificar casos, generan un conjunto de reglas independientes que permiten comparar árboles de decisión y patrones desde datos de entrada. La información de entrada será un conjunto de casos en que se ha asociado una clasificación a un conjunto de variables o atributos (Stark, 2016) .

Algoritmos para *Data Mining*

Los 8 más usados en el mundo son:

Regresión Lineal. - Es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes Xi y término aleatorio E.

Regresión Logística. - Con un análisis de regresión es posible poder predecir el resultado de una variable categórica en función de variables independientes. Este tipo de algoritmo es bastante usado en Ciencias Médicas y Sociales, ya que es posible modelar la probabilidad de un evento en función de otros factores.

Máquinas de soporte Vectorial. - Desarrollado por AT&T, es un algoritmo de aprendizaje supervisado, el cual resuelve problemas de clasificación y regresión, de acuerdo a puntos ubicados en el espacio.

K-Means. - Es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es el método más utilizado en *Data Mining*.

Selvas Aleatorias.- Es una mezcla de árboles predictores, en que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la (Blog, 2016, p. 8) misma distribución para cada uno de estos. *Random Forest* es de los algoritmos de aprendizaje más certero que existe y puede trabajar con una gran cantidad de datos.

Factorización de Matriz. - Es descomponer una matriz como producto de 2 o más matrices de forma canónica. Este algoritmo sirve para resolver problemas de ecuaciones lineales y cálculo de determinantes. Existen múltiples variantes, como es Factorización LU, Cholesky, Gauss-Jordan, etc.

Clasificador *Naive Bayes*. - *Bayes Naive* asume que la presencia o ausencia de una característica no está relacionada con la presencia o ausencia de cualquier otra característica.

Redes Neuronales. - Sin duda quizá el algoritmo más escuchado recientemente, es el que tiene más variantes. Una red neuronal es un grupo interconectado de nodos, simulando una red del cerebro. Esto permite resolver problemas de la misma manera que el cerebro humano.

METODOLOGÍAS DE MINERÍA DE DATOS CRISP-DM: CRISP-DM (Cross Industry Standard Process for Data Mining) es una metodología de minería de datos que proporciona una guía de referencia normalizada del ciclo de vida de un proyecto de análisis de datos.

La metodología CRISP-DM cubre las fases de un proyecto de minería de datos, sus acciones, y las relaciones entre estas acciones, que influyen en la elaboración de modelos. CRISP–DM toma en cuenta al cliente, aunque éste no sea parte del negocio, considera que un proyecto no sólo no acaba cuando se encuentra el mejor modelo, sino que requiere un despliegue y un mantenimiento, que lo relaciona con otros proyectos, y permite hacer una documentación para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

El ciclo de un proyecto de minería de datos posee seis fases como se muestra en la figura.

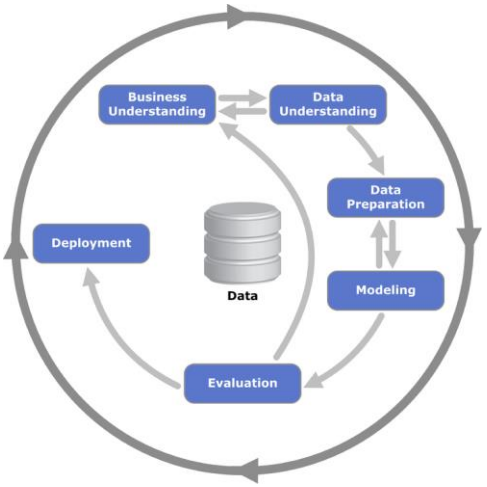


Figura 4. Fase metodología CRISP-DM
Fuente: (Villena, 2016)

El proceso de esta metodología permite hacer un movimiento hacia adelante y hacia atrás entre diferentes fases y las flechas indican las dependencias más importantes y frecuentes. Además, el círculo externo en la figura simboliza la naturaleza cíclica de los proyectos de minería de datos y como el proyecto no se termina una vez que la solución se despliega (Villena, 2016).

Fase I. Business *Understandin*g (Comprensión del negocio). - Es la fase inicial de la metodología en la que se definen las necesidades del cliente para comprender los objetivos de proyecto y así convertirlos en conocimiento. De los datos que el problema de minería de datos se construye un esquema preliminar del diseño que cumpla con los requerimientos establecidos.

Fase II. *Data Understanding* (Comprensión de los datos). - Se analizan los datos para comprenderlos y reconocer los problemas de calidad, descubrir conocimiento preliminar, y/o encontrar los interesantes que permitan formar una hipótesis de acuerdo a la información oculta de acuerdo a la colección de datos iniciales.

Fase III. *Data Preparation* (Análisis y selección de datos). - Realiza el análisis y selección de los datos de la recolección inicial para adaptarlos a las herramientas de modelado que los utilizarán. Los pasos en esta fase consisten en la selección de tablas, registros y atributos, y además en la transformación y limpieza de datos para las herramientas de modelamiento.

Fase IV. *Modeling* (Modelado). - Se encarga de escoger y utilizar las técnicas de modelado más apropiadas para el proyecto, y se calibran sus parámetros a valores óptimos. Para escoger las técnicas se consideraron los siguientes criterios:

- Referencia del problema.
- Disponer de datos adecuados.
- Resolver los requisitos del problema.

- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica

Fase V. *Evaluation*. (Evaluación de resultados obtenidos). - La fase de evaluación está orientada a la evaluación de los resultados obtenidos mediante la construcción de modelos que permitan alcanzar la calidad suficiente desde un enfoque de análisis de datos.

En esta fase se compara el modelo obtenido con los objetivos de negocio determinando argumentos importantes de negocio que no se hayan considerado. Además, en esta fase se determina si los resultados del proceso de análisis de datos deber ser aplicados o no.

Fase VI. *Deployment* (Despliegue). - Con el modelo construido y validado se procede a convertir el conocimiento que se obtuvo en actividades del proceso de negocio, donde se recomiende acciones basadas los resultados obtenidos del modelo, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso.

Generalmente, la creación del modelo no termina con la implantación del modelo, debido a que se debe documentar y presentar los resultados de manera comprensible para el usuario. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados (Villena, 2016).

2.1.2. Fundamentación de la variable dependiente

Escuela Politécnica Nacional

La Escuela Politécnica Nacional, también conocida como EPN, es una universidad pública, de grado y posgrado, ubicada en Quito (Ecuador). Reconocida por la investigación y la educación en ciencias básicas, ingenierías y tecnología, oferta programas doctorales, de maestría y de grado.

Fue fundada, en primera instancia, por el presidente Gabriel García Moreno en 18691 y, luego de su cierre en 1876, fue reabierta casi treinta años después, en la década de 1930. Desde entonces se ha convertido en uno de los centros de estudios superiores más prestigiosos del país. De hecho, la EPN fue acreditada en la categoría A, en el año 2009, por el ex Consejo Nacional de Evaluación y Acreditación de la Educación Superior del Ecuador (CONEA) y en el 2013 por el Consejo de Evaluación, Acreditación y Aseguramiento de Calidad del Sistema de Educación Superior (CEAACES) (EPN, 2018).

Su población estudiantil es alrededor de 10.000 de los cuales alrededor del 30% es femenino y el 70% es masculino ((EPN, 2017).

Escuela de Formación de Tecnólogos

La Escuela de Formación de Tecnólogos con el pasar de los años ha tenido un crecimiento en el número de estudiantes, los mismos que están registrados en el sistema SAEW de la Escuela Politécnica Nacional, por lo cual cuentan con información de calificaciones, datos socio económicos digitalizados desde el año 2010 y datos sobre tutorías académicas desde el año 2016 que se implementó.

Tabla 1

Número de estudiantes ESFOT

Período	Estudiantes
2010-1	766
2010-2	752
2011-1	731
2011-2	713
2012-1	668
2012-1A	565

2012-2A	543
2013	720
2014	788
2015	1004
2016	1240
TOTAL	**8490**

Fuente: (EPN, 2012)(EPN, 2013)(EPN, 2016)(EPN, 2017)

Factores de rendimiento académico

Actualmente se han determinado algunos factores que influyen en el desempeño académico de los estudiantes de la Escuela Politécnica Nacional como: incumplimiento de expectativas, problemas económicos, motivos personales, falta de comprensión al profesor, metodología de enseñanza inconsistentes, entre otros, pero no se le dado el seguimiento adecuado para obtener conocimiento predictivo de la misma y poder conocer el impacto o contribución que tiene cada uno de ellos en el desempeño académico del alumno.

De acuerdo a datos obtenidos de Informes de gestión de la EPN, se estima que en la Escuela de Formación de Tecnólogos dentro de los años 2010 a 2016 se ha matriculados 8490 estudiantes, de los cuales se tiene información sobre calificaciones, datos socioeconómicos y tutorías académicas que permitirán determinar patrones de comportamiento académico.

Tabla 2

Número de registros

Factores	***Número de registros***
Calificaciones	8490
Datos Socio Económicos	8490
Tutorías Académicas	1240
TOTAL	***18220***

PATRONES DE COMPORTAMIENTO

Cuando ciertas reacciones de la persona, se hacen muy frecuentes en determinados ambientes o situaciones, constituyen un patrón de comportamiento. Un patrón de comportamiento es una forma constante que tiene una persona, de pensar, sentir, reaccionar físicamente y actuar en determinada situación (González, 2016).

Un esquema de comportamiento es como el molde que se utiliza para fabricar un anillo en serie, su hechura es siempre la misma y aunque se utilice un material distinto para rellenarlo sigue ofreciendo respuestas estándar (González, 2016).

2.2. Antecedentes del estado del arte

El presente estudio del arte responde a las actividades iniciales de un SMS[1], que corresponde a los criterios de inclusión y la estrategia de búsqueda, para ello se ha revisado bibliografía en los principales repositorios académicos siguiendo las fases de:

Definición de objetivo: En esta fase se relaciona el esquema definido en la formulación del problema, donde se establecieron las preguntas de investigación. que conducen al objetivo de la investigación planteada.

Definición de los criterios de inclusión y exclusión: Los criterios de inclusión y exclusión ubican los estudios con las características apropiadas al tema planteado. Estos criterios son planteados por quien conduce la investigación y discutido con los investigadores.

En base al proyecto planteado se puedo considera como criterios de inclusión a los:

[1] Systematic Mapping Study (SMS): estudio de alcance que analiza un amplio conjunto de estudios primarios (artículos, publicaciones) para identificar qué y cuantas evidencias hay disponibles sobre un determinado tema.

- Artículos sobre la minería de datos orientados a la predicción del fracaso escolar.
- Artículos sobre predicción de deserción de estudiantes.
- Artículos sobre aplicación de minería de datos en análisis académico.
- Artículos sobre clasificación de patrones basado en indicadores de logro académico.
- Artículos sobre modelos de predicción de rendimiento académico.

Además, se puede considerar como criterios de exclusión a los:

- Artículos sobre análisis de factores humanos en la repetición académica.
- Artículos sobre análisis predictivo mediante algoritmos matemáticos.
- Artículos sobre aprendizaje de máquina.
- Artículos de minería de datos que no estén orientados a la predicción de fracaso escolar.
- Artículos que hablen sobre *big data*.
- Artículos sobre el análisis del rendimiento en el estudio de minería de datos.
- Artículos sobre técnica de aprendizaje y enseñanza.

Definición de la estrategia de búsqueda

Revisión inicial: Mediante una revisión inicial de literatura, se analizó a nivel del título, resumen, palabras clave y a groso modo el contenido de un conjunto de estudios que respondan a las preguntas de investigación planteadas, revisando si existen estudios relacionados.

Validación cruzada de estudios: La validación cruzada permite garantizar que los estudios cumplan con los criterios de inclusión y exclusión, como resultado se ha podido constatar que todos los trabajos cumplen con los criterios de inclusión y exclusión, y se ha procedido a realizar un listado de integración del grupo de control.

Integración del Grupo de Control: El grupo de control conforman los estudios que cumplen con las características de la investigación analizado por los investigadores considerando el título del estudio, resumen y palabras claves. Los estudios del grupo de control analizados son los siguientes:

Tabla 3

Estudios por Grupo de Control

Grupo Control	Estudio	Repositorio	Palabras clave
EC1	*Predicting student's achievement based on motivation in vocational school using data mining approach*	*IEE EXPLORE*	*Data mining; educational institutions; human factors; pattern classification; academic achievement indicator; classification algorithm*
EC2	*Modeling Academic Achievement of UUM Graduate Using Descriptive and Predictive Data Mining.*	*Springer, Cham*	*Modeling Academic Achievement ; academic analysis; data Mining ; undergraduate students.*
EC3	*A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes.*	*ACM New York*	*Evaluation metrics; applications; education; risk prediction*
EC4	*Generating descriptive model for student dropout: a review of clustering approach.*	*Scopus*	*Clustering; Dropout; Educational data mining; Retention Student performance*

Continúa

EC5	*Analyzing undergraduate students' performance using educational data mining.*	*ScienceDirect*	*Data mining; Decision trees; Clustering; Performance prediction; Performance progression*
EC6	*Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout.*	*ScienceDirect*	*E-learning; Student dropout prediction; Educational data mining; Logistic regression model; Temporal data; Student dropout prevention*

Construcción de la cadena de búsqueda: Para la construcción de la cadena de búsqueda se analiza los estudios del grupo de control, encontrando palabras comunes entre estudios y las palabras propias que están direccionadas al objetivo de la presente investigación, para ello se formaron los siguientes contextos: minería de datos, políticas educativas e indicadores.

Tabla 4

Construcción de cadenas de búsqueda

Contexto	Palabras claves	EC1	EC2	EC3	EC4	EC5	EC6	Palabra repetida
Mineria de datos	Data Mining	X	X	X		X		4
	pattern classification	X						1
	classification algorithm	X						1
	Educational data mining				X		X	2
	Clustering				X	X		2
	Decision trees					X		1
Politicas educativas	Educational institutions	X						1
	human factors	X						1
	Retention Student performance		X		X			2
	undergraduate students		X		X			2
	Education or Educational Applications			X				1
	Dropout				X			1
	Quality of educational processes					X		1
Indicadores	academic analysis	X	X					2
	academic achievement indicator		X					1
	Predicting academic performance					X		1
	evaluation metrics			X				1
	risk prediction			X				1
	Student dropout prediction						X	1

La cadena de búsqueda se forma con la combinación de las palabras que más se repiten en cada contexto para unir en la cadena se utilizó el conector OR y para concatenar con el contexto se usa el conector AND, estableciendo la siguiente cadena de búsqueda:

((Data mining) AND (academic analysis) AND ((undergraduate students) OR (retention student performance)))

Esta cadena de búsqueda fue aplicada en el repositorio digital SCOPUS, realizando configuraciones como: idioma, fuentes de publicaciones fiables, año de publicación (2016-2018), estudios duplicados, área de interés, tipo de publicación (revistas y congresos), entre otros verificando coincidencias y discrepancias respecto a los criterios de inclusión y exclusión para determinar los 8 estudios primarios relacionados a la problemática planteada y el uso de minería de datos como una de las posibles soluciones.

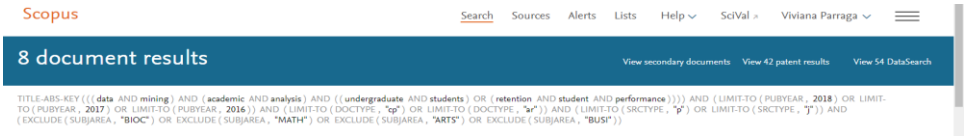


Figura 5. Cadena de búsqueda ejecutada en SCOPUS

Los estudios primarios encontrados se detallan a continuación:

Learning analytics for smart campus: Data on academic performances of engineering undergraduates in Nigerian private university

En este artículo se analizan datos académicos del desempeño de los estudiantes de pregrado seleccionados al azar. Se utiliza estadística descriptiva y distribuciones de frecuencia de los datos de rendimiento académico en tablas y gráficos para facilitar la interpretación de los datos. Además, se realizan análisis de varianza de una vía (ANOVA) y pruebas de comparación múltiple post hoc para determinar si las variaciones en los rendimientos académicos son significativas. Los datos proporcionados en este artículo ayudarán a la comunidad de investigación educativa global y a los

responsables de la política regional a comprender y optimizar el entorno de aprendizaje hacia la realización de campus inteligentes y la educación sostenible (Popoola et al., 2018) .

Predicting Critical Courses Affecting Students Performance: A Case Study

En este documento, se utiliza el algoritmo de inducción del árbol de decisión ID3 para construir modelos de predicción para el rendimiento académico. Los modelos se basan en los registros para mujeres estudiantes en el programa de Licenciatura en el departamento de Tecnología de la Información (TI) de la Universidad King Saud, Riyadh, Arabia Saudita. Los resultados indican que se pueden lograr predicciones confiables basadas en el rendimiento de los estudiantes. También se identificó cursos clave que pueden usarse como predictores de rendimiento. Se cree que los hallazgos encontrados son útiles para los responsables de la toma de decisiones en el departamento de TI (Altujjar, Altamimi, Al-Turaiki, & Al-Razgan, 2016).

Analyzing undergraduate students' performance using educational data mining

Este estudio, utiliza métodos de minería de datos para estudiar el rendimiento de los estudiantes de pregrado. Se han enfocado dos aspectos del rendimiento de los estudiantes. Primero, predecir el rendimiento académico de los estudiantes al final de un programa de estudio de cuatro años. Segundo, estudiando las progresiones típicas y combinándolas con los resultados de predicción. Se han identificado dos grupos importantes de estudiantes: los estudiantes de bajo y alto rendimiento. Los resultados indican que al centrarse en un pequeño número de cursos que son indicadores de un rendimiento particularmente bueno o malo, es posible brindar una advertencia y apoyo oportunos a los estudiantes de bajo rendimiento, y consejos y oportunidades para los estudiantes de alto rendimiento (Asif, Merceron, Ali, & Haider, 2017).

Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department

Este documento discute la construcción de un modelo para predecir el rendimiento de los estudiantes en un curso de programación basado en sus calificaciones en cursos en otras materias. Una clasificación basada en un algoritmo de reglas de asociación se utiliza para construir un clasificador que ayude a evaluar el rendimiento del alumno en el curso de programación. Este modelo tiene como objetivo reducir los niveles de deserción al ayudar al alumno a predecir su probabilidad de éxito en un curso antes de inscribirse en él. Además, los instructores del curso podrán mejorar el rendimiento del alumno en el curso al estimar mejor sus habilidades para aprender el tema y ajustar sus estrategias y métodos de enseñanza (Badr, Algobail, Almutairi, & Almutery, 2016)

Using data mining techniques with open source software to evaluate the various factors affecting academic performance: A case study of students in the faculty of information technology

Este documento de investigación estudia los diferentes factores que podrían afectar los promedios acumulativos de los estudiantes de la Facultad de Informática en las universidades jordanas, verificando la información, y los antecedentes académicos de los estudiantes. También tiene el objetivo de revelar cómo esta información afectará a los estudiantes para obtener altas calificaciones en sus cursos. Se utiliza un software libre de código abierto (WEKA) que admite herramientas y técnicas de minería de datos para decidir qué atributo (s) afectará los promedios acumulativos de los estudiantes. Se encontró que el factor más importante afecta los promedios acumulativos de los estudiantes, es el tipo de aceptación del estudiante. Un modelo y reglas de árbol de decisión también se construyen para determinar cómo los estudiantes pueden obtener altas

calificaciones en sus cursos. La precisión general del modelo fue del 46.8%, que es una tasa aceptada (Alharbi, Cornford, Dolder, & De, 2016).

Generating descriptive model for student dropout: a review of clustering approach

La implementación de la minería de datos es ampliamente considerada como un poderoso instrumento para adquirir nuevos conocimientos a partir de una pila de datos históricos, que normalmente no se estudian. Esta metodología impulsada por los datos ha demostrado su eficacia para mejorar la calidad de la toma de decisiones en varios ámbitos, como problemas empresariales, médicos y de ingeniería compleja. Recientemente, la minería de datos educativos (EDM) ha obtenido una gran atención entre los investigadores educativos y los científicos informáticos. En general, las publicaciones en el campo de la EDM se centran en la comprensión de los tipos de estudiantes y el marketing dirigido, utilizando modelos descriptivos y predictivos para maximizar la retención de los estudiantes. Inspirado en intentos previos, este documento tiene como objetivo establecer el enfoque de agrupamiento como una guía práctica para explorar categorías y características de los estudiantes, con el ejemplo de trabajo en un conjunto de datos reales para ilustrar los procedimientos analíticos y los resultados (Iam-On & Boongoen, 2017a).

Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings

El aumento de la retención de estudiantes ha sido un objetivo común de muchas instituciones académicas, especialmente a nivel universitario. Con esta idea, se han desarrollado varios métodos de minería de datos para la detección temprana de estudiantes en riesgo de deserción, de ahí la aplicación inmediata de medidas de asistencia. Este artículo presenta la investigación más reciente sobre el abandono escolar en la Universidad Mae Fah Luang, Tailandia, y la nueva reutilización del conjunto de clúster basado en enlaces como un marco de transformación de datos para una

predicción más precisa. El estudio empírico sobre la recopilación de datos de tipo mixto relacionado con los detalles demográficos de los estudiantes, el rendimiento académico y el registro de inscripciones, sugiere que el enfoque propuesto suele ser más eficaz que varias técnicas de transformación de referencia, en diferentes clasificadores (Iam-On & Boongoen, 2017b).

Modeling academic achievement of UUM graduate using descriptive and predictive data mining

En la institución educativa superior, la minería de datos se puede utilizar para el proceso de descubrir tendencias ocultas y patrones que ayudan a las instituciones a pronosticar el rendimiento de los estudiantes. El propósito de este estudio es investigar los factores asociados con el rendimiento académico de los estudiantes de pregrado utilizando minería de datos descriptivos y predictivos. Las investigaciones previas indican que los estudiantes y la facultad comparten una percepción común de las habilidades necesarias para el éxito en los programas de grado. Con base en los resultados extraídos de la extracción de datos descriptivos y predictivos, la investigación empírica mediante redes logísticas y neuronales revela que factores como el ingreso familiar, la raza y las habilidades lingüísticas tienen una asociación significativa con el rendimiento académico (Siraj, 2016).

El estado del arte es una modalidad de la investigación documental que permite el estudio del conocimiento escrito en textos dentro de un área específica, permitiendo que la circulación de la información, genere una demanda de conocimiento y establezca comparaciones con otros conocimientos paralelos a este, ofreciendo diferentes posibilidades de comprensión del problema tratado; pues brinda más de una alternativa de estudio (Montoya, 2005).

En conclusión, de acuerdo al estudio del arte revisado, se indica que la minería de datos en el análisis del ámbito educativo es un tema muy extenso y cubierto por algunos artículos como lo

revela este estudio, se puede observar como la minería de datos es aplicada para predecir el rendimiento escolar de los estudiantes y así disminuir índices de deserción de éstos, puesto que los estudios consideran que al descubrir el conocimiento oculto de un gran volumen de datos educativos y aplicarlo adecuadamente para la toma de decisiones se puede garantizar una educación de alta calidad en cualquier institución académica, además que, un análisis de minería de datos se puede utilizar para destacar los problemas de rendimiento desde el principio y proponer acciones correctivas. Además, estos estudios muestran técnicas de minería de datos utilizadas, tales como: regresión logística, árboles de decisión, bosques aleatorios, Naive Bayesian para explorar datos de entornos educativos.

En consecuencia, el presente estudio contribuirá con generación de conocimiento sobre la problemática planteada en la identificación de factores que influyen en el aprovechamiento académico de los estudiantes que de soporte a la toma de decisiones, evitando el bajo rendimiento académico y con esto el rezago estudiantil, mediante una perspectiva más integral desde el proceso de captura de información, consolidación automática y así proponer un modelo de pronóstico basándose en el análisis y uso de técnicas de minería de datos contribuyendo con nuevos enfoques de análisis a los estudios actuales, considerando las características de la gestión educativa.

2.3. Marco conceptual

En base a los requerimientos del proyecto, se procedió a seleccionar las herramientas.

POWER DESIGNER

Herramienta de modelamiento propiedad de la empresa SAP[2], que permite realizar el análisis, diseño inteligente, visualización y manipulación de metadatos para la construcción sólida de bases de datos.

Brinda un enfoque orientado a modelos a nivel físico y conceptual, facilitando el proceso de implementación de arquitecturas efectivas de información basándose en tecnologías actuales.

Está compuesto de algunas técnicas básicas de modelamiento mediante herramientas de desarrollo, como *Sybase Powerbuilder*, Eclipse, Java, .NET y *Sybase WorkSpace*. Proporciona respuestas de diseño y análisis de bases de datos («PowerDesigner - EcuRed», s. f.).

Mediante esta herramienta se diseñó el modelo entidad relación y multidimensional de acuerdo a los datos entregados para posteriormente utilizarlos en la implementación de la bodega de datos.

MYSQL (MY STRUCTURED QUERY LANGUAGE)

Sistema de gestión de base de datos que permite crear base de datos y tablas, insertar datos, modificarlos, eliminarlos, ordenarlos, hacer consultas y realizar muchas operaciones (Parraga, 2015).

También MySQL es conocida por brindar alta velocidad y simplicidad al momento de desarrollar búsqueda de datos. Es multiplataforma puesto que trabaja en Mac, Windows, Linux, BSD, Open Solaris, Perl y Phyton entre otras (Parraga, 2015).

[2] SAP (Sistemas, Aplicaciones & Productos en Procesamiento de Datos") es una empresa multinacional de software europea con sede en Alemania.

MySQL se puede utilizar de forma libre y es de código fuente abierta, pero si se desea un uso más autónomo, servicios y soportes extras, se puede adquirir una licencia, debido a que MySQL es patrocinado por un grupo privado (Parraga, 2015)..

MySQL se considera una herramienta orientada a la gestión de las bases de datos asociadas a aplicaciones web. Es por esto que forma parte del popular servidor XAMPP (Parraga, 2015).

En vista de todas las ventajas que brinda este gestor de base de datos, se lo utilizó para almacenar aquí todos los datos luego de haberlos adecuado en el proceso ETL para después ser utilizado por la herramienta de minería de datos.

DATA INTEGRATION DE PENTAHO

Pentaho es una suite de herramientas de Inteligencia de Negocios de código abierto, programadas en código 100% Java. Dentro de las que destaca un producto para la integración de datos llamado PDI.

Brinda un conjunto amplio de herramientas que facilitan la integración de información para el análisis inteligente de los datos, puesto permite una rápida y eficiente extracción, transformarlos, limpieza, validación, carga, entre otros de los mismos sin importar donde se encuentren.

Ofrece grandes capacidades en el proceso de manejo de procesos ETL, además, análisis multidimensionales de información e informes interactivos.

Según los requerimientos del usuario, las herramientas permiten utilizar los módulos de manera conjunta o independientemente. Transforma e integra datos entre sistemas de información existentes y los *Data marts* que compondrán el sistema BI. Se compone de 4 módulos:

SPOON: Que brinda una interfaz gráfica de diseño de transformaciones.

PAN: Que permite ejecutar transformaciones diseñadas con *Spoon* utilizando línea de comandos.

CARTE: para ejecutar trabajos y transformaciones en un servidor Web de forma remota (López & Galindo, 2013).

Luego de realizar la comparación de algunas herramientas ETL se consideró era la más adecuada para realizar el proceso de extracción, transformación y carga de los datos del proyecto.

RAPIDMINER

Esta herramienta de software libre y de código abierto que forma parte del proyecto Rapid-i que cuenta con dos componentes:

RapidMiner: Versión *stand-alone* para analistas. Implementa todos los operadores de *data mining*, modelos predictivos, modelos descriptivos, transformación de datos, series de tiempo, etc.

RapidAnalytics: Versión servidor de RapidMiner. Permite trabajo colaborativo, escalable y concurrente de múltiples usuarios, capacidad de delegar en bases de datos (In-*Database Mining*).

RapidMiner es una herramienta para desarrollar análisis de datos utilizando procesos encadenados a través de operadores visualizados en un entorno gráfico con herramientas de visualización de datos. Esta herramienta multiplataforma está desarrollada en Java y permite a los experimentos componerse de un gran número de operadores anidables arbitrariamente.

Además, esta aplicación presenta un conjunto de más de operadores que permiten realizar procedimientos principales de máquina de aprendizaje, combinando esquemas de aprendizaje y evaluadores de atributos del entorno de aprendizaje Weka (Garcia, 2013).

Tomando como referencia los beneficios que esta herramienta ofrece, dentro del proyecto fue utilizada para el análisis de los diferentes algoritmos de minería de datos, para la creación del modelo y para la validación del mismo.

CAPÍTULO III

MEMORIA TÉCNICA METODOLÓGICA

3.1. Metodología de Investigación

La investigación de este trabajo se elaboró y ejecutó utilizando la metodología de investigación científica basada en el diseño (Gonzalez & Pomares Quimbaya, 2012).

Esta metodología está enfocada a las ciencias de la computación y los sistemas de información, en los que se pueden obtener contribuciones científicas o técnicas genéricas, pero de utilidad borrosa o no declarada y aquellas orientadas a la relevancia que son útiles, pero no suficientemente formales, transparentes o validadas con criterio científico (Gonzalez & Pomares Quimbaya, 2012).

Se plantea utilizar la investigación científica basada en el diseño que busca hacer aportes significativos en un área del conocimiento y a la vez dar solución a problemas relevantes, utilizando el análisis de problemas aún no resueltos en un ambiente del mundo real y su resolución de una manera novedosa y rigurosa a través del diseño de artefactos (Winter, Zhao, Aierσ, & DESRIST, 2012).

Esta investigación se distingue de la investigación aplicada por el énfasis en la construcción de artefactos innovadores y la retroalimentación que proveen para revisar, extender o re-contextualizar los fundamentos teóricos.

Algunos autores han propuesto una arquitectura general para proyectos de investigación centrados en diseño, mediante la articulación de tres ciclos: rigor, relevancia y diseño, se puede notarlo en la figura 6.

Figura 6. Investigación mediante ciclos
Fuente: (Gonzalez & Pomares Quimbaya, 2012)

Los tres ciclos consisten en primero integrar el entorno, lugar donde se dará solución al problema utilizando los requerimientos de personas, tecnología y organización para que mediante la construcción y evaluación de un diseño se contribuya a la base del conocimiento con teorías, modelos, métodos, experiencia y artefactos existentes.

Se debe tomar en cuenta que este es un proceso cíclico en cada una de sus fases y de manera general y que si un ciclo se está moviendo los otros dos lo harán a la vez, no como fases de investigación (Gonzalez & Pomares Quimbaya, 2012).

Además, se utilizaron los tipos de investigación experimental y descriptiva. La experimental permitió relacionar las variables de estudio con la que se determinó las herramientas y modelos necesarios para hacer la minería de datos. Y la descriptiva recopiló datos relevantes de estudios posteriores sobre el diseño de minería de datos.

3.2. Ejecución del proceso de investigación

El proceso de investigación se realizó haciendo referencia a la metodología de investigación seleccionada como es la investigación científica basada en el diseño.

Mediante este proceso se tuvieron tres ciclos:

ENTORNO: Se realizó el análisis de la información existente en las bases de datos de la Escuela Politécnica Nacional, realizando el diseño de la bodega de datos que cumpla con los propósitos del trabajo, utilizando la herramienta Power Designer para realizarlo. Además, se determinó la herramienta ETL, para hacer la limpieza y transformación de la misma, seleccionando la herramienta Data Integration de Pentaho, por ser una herramienta versátil para el diseño y creación de una bodega de datos. También, se identificaron los algoritmos de minería de datos utilizados en el modelo diseñado utilizando la herramienta de minería de datos RapidMiner.

Obtenidos los requerimientos para el desarrollo del proyecto se pasa a la segunda fase:

DISEÑO: Fase en la que se creó el modelo de minería de datos mediante los requerimientos definidos en la fase de entorno para determinar los patrones que afectan el desempeño de los alumnos. Dicho modelo fue evaluado a través de la técnica de validación matriz de confusión, que determino un nivel de confianza en un resultado predictivo encontrado.

Además, aquí se utilizó la metodología de minería de datos CRISP-DM que proporcionó una guía de referencia normalizada del ciclo de vida de un proyecto de análisis de datos, mediante seis fases:

Fase I. *Business Understanding* (Comprensión del negocio)

En la Escuela de Formación de Tecnólogos se han determinado algunos factores que influyen en el desempeño académico de los estudiantes como: incumplimiento de expectativas, problemas económicos, motivos personales, falta de comprensión al profesor, metodología de enseñanza inconsistentes, entre otros, pero no se le dado el seguimiento adecuado para obtener conocimiento predictivo de la misma y poder conocer el impacto o contribución que tiene cada uno de ellos en el desempeño académico del alumno. Dicha institución con el pasar de los años ha tenido un crecimiento en el número de estudiantes, los mismos que están registrados en el sistema SAEW de

la Escuela Politécnica Nacional, por lo cual se cuenta con información de calificaciones, datos socio económicos digitalizados desde el año 2010 y datos sobre tutorías académicas desde el año 2016 (que se implementó) de los mismos.

Fase II. *Data Understanding* (Comprensión de los datos)

Se analizó los datos obtenidos del sistema SAEW otorgados por la Dirección de Gestión de la Información y Procesos (DGIP) de la EPN de los estudiantes de la Escuela de Formación de Tecnólogos de los últimos 8 períodos ordinarios (2013-B hasta 2017-A) sobre:

- Datos Generales: Código Anonimizado, Edad (actual), Sexo, Lugar Nacimento (Pais, Provincia, Canton), TipoColegio.
- Datos de Problemas Académicos (Tutorías Académicas): Código Anonimizado, Con o sin Factores, Factores, Otros Factores, PeríodoAcadémico, Facultad, Carrera.
- Datos Socioeconómicos: Código Anonimizado, Período Académico, Quintil, #Miembros Núcleo Familiar, Total Ingresos, Ocupación Estudiante, Remuneración Estudiante, Ocupación Cónyuge, Remuneración Cónyuge, Ocupación Padre, Remuneración Padre, Ocupación Madre, Remuneración Madre.
- Calificaciones: Código Anomizado, Nombre Materia, Tipo de Aprobación Materia, Creditos, Calificacion 1, Calificación 2, Calificacion 3, Calificación Final, Aprueba/Falla, Facultad, Carrera, Período Académico.
- Becas: Código Anonimizado, Período Académico, TipoBeca, Facultad, Carrera.

Fase III. *Data Preparation* (Análisis y selección de datos)

La información entregada por la DGIP se encontraba en archivos de Excel, las cuales no se encontraban en las condiciones para ser analizadas, debido a que se encontraban en archivos

diferentes, en diferentes hojas de cálculo, con diferentes formatos, con uso de letras ñ y tildes, entre otros, así que se realizó un proceso ETL. Este proceso permitió organizar los datos desde las diferentes fuentes, reformatearlos y cargarlos en otra base de datos con el objeto de analizarlos.

Fase IV. *Modeling* (Modelado)

Con la bodega de datos obtenido del proceso ETL en una nueva base de datos, se procedió a realizar el respectivo modelo de minería de datos, comparando, analizando y buscando el modelo que mejor resultados entregue a la información de los estudiantes. Esta es la fase medular ya que permitió determinar la relación existente entre las variables y el atributo a predecir. También, se seleccionó una técnica de modelado.

Fase V. *Evaluation.* (Evaluación de resultados obtenidos)

Con el modelo creado, en esta fase se realizó una revisión de forma técnica del mismo, con respecto a la precisión y porcentaje de error de acuerdo a los factores establecidos. Aquí se evaluó el nivel de satisfacción de los objetivos del proyecto. Las tareas realizadas aquí fueron evaluar los resultados, revisar el proceso de minería de datos para finalmente decidir que el proyecto de minería de datos concluyó y entrar en la fase de despliegue.

Fase VI. Deployment (Despliegue)

En esta fase se estructuro las recomendaciones para aplicar el modelo desarrollados. Considerando tareas de planificación, monitoreo y mantenimiento del modelo de minería de datos con el propósito de mejorar sus prestaciones. Finalmente, se genera un reporte final del proyecto en relación a evaluar lo que ocurrió correctamente y lo que necesita ser mejorado.

Con el diseño realizados y evaluador se realiza la contribución de acuerdo a la metodología donde es posible pasar a la última fase:

BASES DE CONOCIMIENTO: En esta fase se observó los resultados obtenidos en la investigación como es el modelo de minería de datos de los estudiantes de la Escuela de Formación de Tecnólogos de la Escuela Politécnica Nacional, para identificar los patrones de comportamiento del alumnado que influyen en el aprovechamiento académico, de acuerdo a lo planteado en el objetivo general.

CAPÍTULO IV

RESULTADOS

4.1 Informe de Resultados

4.1.1. Análisis y selección de datos

Los registros históricos de datos académicos, problemas académicos, socioeconómicos y becas, registrados en las bases de datos de la EPN, otorgados por la DGIP de los último 8 semestres, fueron la principal fuente de datos para llevar a cabo esta investigación.

Las variables a estudiar, por alumno, son: lugar de nacimiento, tipo de colegio del que proviene, con o sin factores de rendimiento, quintil, número de miembros núcleo familiar, total ingresos, ocupación estudiante, remuneración estudiante, ocupación cónyuge, remuneración cónyuge, ocupación padre, remuneración padre, ocupación madre, remuneración madre, calificación final, aprueba/falla, tipo de beca.

La selección de las variables a utilizar se basó principalmente en los factores disponibles en las bases de datos de la EPN considerados como relevantes para el estudio de la deserción escolar.

Con los datos de alumnos obtenidos se buscó encontrar los principales indicadores aplicando correctamente los diferentes algoritmos de minería de datos para poder validar que las variables a utilizar sean las correctas.

Estas variables fueron entregadas agrupadas en 5 grupos con una serie de atributos cada una, que serán útiles para la creación de la bodega de datos, descritos en la Tabla 5.

Tabla 5

Datos entregados por la DGIP

Datos Generales	Código Anonimizado
	Edad (actual)
	Sexo
	Lugar Nacimento (Pais, Provincia, Canton)
	TipoColegio.
Datos de Problemas Académicos	Código Anonimizado
(Tutorías Académicas)	Con o sin Factores
	Factores
	PeríodoAcadémico
	Facultad
	Carrera.
Datos Socioeconómicos	Código Anonimizado
	Período Académico
	Quintil
	#Miembros Núcleo Familiar
	Total Ingresos
	Ocupacion Estudiante
	Remuneración Estudiante
	Ocupacion Cónyuge
	Remuneración Cónyuge
	Ocupacion Padre
	Remuneración Padre
	Ocupacion Madre
	Remuneración Madre.

Continúa

Calificaciones	Código Anomizado
	Nombre Materia
	Tipo de Aprobación Materia
	Creditos
	Calificacion 1
	Calificación 2
	Calificacion 3
	Calificación Final
	Aprueba/Falla
	Facultad
	Carrera
	Período Académico.
Becas	Código Anonimizado
	Período Académico
	TipoBeca
	Facultad
	Carrera

4.1.2. Preparación de los datos

El análisis y selección de los datos fue el inicio de la creación de la bodega de datos, puesto que con éste se realizó el diagrama entidad relación que se muestra en la figura, que fue la base para la creación del modelo multidimensional.

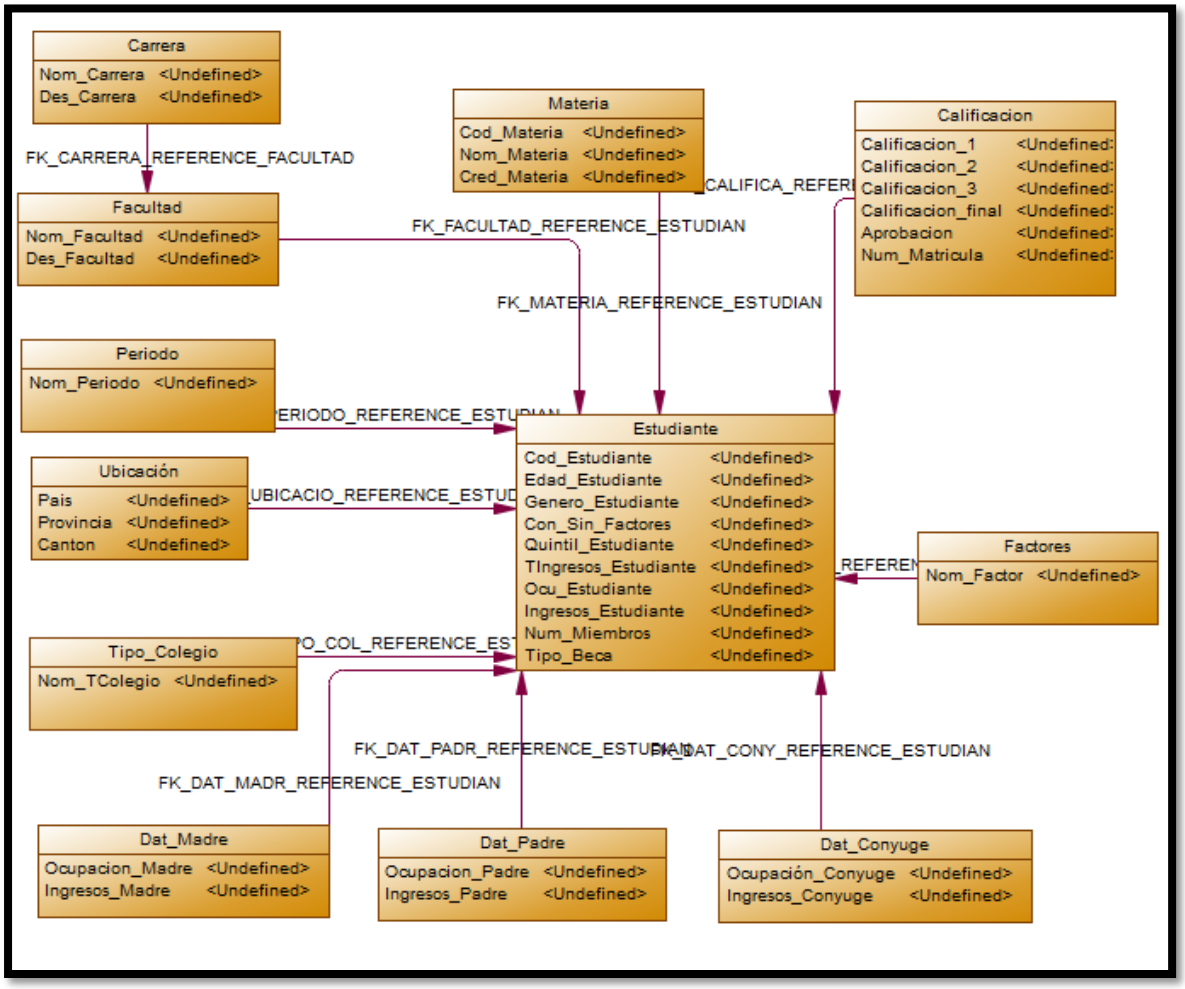


Figura 7. Diagrama Entidad – Relación

Con el propósito de obtener una vista viable que permita construir un modelo enfocado al objetivo del proyecto, se realizó el diagrama del modelo multidimensional como se muestra en la Figura 8, mediante el cual se creó la bodega de datos.

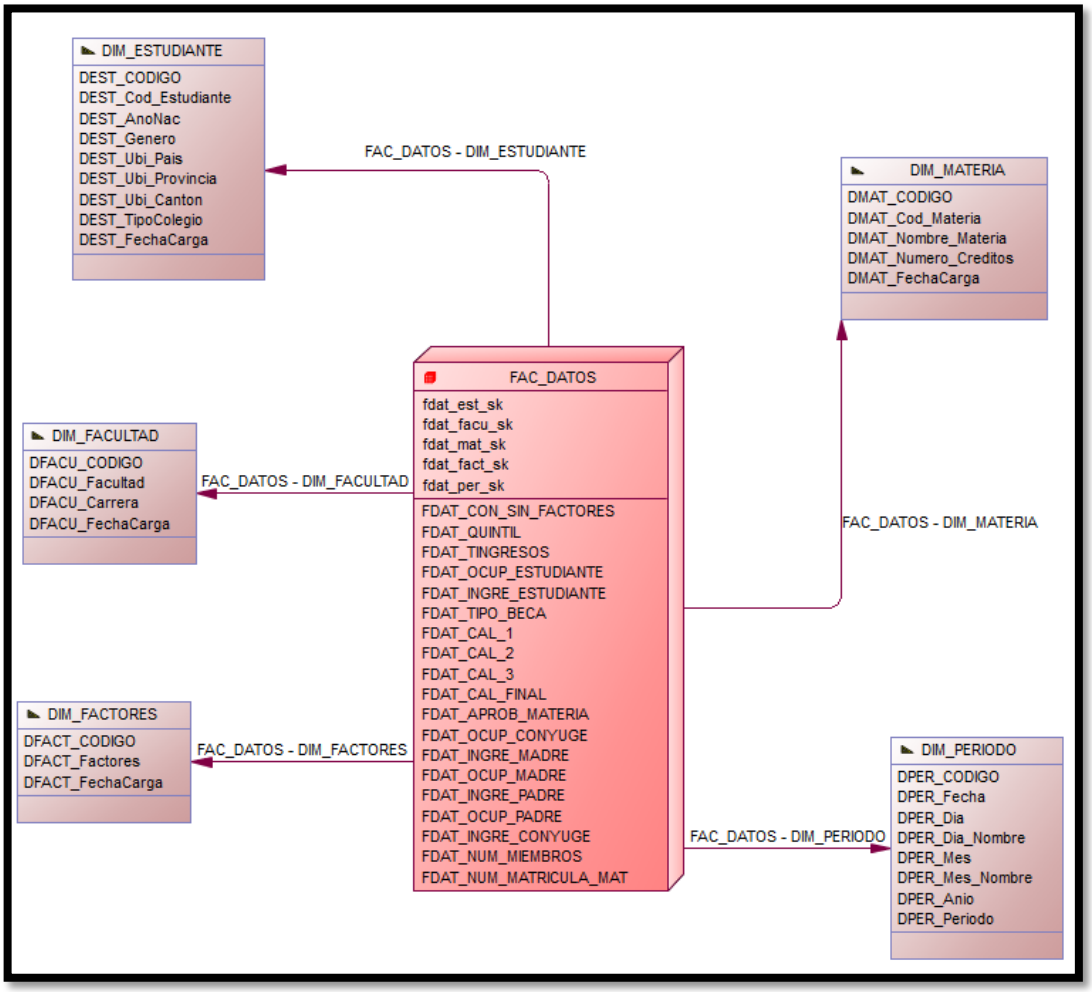


Figura 8. Diagrama modelo multidimensional

La bodega de datos está compuesta por 5 dimensiones y una tabla de hechos descrito a continuación:

DIMENSIÓN ESTUDIANTE (DIM_Estudiante). - Tabla en la cual se almacenó toda la información general y propia de cada uno de los estudiantes, que no cambia por facultad, carrera, período entre otros, de los datos otorgados.

DIMENSIÓN FACULTAD (DIM_Facultad). - Tabla que tiene toda la información referente a facultades y carrera en las cuales los estudiantes se han matriculado en los períodos académicos de estudio, dados por la DGIP.

DIMENSIÓN FACTORES (DIM_Factores). - Tabla poblada con todos los factores encontrados en los datos de factores académicos registrados por los profesores en el sistema SAEW de cada estudiante, entregados por la EPN.

DIMENSIÓN MATERIA (DIM_Materia). - Tabla en la que se almacenan todas las asignaturas que poseen calificaciones de los estudiantes, de acuerdo a los datos analizados.

DIMENSIÓN PERIODO (DIM_Periodo). - Tabla que hace referencia a la línea del tiempo, dentro de los datos otorgados, puesto a través de esta dimensión se puede diferenciar los datos de acuerdo al período académico.

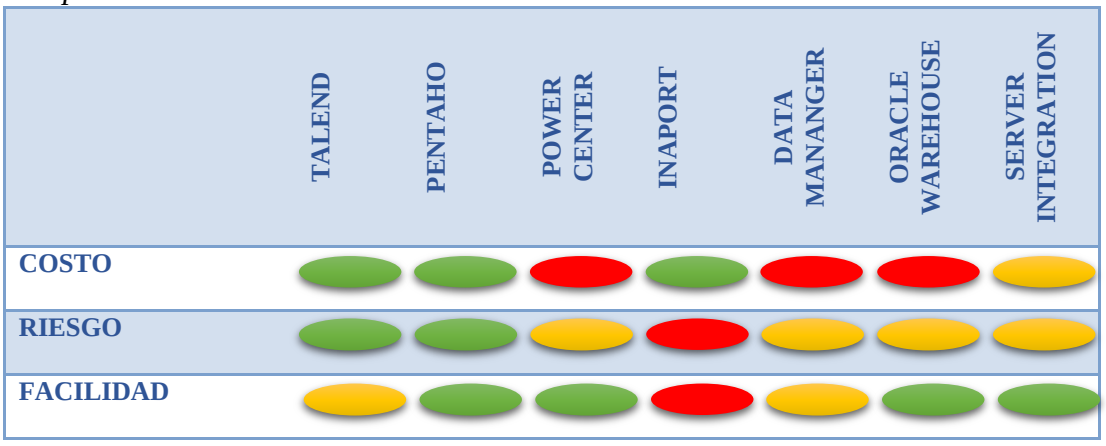
TABLA DE HECHOS (FAC_DATOS). - Tabla central en la que se almacenan todos los datos cambiantes y en su mayoría cualitativos, que permitirán definir las variables a analizar para determinar el comportamiento de los estudiantes.

4.1.3. Proceso ETL

Primero se seleccionó la herramienta ETL para realizar el proceso de extracción, transformación y carga, mediante la comparación de algunas herramientas que brindaban este servicio.

Tabla 6

Comparación de herramientas ETL

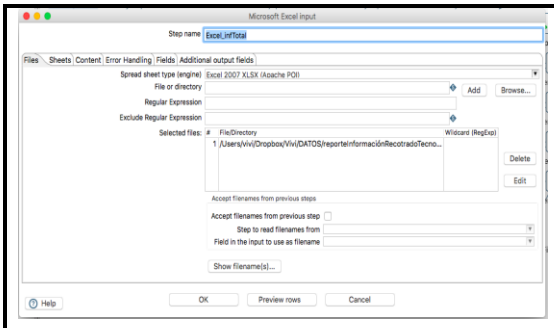
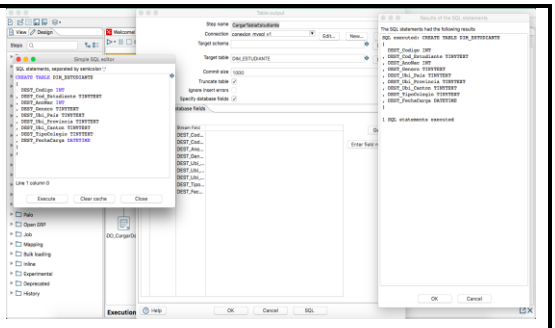


Continúa

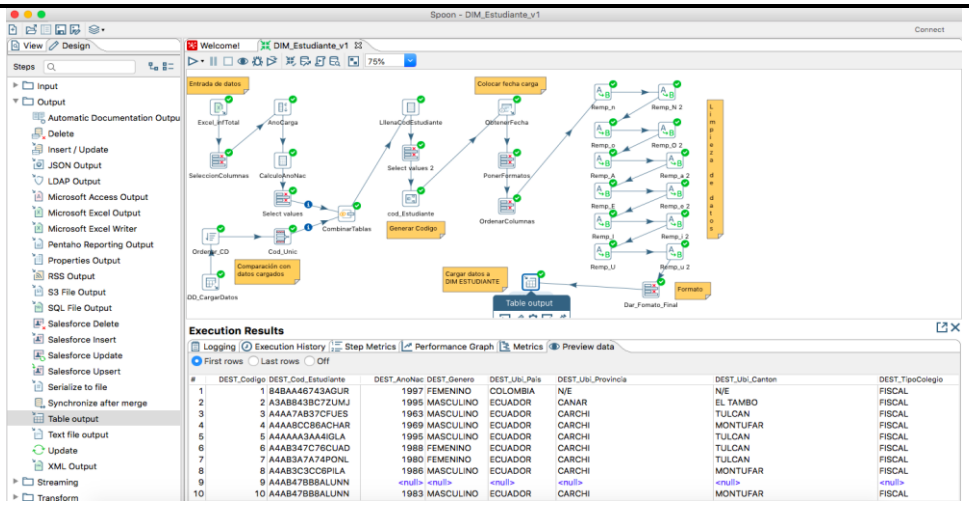
SOPORTE	
IMPLEMENTACIÓN	
VELOCIDAD	
CALIDAD DATA	
MONITOREO	
CONECTIVIDAD	

Fuente:(Jorge Bustillos, 18:53:28 UTC)

Luego se prosiguió con el proceso de extracción desde cada uno de los archivos de Excel, transformación y carga (ETL), seleccionando los datos útiles para la investigación y realizando la respectiva limpieza y transformación de los mismos. Para esto se seleccionó la herramienta Data Integration de Pentaho, mediante la cual se realizó el proceso ETL de cada una de las dimensiones del modelo multidimensional, cargando los datos a la nueva base de datos.

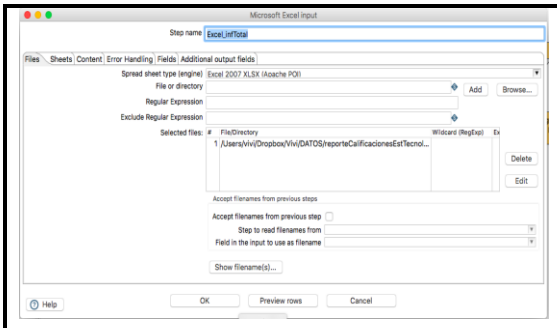
	
Fuente origen	Fuente destino

Continúa

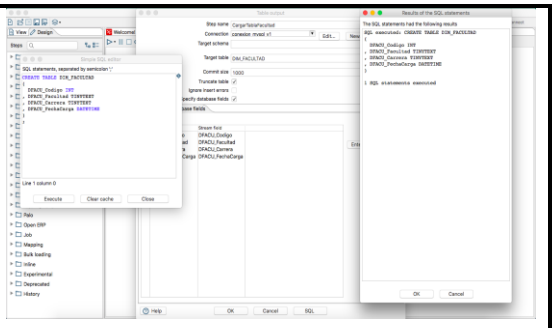


Transformación y carga

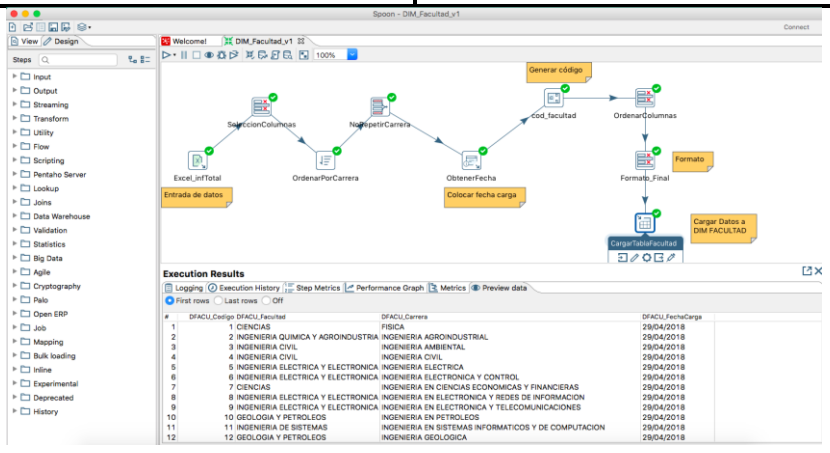
Figura 9. Proceso ETL dimensión estudiante



Fuente origen

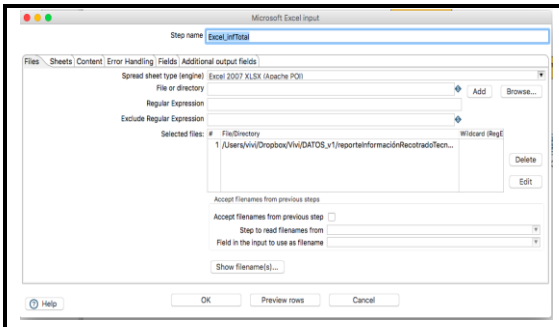


Fuente destino

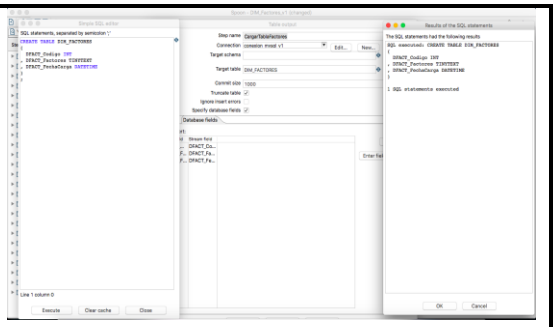


Transformación y carga

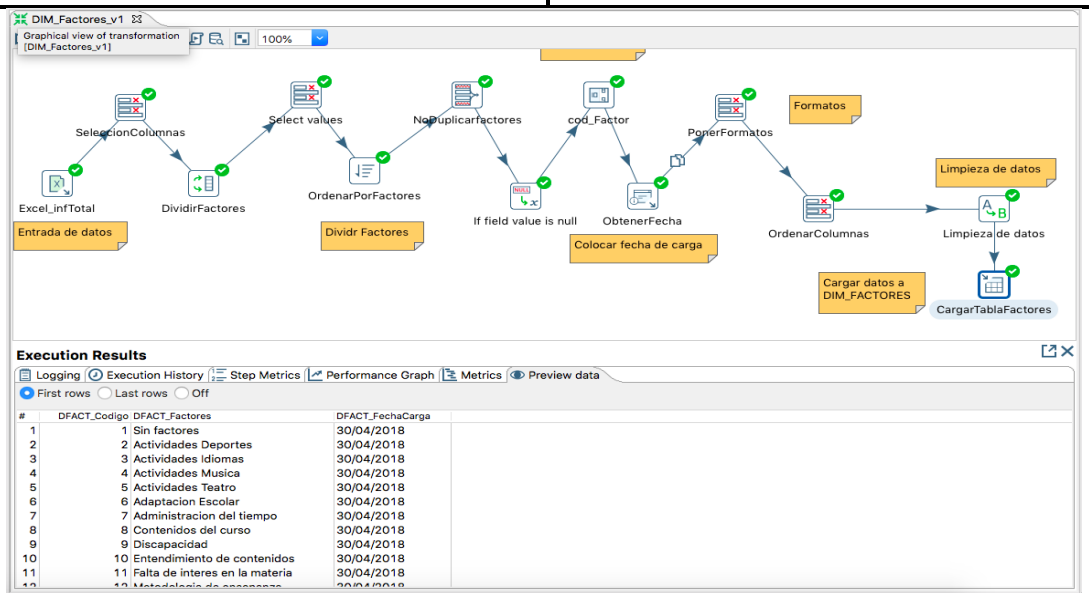
Figura 10. Proceso ETL dimensión facultad



Fuente origen



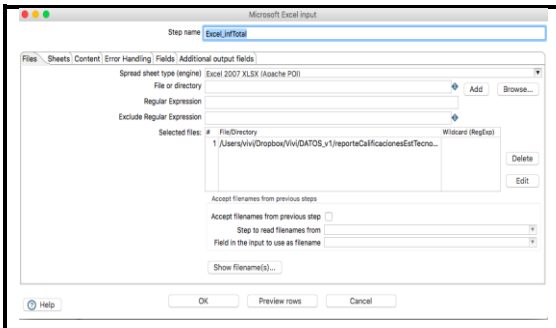
Fuente destino



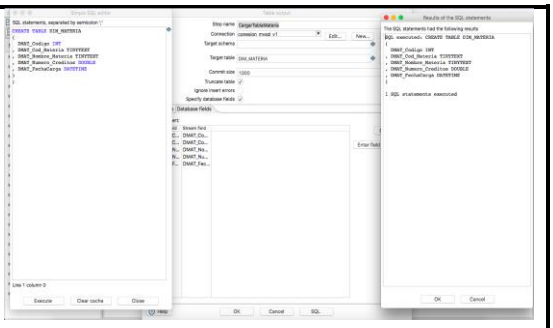
#	DFACT_Codigo	DFACT_Factores	DFACT_FechaCarga
1	1	Sin factores	30/04/2018
2	2	Actividades Deportes	30/04/2018
3	3	Actividades Idiomas	30/04/2018
4	4	Actividades Musica	30/04/2018
5	5	Actividades Teatro	30/04/2018
6	6	Adaptacion Escolar	30/04/2018
7	7	Administracion del tiempo	30/04/2018
8	8	Contenidos del curso	30/04/2018
9	9	Discapacidad	30/04/2018
10	10	Entendimiento de contenidos	30/04/2018
11	11	Falta de interes en la materia	30/04/2018

Transformación y carga

Figura 11. Proceso ETL dimensión factores

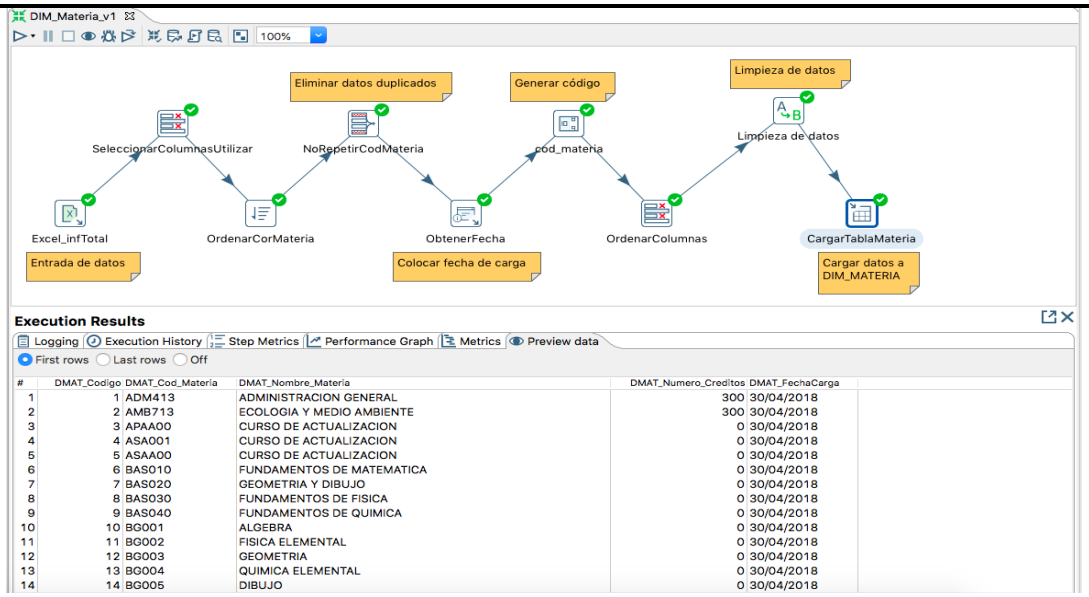


Fuente origen



Fuente destino

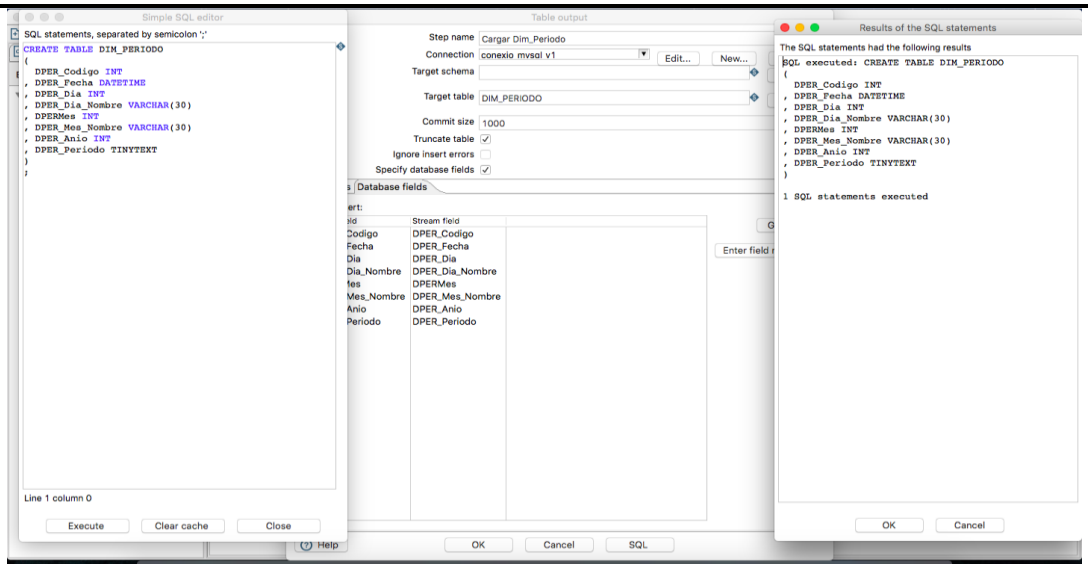
Continúa



#	DMAT_Codigo	DMAT_Cod_Materia	DMAT_Nombre_Materia	DMAT_Numero_Creditos	DMAT_FechaCarga
1	1	ADM413	ADMINISTRACION GENERAL	300	30/04/2018
2	2	AMB713	ECOLOGIA Y MEDIO AMBIENTE	300	30/04/2018
3	3	APAA00	CURSO DE ACTUALIZACION	0	30/04/2018
4	4	ASA001	CURSO DE ACTUALIZACION	0	30/04/2018
5	5	ASAA00	CURSO DE ACTUALIZACION	0	30/04/2018
6	6	BAS010	FUNDAMENTOS DE MATEMATICA	0	30/04/2018
7	7	BAS020	GEOMETRIA Y DIBUJO	0	30/04/2018
8	8	BAS030	FUNDAMENTOS DE FISICA	0	30/04/2018
9	9	BAS040	FUNDAMENTOS DE QUIMICA	0	30/04/2018
10	10	BG001	ALGEBRA	0	30/04/2018
11	11	BG002	FISICA ELEMENTAL	0	30/04/2018
12	12	BG003	GEOMETRIA	0	30/04/2018
13	13	BG004	QUIMICA ELEMENTAL	0	30/04/2018
14	14	BG005	DIBUJO	0	30/04/2018

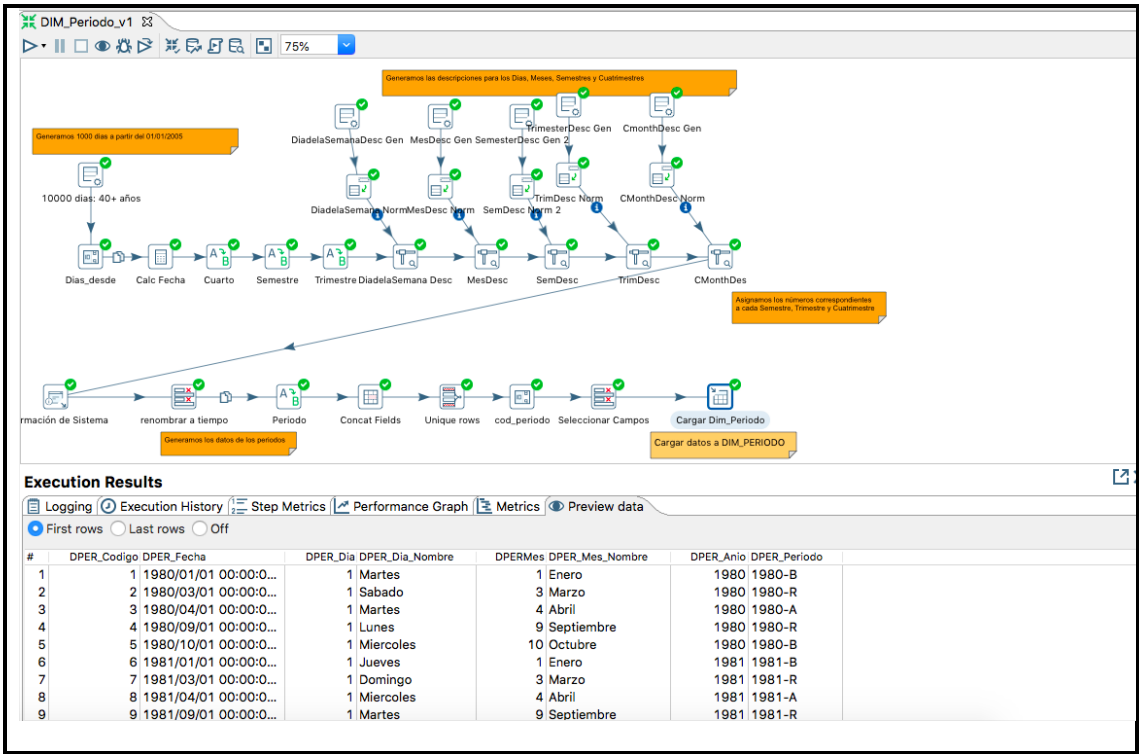
Transformación y carga

Figura 12. Proceso ETL dimensión materia



Fuente destino

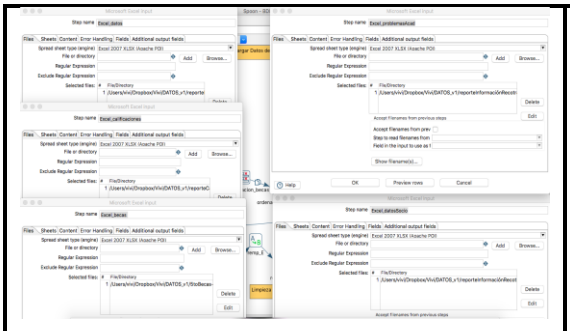
Continúa



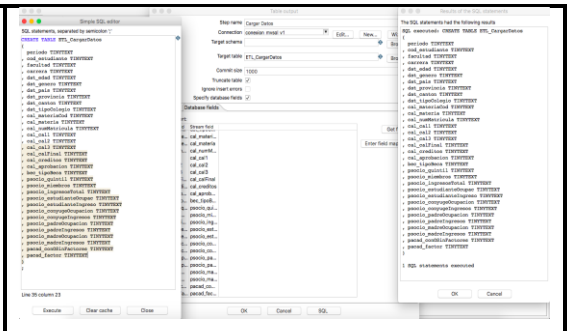
Transformación y carga

Figura 13. Proceso ETL dimensión periodo

Para el ETL de la tabla de hecho, se lo realizó en dos procesos, uno en el cual primero se realizó la carga de todos los datos a una tabla de la nueva base de datos con la respectiva transformación denominado ETL_CargarDatos y limpieza de los mismos y otro mediante el cual se relacionaba con las demás dimensiones, utilizando dichos datos cargados.

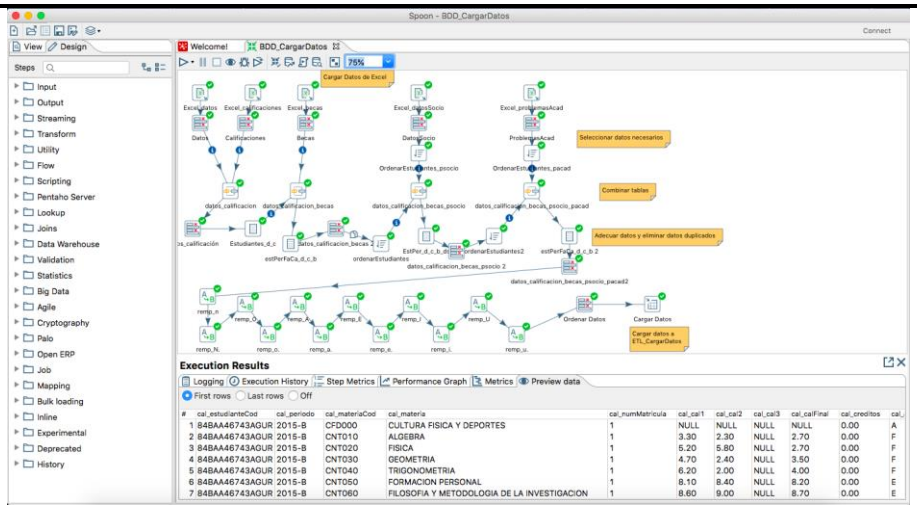


Fuente origen



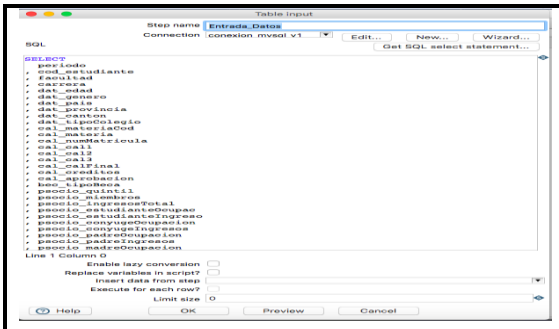
Fuente destino

Continúa

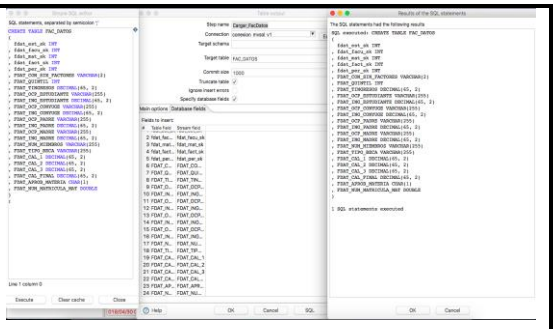


Transformación y carga

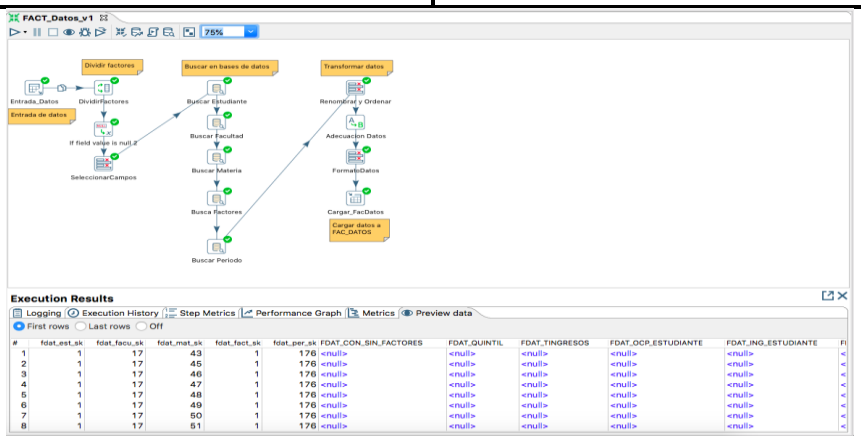
Figura 14. Proceso ETL cargar datos



Fuente origen



Fuente destino



Transformación y carga

Figura 15. Proceso ETL tabla de hechos

4.1.4. Base de datos

Se procedió a crear una base de datos MySQL denominada Tesis_BDD_v1 utilizando la herramienta XAMP que permitió inicializar el servidor gestor de base de datos MySQL y el servidor web Apache, para le respectiva configuración de la base de datos en una página web local.

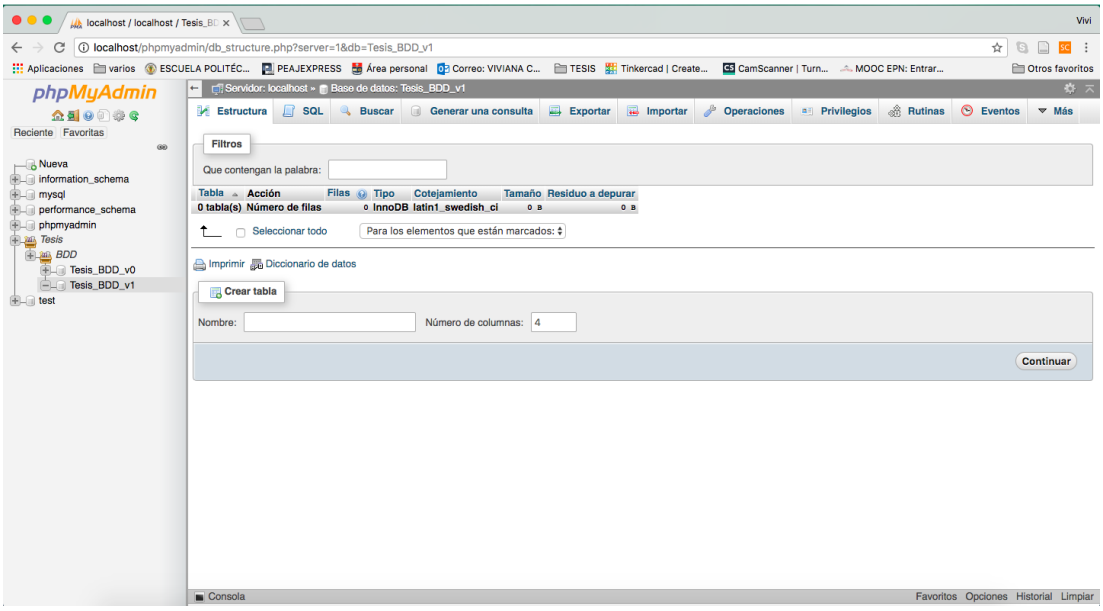


Figura 16. Creación de la base de datos en MySQL

Desde cada uno de los procesos ETL de las dimensiones se cargó los datos a las respectivas tablas de la base de datos en MySQL, mediante la conexión creada en la herramienta *Data Integration* de *Pentaho*, como lo muestra la figura.

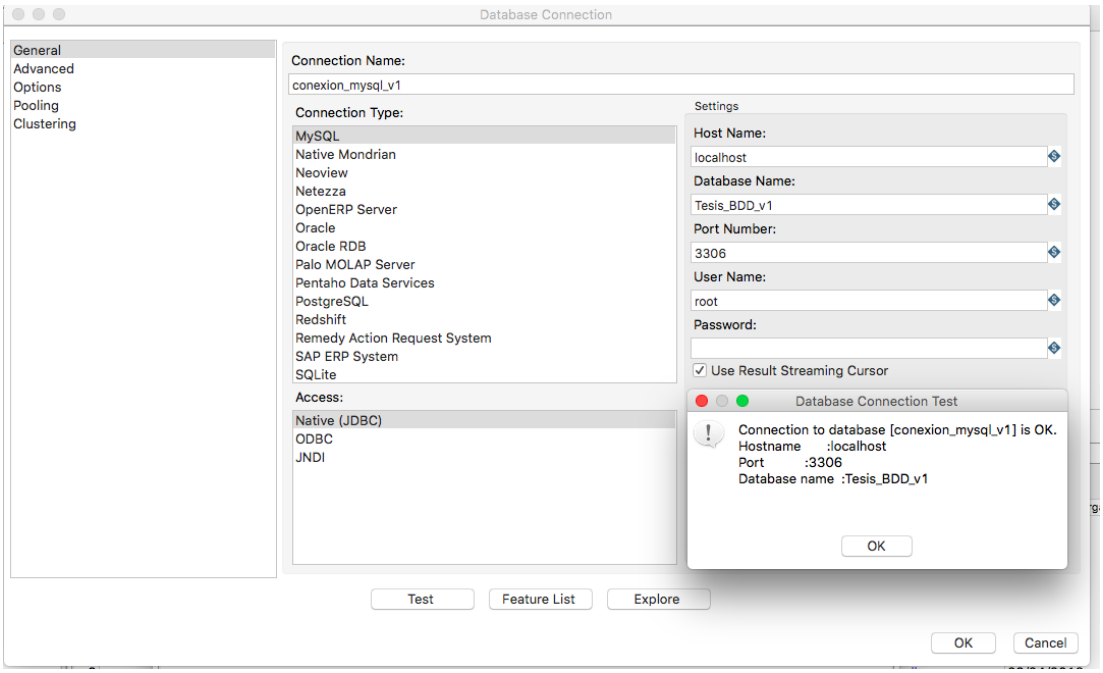
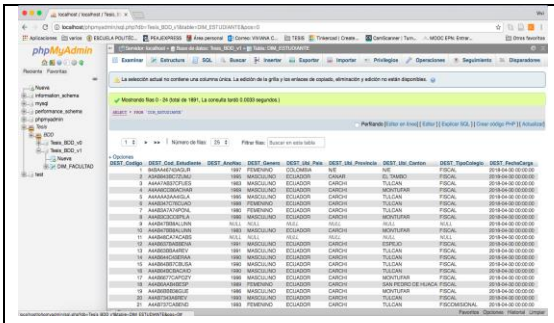
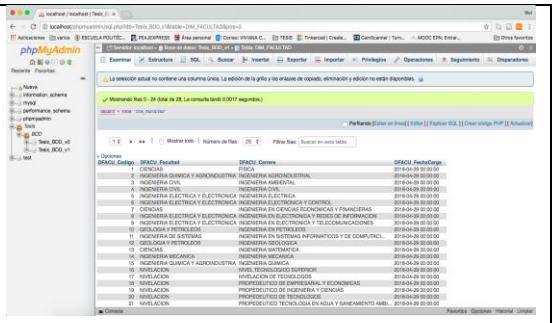


Figura 17. Conexión con la base de datos MySQL mediante *Data Integration*

De esta manera la base de datos se fue cargando, mientras se ejecutaban cada uno de los procesos ETL.

	
DIM_ESTUDIANTE	DIM_FACULTAD

Continúa

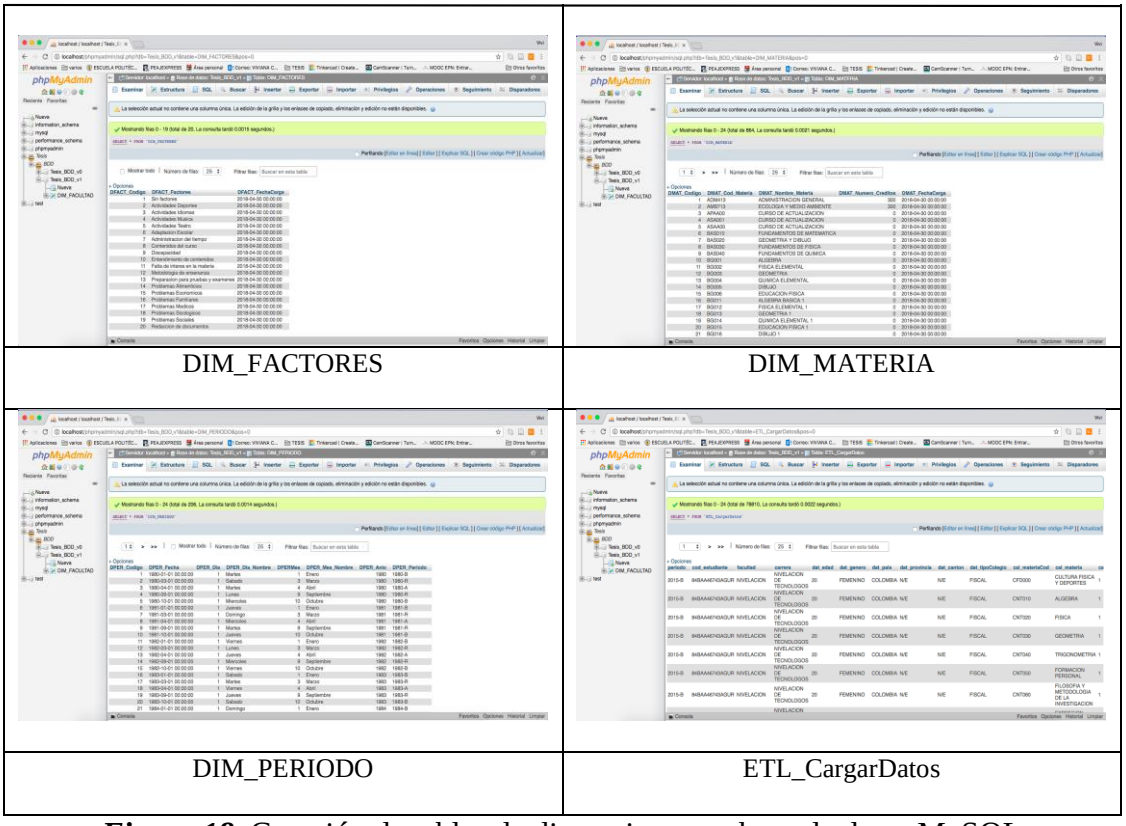


Figura 18. Creación de tablas de dimensiones en base de datos MySQL

Por último, se cargó la tabla de hechos FAC_DATOS, que completo el proceso de cargar los datos a la base de datos escogida, para utilizarlos en la creación del modelo de minería de datos.

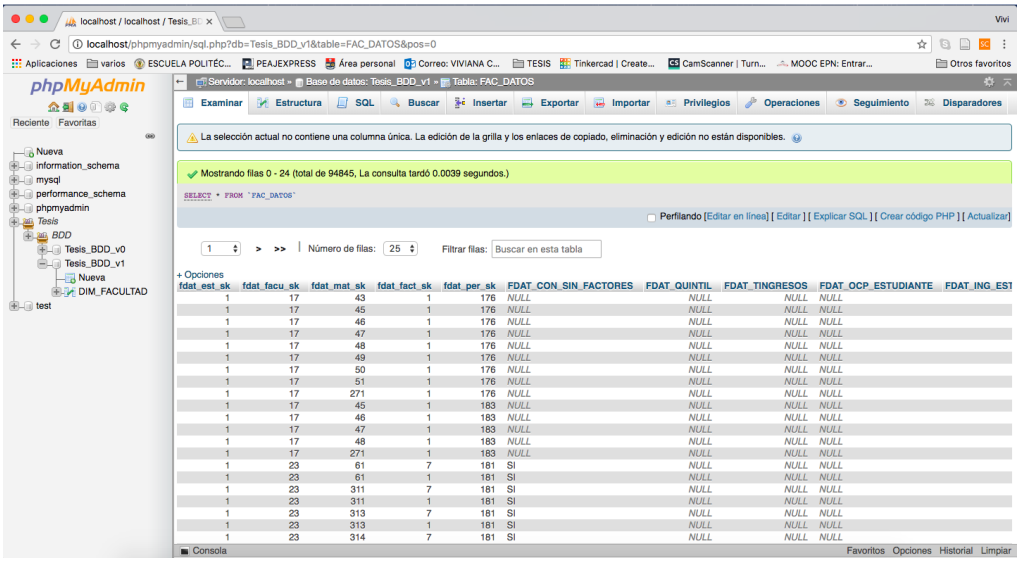


Figura 19. Datos cargados a FAC_DATOS de la base de datos Tesis_BDD_V1

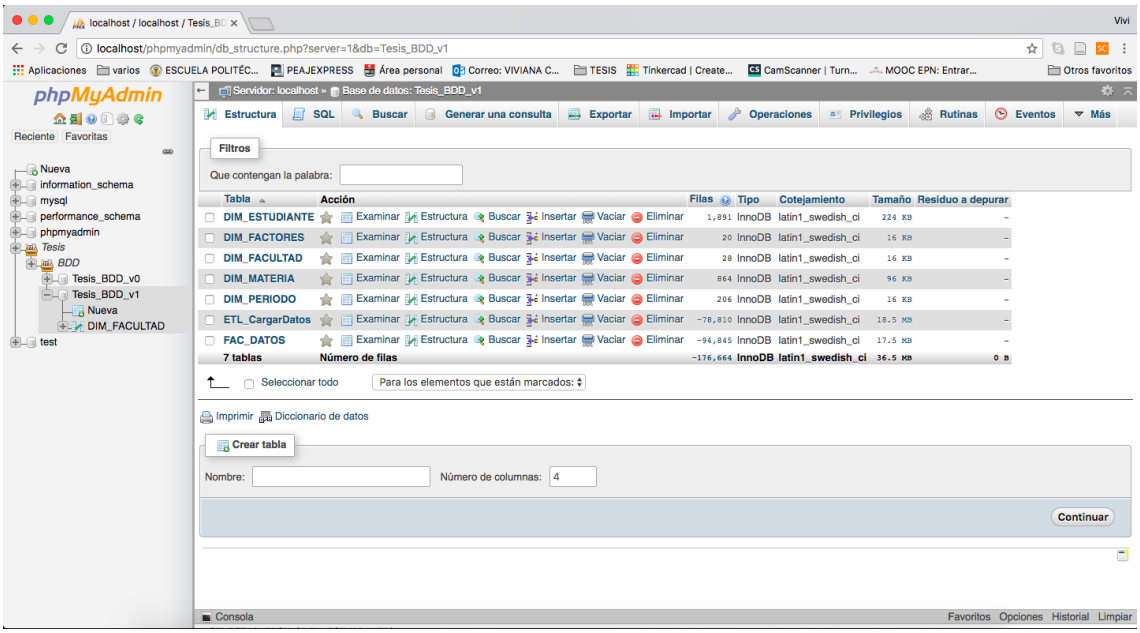


Figura 20. Base de datos Tesis_BDD_v1 con todas las tablas creadas

4.1.5. Auto Model RapidMiner

RapidMiner Auto Model, una nueva adición a RapidMiner Studio desde la versión 8.1 que acelera el proceso de construcción de modelos de aprendizaje automático. Auto Model genera un proceso de RapidMiner Studio detrás de escena, por lo que el análisis de datos puede afinar y probar modelos antes de ponerlos en producción.

Auto Model es una opción de RapidMiner que ayuda a construir modelos perfectos para los datos, incluye la preparación de datos y la optimización del modelo. El resultado no es simplemente una caja negra, sino un proceso RapidMiner que se puede examinar y modificar, optimizándolo aún más.

Es así, como el proceso de minería comenzó tomando los datos de la bodega de datos, que previamente fueron sometidas al proceso ETL para conseguir información confiable y organizada, para analizarlos mediante esta herramienta.

Utilizando la herramienta RapidMiner se procedió a agrupar las dimensiones en una sola tabla para analizarla en la opción Auto Model y seleccionar los atributos de análisis.

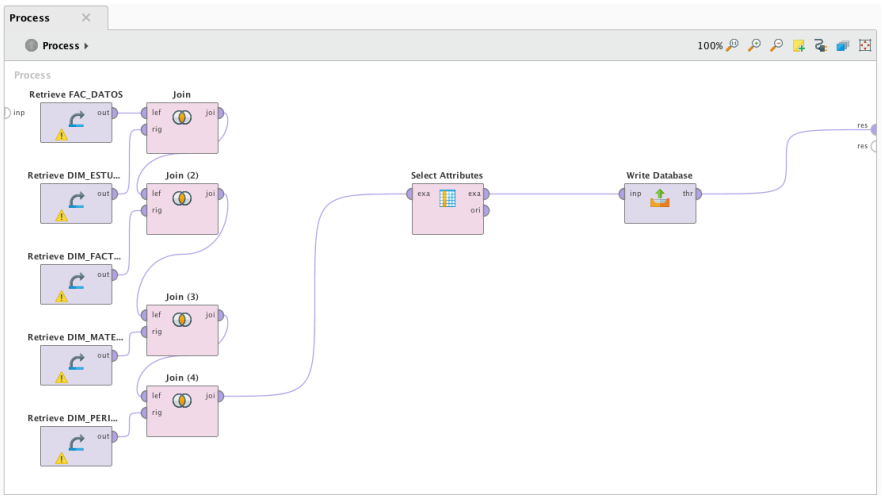


Figura 21. Unión de tablas en RapidMiner

Creada la tabla de todos los datos que abarcan el análisis, se realizó el proceso de Auto Model para determinar las mejores técnicas de minería de datos, de acuerdo a la información proporcionada y la relación entre los atributos que se tienen.

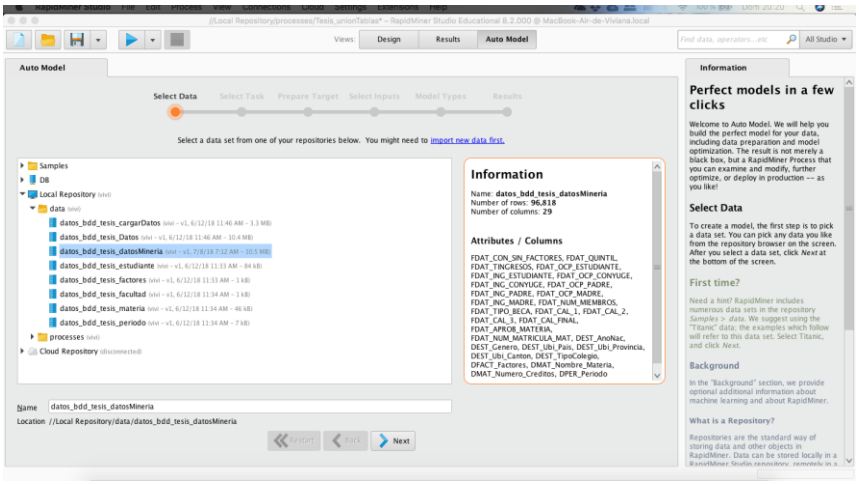


Figura 22. Paso 1 Auto Model RapidMiner

El objetivo principal del proyecto es determinar los factores que influyen en el aprovechamiento académico de los estudiantes por cuanto se determinó que el atributo a predecir es el rendimiento escolar, definido por la calificación final obtenida en las materias que ha tomado.

Es así como en el segundo paso del proceso de Auto Model se indicó la tarea a realizar en el proceso de minería de datos, en este caso se busca predecir los valores de la columna FDAT_CAL_FINAL que están asociados al rendimiento académico de los estudiantes.

Mediante esta selección RapidMiner desarrolla un modelo de aprendizaje automático que predice los valores de la columna seleccionada en función de los valores de las otras columnas.

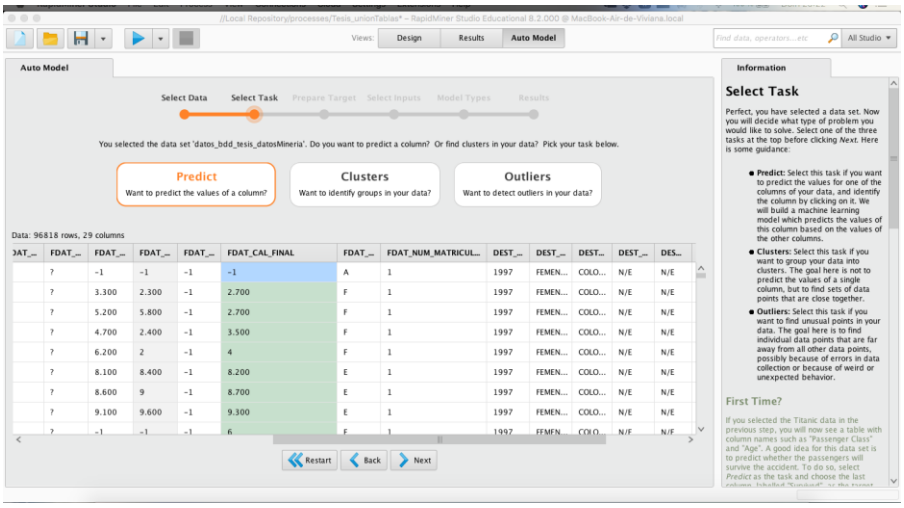


Figura 23. Paso 2 Auto Model RapidMiner

Como tercer paso, RapidMiner permite preparar el objetivo para obtener mejores resultados, en este caso al ser un atributo numérico el seleccionado se tendría como respuesta una regresión, con un valor exacto calculado. Mediante una regresión no se observaría el comportamiento de los diferentes rendimientos académicos que podría tener el estudiante de acuerdo a los factores que tenga. Es así, que en este paso se transforma el atributo a predecir, en uno categórico, divido en 3 rangos, discretizados por frecuencia, para determinar si el estudiante está en un rango alto, bajo o medio de acuerdo a los factores asociados.

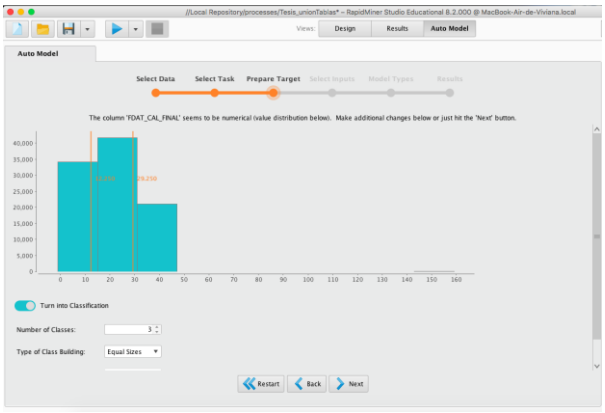


Figura 24. Paso 3 Auto Model RapidMiner

A continuación, el proceso de Auto Model permite seleccionar cuales son los datos de entrada para el modelo, centrándose en la calidad de los mismos de acuerdo al atributo a predecir. Es por esto, que se presenta una tabla con los atributos cargados de la base de datos con sus respectivos porcentajes de: correlación con el atributo destino, valores diferentes, valores idénticos y valores perdidos. Además, la herramienta recomienda que atributos son los que permitirían obtener el mejor modelo, de acuerdo a los datos observados se procede a seleccionar cuales atributos se tomarán en cuenta en el modelo a crear.

Figura 25. Paso 4 Auto Model RapidMiner

En función de los datos y las configuraciones realizadas anteriormente, Auto Model muestra los modelos pertinentes de aprendizaje automático. Además, si se selecciona más de un modelo, los resultados incluyen una comparación de rendimientos. Dependiendo del conjunto de datos, algunos modelos pueden ser deseleccionados, para evitar tiempos de ejecución largos.

Los modelos analizados fueron:

- *Naive Bayes*: un clasificador probabilístico simple y rápido basado en el teorema de Bayes.
- Árbol de decisiones: encuentra modelos arbóreos simples que son fáciles de entender.
- Modelo lineal generalizado (GLM): generalización de modelos de regresión lineal.

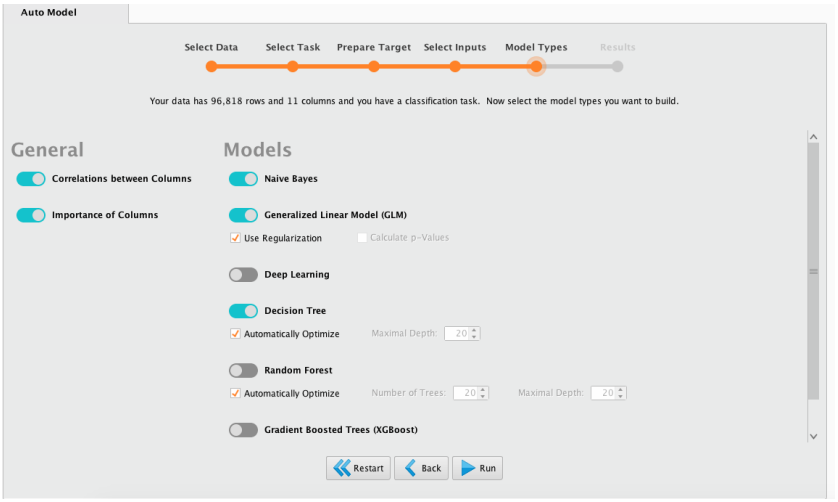


Figura 26. Paso 5 Auto Model RapidMiner

Por último, luego seleccionar la opción ejecutar, la herramienta presentó los resultados obtenidos de manera general de todos los modelos seleccionados y de manera independiente.

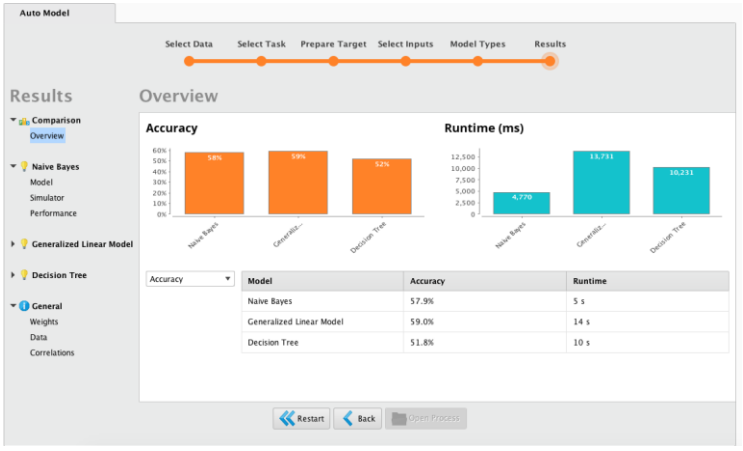


Figura 27. Resultado Auto Model

Además, al momento de seleccionar un modelo determinado, se permitió observar el proceso creado para crear dicho modelo.

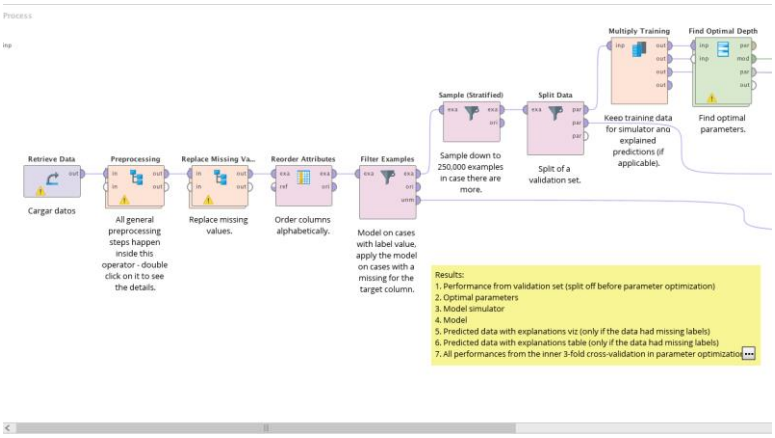


Figura 28. Proceso creado por Auto Model

4.1.6. Identificación de algoritmo de minería de datos

Para determinar los atributos con los cuales se realizó el modelo, se procedió a analizar la tabla que Auto Model presentó en su cuarto paso, que se visualiza a continuación:

Tabla 7

Datos de los atributos determinado por Auto Model

	Atributo	Correlación	Estabilidad	Datos Faltantes	Recomendado Auto Model
1	FDT_QUINTIL	0,17 %	29.07%	95.81%	NO
2	FDAT_TINGRESOS	1.11%	6.50%	95.81%	NO
3	FDAT_OCP_ESTUDIANTE	0,02%	90.91%	95.81%	NO
4	FDAT_ING_ESTUDIANTE	2.13%	91.16%	95.81%	NO
5	FDAT_OCP_CONYUGE	0.01%	96.23%	95.81%	NO
6	FDAT_ING_CONYUGE	0.19%	96.50%	95.81%	NO
7	FDAT_OCP_PADRE	0.01%	42.52%	95.81%	NO
8	FDAT_ING_PADRE	0.11%	42,23%	95.81%	NO
9	FDAT_OCP_MADRE	0.02%	36.44%	95.81%	NO
10	FDAT_ING_MADRE	0.01%	60.68%	95.81%	NO
11	FDAT_NUM_MIEMBROS	0.02%	30.99%	95.81%	NO
12	FDAT_TIPO_BECA	0.37%	95.53%	95.89%	NO
13	DEST_Ubi_Pais	0.05%	99.64%	3.05%	NO
14	FDAT_CON_SIN_FACTORES	3.73%	82,22%	65.91%	SI
15	FDAT_CAL_1	12.38%	21.41%	0%	SI
16	FDAT_CAL_2	15.70%	24.03%	0%	SI
17	FDAT_CAL_3	16.94%	72%	0%	SI
18	FDAT_APROB_MATERIA	20.94%	46.86%	0%	SI

Continúa

19	FDAT_NUM_MATRICULA_MAT	0.92%	81.48%	0%	SI
20	DEST_AnoNac	0.44%	11.82%	3.05%	SI
21	DEST_Genero	0.06%	74.06%	3.05%	SI
22	DEST_Ubi_Provincia	0.04%	79.92%	3.05%	SI
23	DEST_Ubi_Canton	0.04%	73.47%	3.05%	SI
24	DEST_TipoColegio	0.03%	74.33%	3.05%	SI
25	DFACT_Factores	0.78%	82.43	0%	SI
26	DMAT_Nombre_Materia	2.06%	4.37%	0%	SI
27	DMAT_Numero_Creditos	1.13%	25.03%	0%	SI
28	DPER_Periodo	0.14%	13.53%	0%	SI

Mediante esta tabla se determinó seis casos de estudio:

1. Recomendados por RapidMiner
2. En base a la experiencia docente
3. Atributos con mejor estabilidad de datos
4. Atributos con mejor correlación con el atributo de salida
5. Atributos con menor porcentaje de datos faltantes
6. Considerando como atributo a predecir FDAT_APROB_MATERIA que está directamente relacionado con el atributo FDAT_CAL_FINAL

Considerando cada uno de estos casos se seleccionó los atributos para construir los modelos en Auto Model, obteniendo un grado de precisión por cada modelo seleccionado y estimando un promedio de estos.

Tabla 8

Casos de estudio

CASOS	CASO 1	CASO 2	CASO 3	CASO 4	CASO 5	CASO 6
ATRIBUTOS SELECCIONADOS	FDAT_CON_SIN_FACTORES	FDT_QUINTIL	DFACT_Factores	FDAT_APROB_MATERIA	FDAT_APROB_MATERIA	FDAT_CON_SIN_FACTORES
	FDAT_CAL_1	FDAT_TIPO_BECA	DEST_Ubi_Pais	FDAT_CAL_3	FDAT_CAL_3	FDAT_CAL_1
	FDAT_CAL_2	FDAT_NUM_MATRICULA_MAT	FDAT_ING_CONYUGE	FDAT_CAL_2	FDAT_CAL_2	FDAT_CAL_2
	FDAT_CAL_3	DEST_AnoNac	FDAT_OCP_CONYUGE	FDAT_CAL_1	FDAT_CAL_1	FDAT_CAL_3
	FDAT_APROB_MATERIA	DEST_Genero	FDAT_TIPO_BECA	FDAT_CON_SIN_FACTORES	DMAT_Nombre_Materia	FDAT_APROB_MATERIA
	FDAT_NUM_MATRICULA_MAT	DEST_Ubi_Provincia	FDAT_ING_ESTUDIANTE	FDAT_ING_ESTUDIANTE	DMAT_Numero_Creditos	FDAT_NUM_MATRICULA_MAT
	DEST_AnoNac	DEST_TipoColegio	FDAT_OCP_ESTUDIANTE	DMAT_Nombre_Materia	FDAT_NUM_MATRICULA_MAT	DEST_AnoNac
	DEST_Genero	DFACT_Factores	FDAT_CON_SIN_FACTORES	DMAT_Numero_Creditos	DFACT_Factores	DEST_Genero
	DEST_Ubi_Provincia	DMAT_Nombre_Materia	FDAT_NUM_MATRICULA_MA	FDAT_TINGRESOS	DPER_Periodo	DEST_Ubi_Provincia
	DEST_Ubi_Canton	DMAT_Numero_Creditos	DEST_Ubi_Provincia		DEST_Ubi_Pais	DEST_Ubi_Canton
	DEST_TipoColegio	DPER_Periodo	DEST_TipoColegio		DEST_AnoNac	DEST_TipoColegio
	DFACT_Factores		DEST_Genero		DEST_Genero	DFACT_Factores
	DMAT_Nombre_Materia		DEST_Ubi_Canton		DEST_Ubi_Provincia	DMAT_Nombre_Materia
	DMAT_Numero_Creditos		FDAT_CAL_3		DEST_Ubi_Canton	DMAT_Numero_Creditos
	DPER_Periodo		FDAT_ING_MADRE		DEST_TipoColegio	DPER_Periodo
PROMEDIO PRECISIÓN MODELOS	80,32%	59%	53%	78,57%	80,60%	92,77%

A partir, de los resultados obtenidos se determinó que:

- La estabilidad de los datos no aporta a la creación de un modelo adecuado.
- Es preferible remplazar los datos faltantes con datos promedio para que no afecte el modelo.
- Tomar como variable a predecir el atributo FDAT_APROB_MATERIA genera una mejor precisión al trabajar con datos nominales.

Como resultado de este análisis se tomó como atributos de entrada y salida para la creación de un modelo óptimo los siguientes:

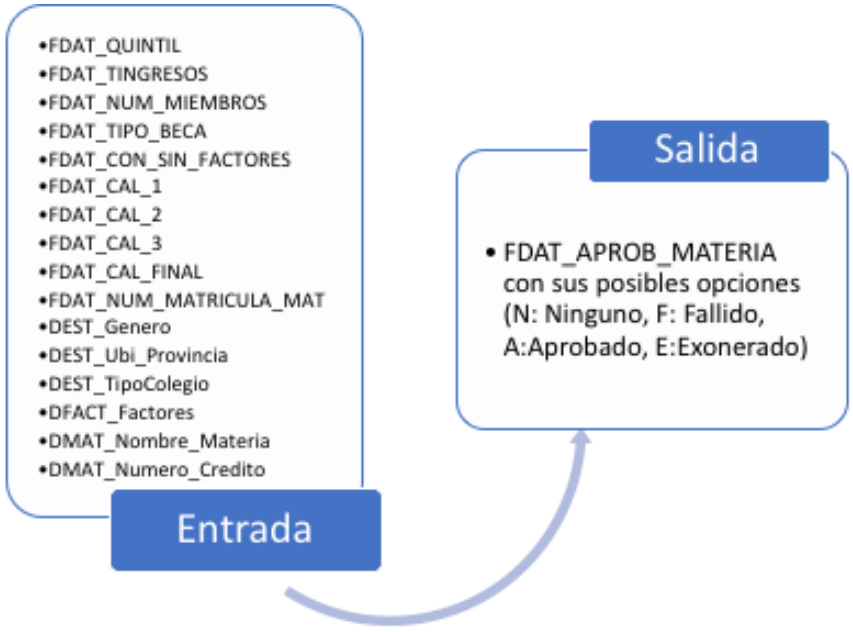


Figura 29. Atributos de entrada y salida del modelo

A continuación, se comparó los resultados de precisión obtenidos en cada uno de los casos de estudio, luego de analizarlos en RapidMiner Auto Model determinando el mejor algoritmo para modelo a crear.

Tabla 9

Comparación de modelos por precisión

PRECISIÓN	NAIVE BAYES	MODELO LINEAL	DECISION TREE
CASO 0	71,06	80,70	89,20
CASO 1	57,90	59,00	51,80
CASO 2	47,80	53,30	57,90
CASO 3	68,80	79,50	87,40
CASO 4	72,00	80,60	89,20
CASO 5	88,00	93,20	97,10
CASO 6	60,10	60,20	50,90
CASO APROBADO	87,40	92,60	96,90
	69,13	74,89	77,55

De acuerdo, a los resultados obtenidos en la tabla anterior se puede observar que el algoritmo que mejores resultados presenta es DECISION TREE.

Además, se realizó el estudio de la opción del atributo FDAT_APROB_MATERIA que cada uno de los algoritmos estima es el más probable.

Tabla 10

Comparación de modelos por atributo probable

RANGO PROBABLE	NAIVE BAYES	MODELO LINEAL	DECISION TREE
CASO 0	E	E	A
CASO 1	F	E	E
CASO 2	A	F	F
CASO 3	E	E	A
CASO 4	E	E	A
CASO 5	A	E	F
CASO 6	E	E	E
CASO APROBADO	A	E	F
F	1 vez	1 vez	3 veces
A	3 veces	0 veces	3 veces
E	4 veces	7 veces	2 veces

Como se visualiza en la tabla anterior el algoritmo DECISION TREE brinda resultados aleatorios entre las diferentes opciones del atributo a predecir, permitiendo ampliar las posibilidades de respuesta del modelo en los diferentes rendimientos académicos de los estudiantes.

Es así, como se identifica el algoritmo de análisis a DECISION TREE para el modelo a crear, mediante la herramienta de minería de datos RapidMiner Auto Model.

4.1.7. Modelo de minería de datos

Mediante el proceso de minería de datos se consiguió seleccionar los comportamientos de los alumnos de acuerdo a los factores dados, para detectar posibles tendencias o patrones de comportamiento relacionadas con el desempeño académico.

Utilizando métodos matemáticos de análisis mediante diferentes algoritmos y técnicas, tales como regresión, algoritmos genéticos, clustering, redes neuronales, inteligencia artificial, árboles de decisión, reglas de asociación, entre otras, la Minería de Datos es de gran ayuda para realizar el análisis inteligente de grandes volúmenes de información. La minería de datos relacionada con la educación se denomina “Minería de datos educativa” (CEDANO & CASTRO, 2015).

La técnica de clasificación implementada mediante árboles de decisión o redes neuronales es una de las más utilizada en minería de datos. La clasificación se realiza mediante in aprendizaje propio, donde se entrena los satos utilizando diferentes algoritmos, para sus posteriores pruebas y comprobación de resultados (Blog, 2016).

En esta etapa del proyecto se aplicó el algoritmo seleccionado en el proceso anterior determinar los patrones de comportamiento que afectan en el aprovechamiento académico de los estudiantes.

Antes de modelar con los atributos seleccionados se modificó los datos, remplazando los valores faltantes con un promedio de los existentes para obtener mejores resultados creando una nueva tabla en la base de datos denominada Tesis_DatosMineria1.

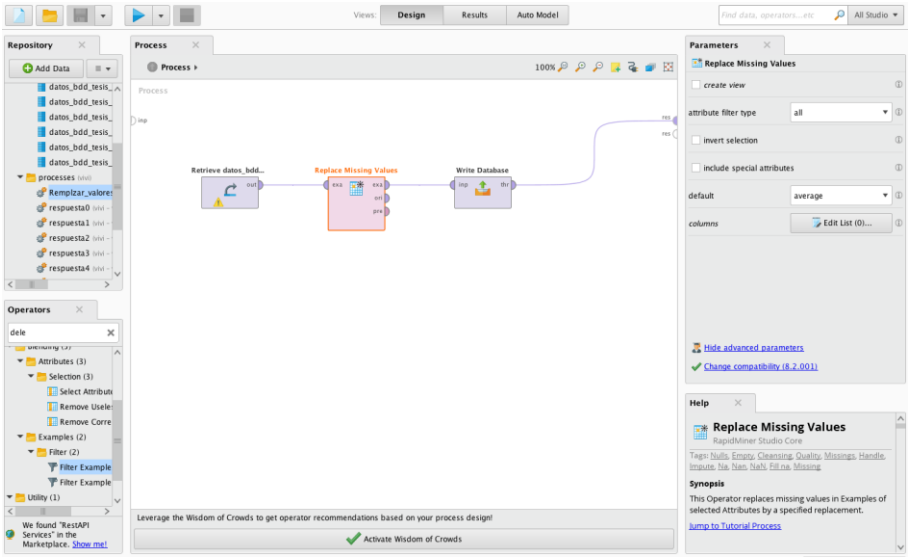


Figura 30. Proceso para remplazar valores nulos

A continuación, se seleccionó dicha tabla en la opción Auto Model para comenzar con la creación del modelo.

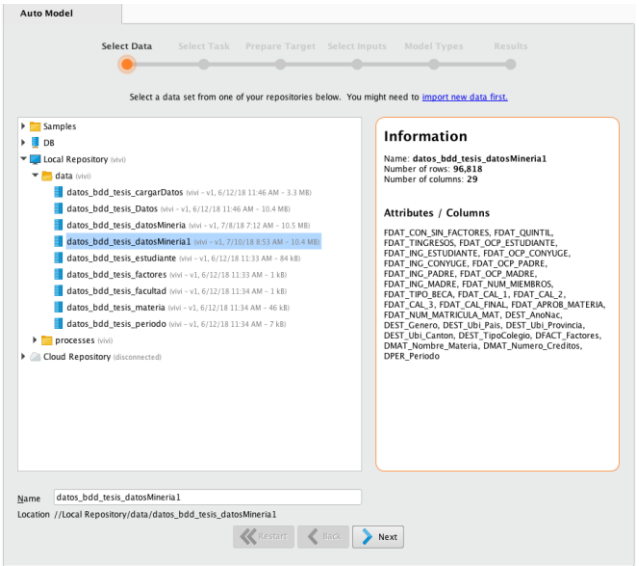


Figura 31. Selección de datos para el modelo

Luego se definió como atributo a predecir a FDAT_APROB_MATERIA:

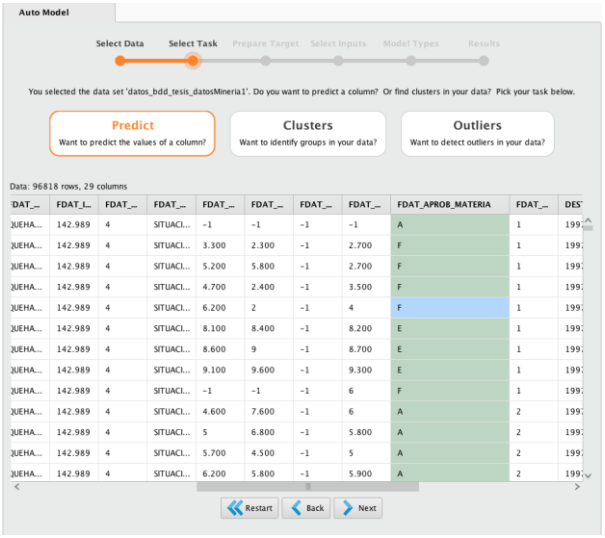


Figura 32. Selección de atributo a predecir

Como siguiente paso se comprobó las opciones que se tendría en la variable a predecir, y se colocó nombres más descriptivos para identificarlos.

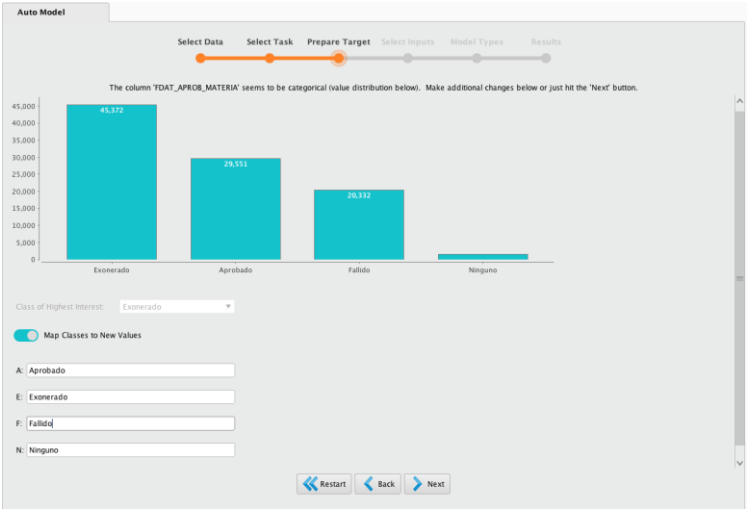


Figura 33. Opciones de la variable a predecir

Seguidamente, se estableció los atributos de entrada, que anteriormente se habían seleccionado.

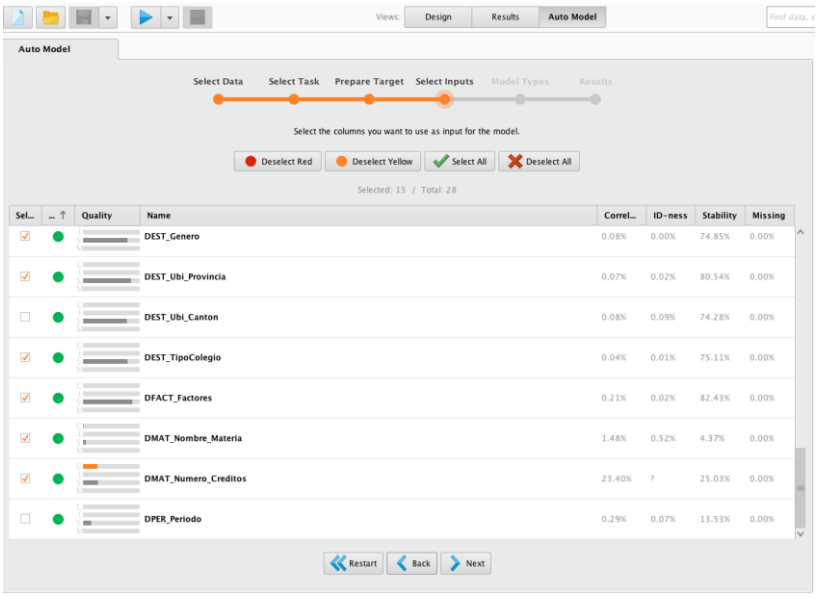


Figura 34. Selección de atributos de entrada

Por último, se seleccionó el algoritmo de modelamiento, en este caso se escogió los tres algoritmos, únicamente para ratificar la mejor precisión alcanzada con el algoritmo DECISION TREE.

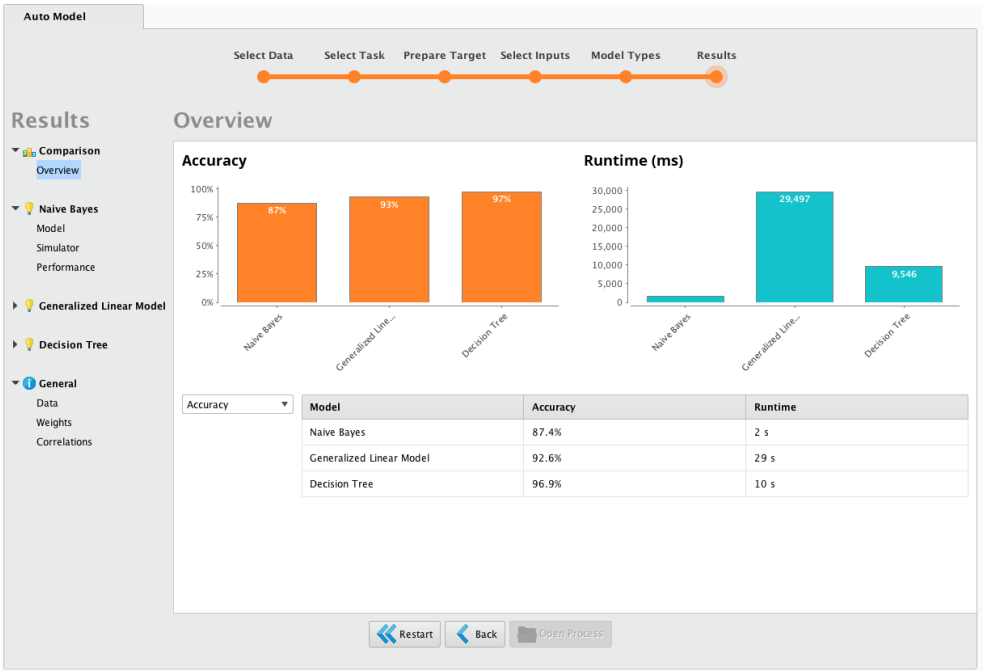


Figura 35. Comparación de algoritmos

Como ejemplo, seleccionando el algoritmo Naive Bayes se pudieron observar resultados como:

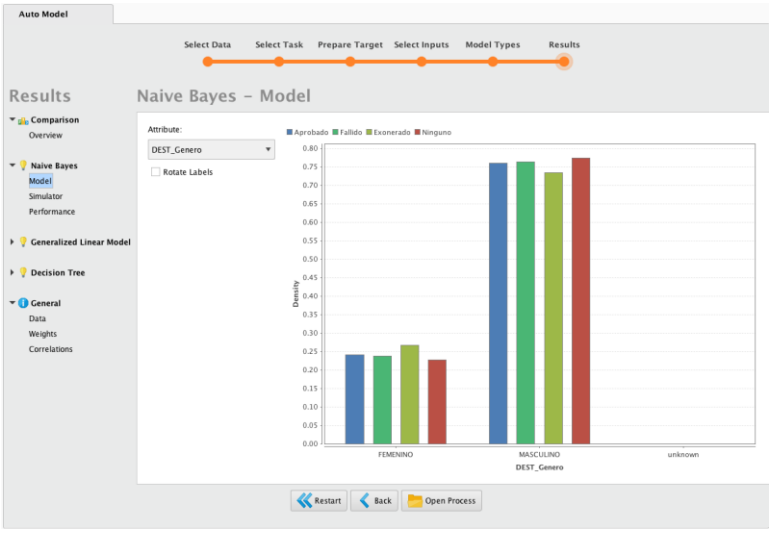


Figura 36. Resultado con algoritmo Naive Bayes

Donde se puede observar que un patrón de comportamiento es que las estudiantes de género femenino tienden a exonerarse mientras que el género masculino tiende a perder y aprobar la materia.

Mientras que seleccionando el algoritmo Modelo lineal generalizado se obtuvo resultados como:

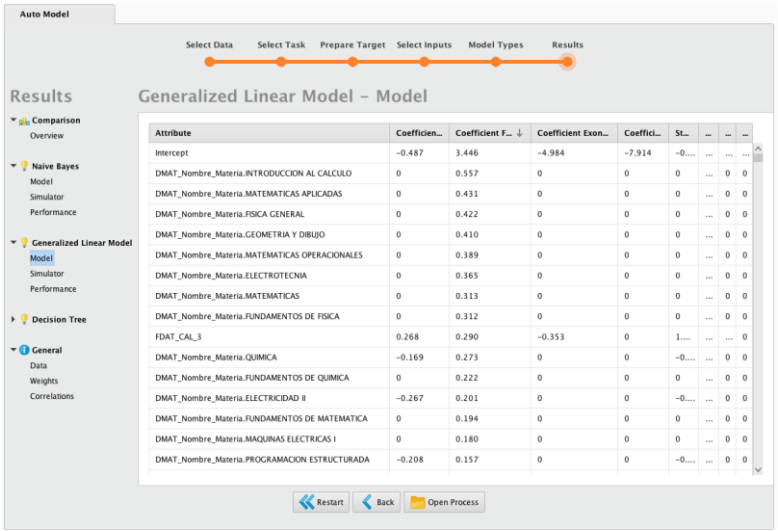


Figura 37. Resultado con algoritmo Lineal Generalizado

Por ejemplo, mediante este modelo, se puede notar que las materias básicas, orientadas a las Matemáticas tienen un coeficiente alto en la opción de fallar la materia.

Por último, al momento de analizar el algoritmo DECISION TREE que fue el seleccionado se obtuvo el siguiente modelo de árbol de decisión.

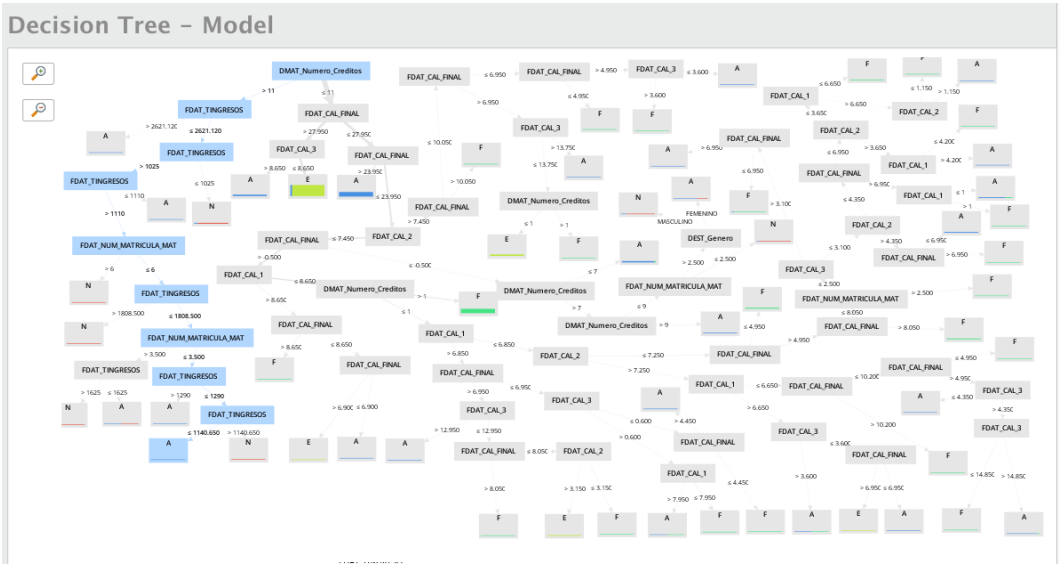


Figura 38. Resultado con algoritmo Decision Tree

De acuerdo al árbol de decisión obtenido, se nota que la opción “NINGUNO” no aporta al modelo creado por cuanto se procedió a abrir el proceso de RapidMiner para adecuar el mismo eliminando esta opción mediante un filtro.

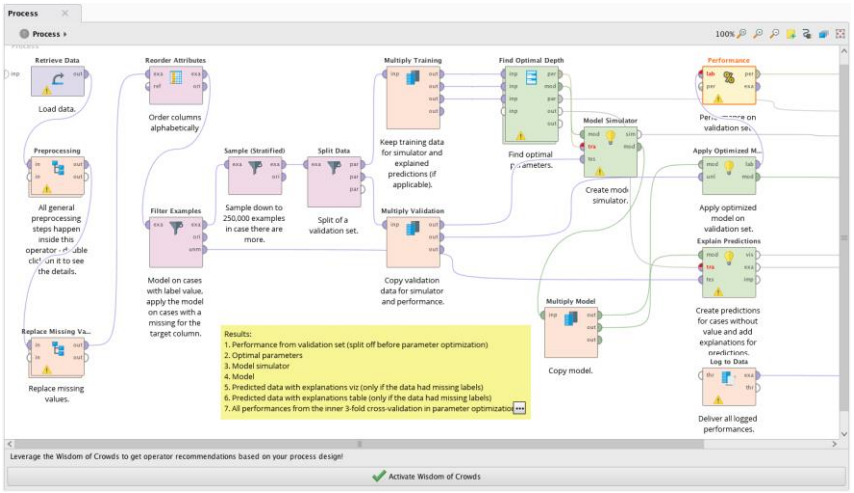


Figura 39. Proceso del algoritmo Decision Tree

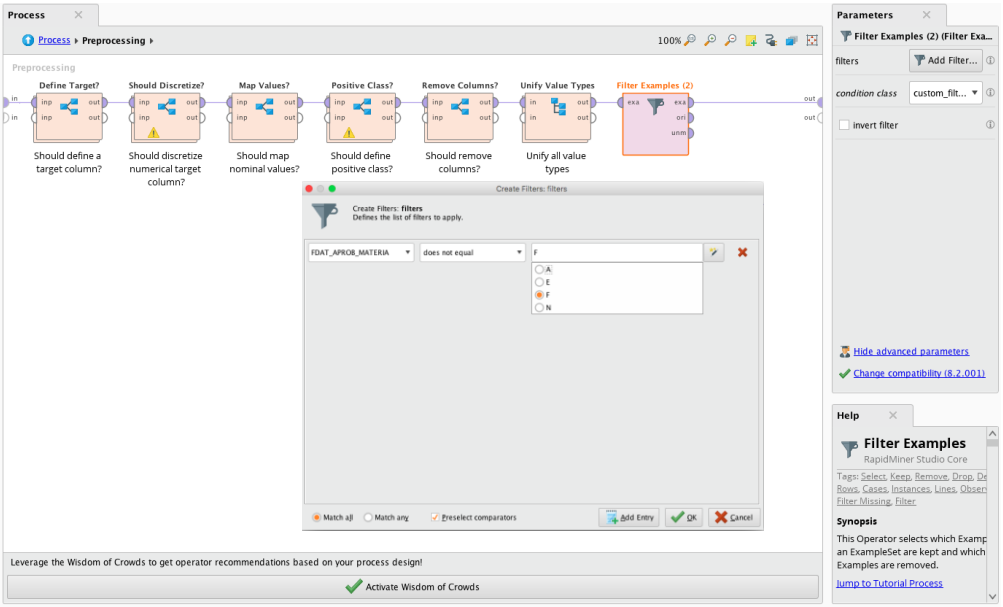


Figura 40. Proceso para eliminar opción “Ninguno”

Con el proceso creado, se logró obtener un árbol de decisión más óptimo, como se visualiza en la siguiente figura.

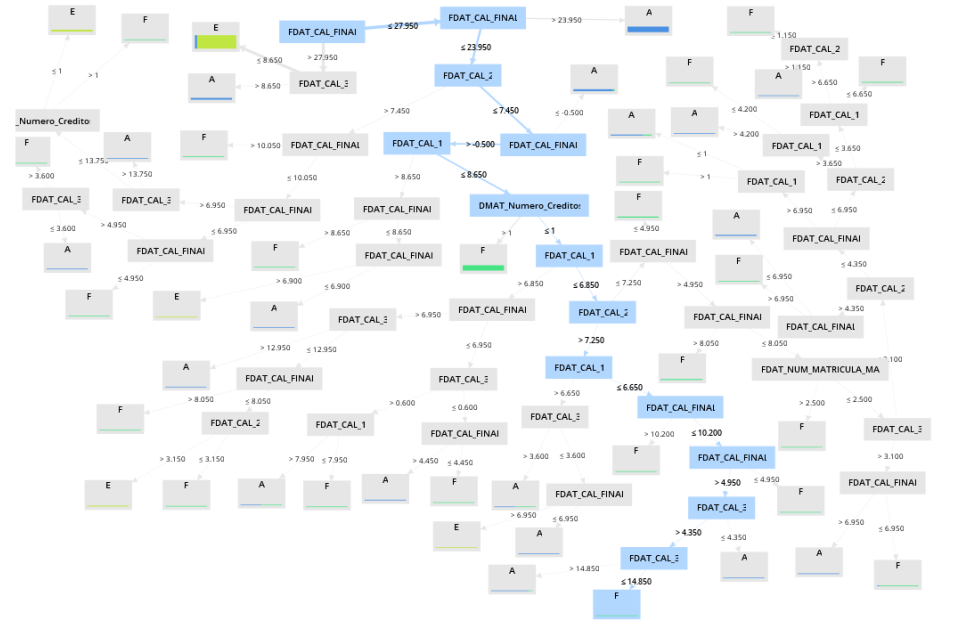


Figura 41. Modelo utilizando árbol de decisión

Mediante este modelo, se pueden estimar un conjunto de reglas de decisión que indican los patrones de comportamiento de los estudiantes, por ejemplo, para que un estudiante falle una materia intervienen factores como:

- Calificación final
- Calificación 1
- Calificación 2
- Calificación 3
- Número de créditos de la materia

También, se obtuvo las reglas de decisión generadas por el árbol de decisión creado.

```
Tree
DMAT_Numero_Creditos > 11
|   FDAT_TINGRESOS > 2621.120: Aprobado {Aprobado=2, Fallido=0, Exonerado=0, Ninguno=0}
|   FDAT_TINGRESOS ≤ 2621.120
|   |   FDAT_TINGRESOS > 1025
|   |   |   FDAT_TINGRESOS > 1110
|   |   |   |   FDAT_NUM_MATRICULA_MAT > 6: Ninguno {Aprobado=0, Fallido=0, Exonerado=0, Ninguno=4}
|   |   |   |   FDAT_NUM_MATRICULA_MAT ≤ 6
|   |   |   |   |   FDAT_TINGRESOS > 1808.500: Ninguno {Aprobado=0, Fallido=0, Exonerado=0,
Ninguno=3}
|   |   |   |   |   FDAT_TINGRESOS ≤ 1808.500
|   |   |   |   |   |   FDAT_NUM_MATRICULA_MAT > 3.500
|   |   |   |   |   |   |   FDAT_TINGRESOS > 1625: Ninguno {Aprobado=0, Fallido=0, Exonerado=0,
Ninguno=2}
|   |   |   |   |   |   |   FDAT_TINGRESOS ≤ 1625: Aprobado {Aprobado=1, Fallido=0, Exonerado=0,
Ninguno=1}
|   |   |   |   |   |   FDAT_NUM_MATRICULA_MAT ≤ 3.500
|   |   |   |   |   |   |   FDAT_TINGRESOS > 1290: Aprobado {Aprobado=5, Fallido=0, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   |   FDAT_TINGRESOS ≤ 1290
```

Continúa 

```
|   |   |   |   |   |   |   FDAT_TINGRESOS > 1140.650: Ninguno {Aprobado=0, Fallido=0,
Exonerado=0, Ninguno=2}
|   |   |   |   |   |   |   FDAT_TINGRESOS ≤ 1140.650: Aprobado {Aprobado=2, Fallido=0,
Exonerado=0, Ninguno=0}
|   |   |   FDAT_TINGRESOS ≤ 1110: Aprobado {Aprobado=5, Fallido=0, Exonerado=0, Ninguno=0}
|   |   FDAT_TINGRESOS ≤ 1025: Ninguno {Aprobado=132, Fallido=0, Exonerado=0, Ninguno=1056}
DMAT_Numero_Creditos ≤ 11
|   FDAT_CAL_FINAL > 27.950
|   |   FDAT_CAL_3 > 8.650: Aprobado {Aprobado=2960, Fallido=0, Exonerado=0, Ninguno=0}
|   |   FDAT_CAL_3 ≤ 8.650: Exonerado {Aprobado=1767, Fallido=0, Exonerado=32214, Ninguno=0}
|   FDAT_CAL_FINAL ≤ 27.950
|   |   FDAT_CAL_FINAL > 23.950: Aprobado {Aprobado=13096, Fallido=0, Exonerado=22, Ninguno=0}
|   |   FDAT_CAL_FINAL ≤ 23.950
|   |   |   FDAT_CAL_2 > 7.450
|   |   |   |   FDAT_CAL_FINAL > 10.050: Fallido {Aprobado=0, Fallido=238, Exonerado=0, Ninguno=0}
|   |   |   |   FDAT_CAL_FINAL ≤ 10.050
|   |   |   |   |   FDAT_CAL_FINAL > 6.950
|   |   |   |   |   |   FDAT_CAL_3 > 13.750: Aprobado {Aprobado=20, Fallido=0, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   FDAT_CAL_3 ≤ 13.750
|   |   |   |   |   |   |   DMAT_Numero_Creditos > 1: Fallido {Aprobado=0, Fallido=3, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   |   DMAT_Numero_Creditos ≤ 1: Exonerado {Aprobado=0, Fallido=0, Exonerado=3207,
Ninguno=0}
|   |   |   |   |   FDAT_CAL_FINAL ≤ 6.950
|   |   |   |   |   |   FDAT_CAL_FINAL > 4.950
|   |   |   |   |   |   |   FDAT_CAL_3 > 3.600: Fallido {Aprobado=0, Fallido=7, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   |   FDAT_CAL_3 ≤ 3.600: Aprobado {Aprobado=143, Fallido=0, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   FDAT_CAL_FINAL ≤ 4.950: Fallido {Aprobado=0, Fallido=16, Exonerado=0,
Ninguno=0}
|   |   |   FDAT_CAL_2 ≤ 7.450
|   |   |   |   FDAT_CAL_FINAL > -0.500
```

Continúa 

```
|   |   |   |   |   FDAT_CAL_1 > 8.650
|   |   |   |   |   |   FDAT_CAL_FINAL > 8.650: Fallido {Aprobado=0, Fallido=43, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   FDAT_CAL_FINAL ≤ 8.650
|   |   |   |   |   |   |   FDAT_CAL_FINAL > 6.900: Exonerado {Aprobado=1, Fallido=0, Exonerado=163,
Ninguno=0}
|   |   |   |   |   |   |   FDAT_CAL_FINAL ≤ 6.900: Aprobado {Aprobado=4, Fallido=0, Exonerado=0,
Ninguno=0}
|   |   |   |   |   FDAT_CAL_1 ≤ 4.650
|   |   |   |   |   |   DMAT_Numero_Creditos > 1: Fallido {Aprobado=0, Fallido=12409, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   DMAT_Numero_Creditos ≤ 1
|   |   |   |   |   |   |   FDAT_CAL_1 > 6.850
|   |   |   |   |   |   |   |   FDAT_CAL_FINAL > 6.950
|   |   |   |   |   |   |   |   |   |   FDAT_CAL_3 > 12.950: Aprobado {Aprobado=35, Fallido=0,
Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   FDAT_CAL_3 ≤ 12.950
|   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL > 8.050: Fallido {Aprobado=0, Fallido=56,
Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL ≤ 8.050
|   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 > 3.150: Exonerado {Aprobado=0, Fallido=0,
Exonerado=660, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 ≤ 3.150: Fallido {Aprobado=0, Fallido=9,
Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   FDAT_CAL_FINAL ≤ 6.950
|   |   |   |   |   |   |   |   |   FDAT_CAL_3 > 0.600
|   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 > 7.950: Aprobado {Aprobado=1, Fallido=1,
Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 ≤ 7.950: Fallido {Aprobado=1, Fallido=67,
Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   FDAT_CAL_3 ≤ 0.600
|   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL > 4.450: Aprobado {Aprobado=236, Fallido=0,
Exonerado=0, Ninguno=0}
```

Continúa 

| | | | | | | | | | FDAT_CAL_FINAL ≤ 4.450: Fallido {Aprobado=0, Fallido=16, Exonerado=0, Ninguno=0}

| | | | | | | FDAT_CAL_1 ≤ 6.850

| | | | | | | | FDAT_CAL_2 > 7.250

| | | | | | | | | FDAT_CAL_1 > 6.650

| | | | | | | | | | FDAT_CAL_3 > 3.600: Aprobado {Aprobado=1, Fallido=1, Exonerado=0, Ninguno=0}

| | | | | | | | | | FDAT_CAL_3 ≤ 3.600

| | | | | | | | | | | FDAT_CAL_FINAL > 6.950: Exonerado {Aprobado=0, Fallido=0, Exonerado=32, Ninguno=0}

| | | | | | | | | | | FDAT_CAL_FINAL ≤ 6.950: Aprobado {Aprobado=2, Fallido=0, Exonerado=0, Ninguno=0}

| | | | | | | | | FDAT_CAL_1 ≤ 6.650

| | | | | | | | | | FDAT_CAL_FINAL > 10.200: Fallido {Aprobado=0, Fallido=5, Exonerado=0, Ninguno=0}

| | | | | | | | | | FDAT_CAL_FINAL ≤ 10.200

| | | | | | | | | | | FDAT_CAL_FINAL > 4.950

| | | | | | | | | | | | FDAT_CAL_3 > 4.350

| | | | | | | | | | | | | FDAT_CAL_3 > 14.850: Aprobado {Aprobado=9, Fallido=1, Exonerado=0, Ninguno=0}

| | | | | | | | | | | | | FDAT_CAL_3 ≤ 14.850: Fallido {Aprobado=0, Fallido=4, Exonerado=0, Ninguno=0}

| | | | | | | | | | | | FDAT_CAL_3 ≤ 4.350: Aprobado {Aprobado=66, Fallido=0, Exonerado=0, Ninguno=0}

| | | | | | | | | | | FDAT_CAL_FINAL ≤ 4.950: Fallido {Aprobado=0, Fallido=3, Exonerado=0, Ninguno=0}

| | | | | | | | FDAT_CAL_2 ≤ 7.250

| | | | | | | | | FDAT_CAL_FINAL > 4.950

| | | | | | | | | | FDAT_CAL_FINAL > 8.050: Fallido {Aprobado=26, Fallido=998, Exonerado=0, Ninguno=0}

| | | | | | | | | FDAT_CAL_FINAL ≤ 8.050

| | | | | | | | | | | FDAT_NUM_MATRICULA_MAT > 2.500: Fallido {Aprobado=0, Fallido=3, Exonerado=0, Ninguno=0}

| | | | | | | | | | FDAT_NUM_MATRICULA_MAT ≤ 2.500

Continúa

```
|   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_3 > 3.100
|   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL > 6.950: Aprobado {Aprobado=50,
Fallido=0, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL ≤ 6.950: Fallido {Aprobado=15,
Fallido=179, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_3 ≤ 3.100
|   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 > 4.350
|   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL > 6.950: Fallido {Aprobado=0,
Fallido=10, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL ≤ 6.950: Aprobado
{Aprobado=1150, Fallido=0, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 ≤ 4.350
|   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL > 6.950
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 > 1: Fallido {Aprobado=0,
Fallido=50, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 ≤ 1: Aprobado {Aprobado=773,
Fallido=207, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL ≤ 6.950
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 > 3.650
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 > 4.200: Aprobado
{Aprobado=35, Fallido=0, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 ≤ 4.200: Fallido
{Aprobado=0, Fallido=4, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 ≤ 3.650
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 > 6.650
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 > 1.150: Aprobado
{Aprobado=3, Fallido=0, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_2 ≤ 1.150: Fallido
{Aprobado=0, Fallido=4, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   FDAT_CAL_1 ≤ 6.650: Fallido
{Aprobado=1, Fallido=444, Exonerado=0, Ninguno=0}
|   |   |   |   |   |   |   |   |   FDAT_CAL_FINAL ≤ 4.950: Fallido {Aprobado=0, Fallido=1343,
Exonerado=0, Ninguno=0}
|   |   |   |   FDAT_CAL_FINAL ≤ -0.500
```

Continúa

```
|   |   |   |   |   DMAT_Numero_Creditos > 7
|   |   |   |   |   |   DMAT_Numero_Creditos > 9: Aprobado {Aprobado=550, Fallido=0, Exonerado=0,
Ninguno=0}
|   |   |   |   |   |   DMAT_Numero_Creditos ≤ 9
|   |   |   |   |   |   |   FDAT_NUM_MATRICULA_MAT > 2.500
|   |   |   |   |   |   |   |   DEST_Genero = FEMENINO: Aprobado {Aprobado=7, Fallido=0, Exonerado=0,
Ninguno=3}
|   |   |   |   |   |   |   |   DEST_Genero = MASCULINO: Ninguno {Aprobado=4, Fallido=0, Exonerado=0,
Ninguno=20}
|   |   |   |   |   |   |   FDAT_NUM_MATRICULA_MAT ≤ 2.500: Ninguno {Aprobado=13, Fallido=0,
Exonerado=0, Ninguno=159}
|   |   |   |   |   DMAT_Numero_Creditos ≤ 7: Aprobado {Aprobado=2525, Fallido=145, Exonerado=0,
Ninguno=0}
```

Estas reglas expresan de forma analítica el flujo que sigue el árbol de decisión. Además, entre corchetes ({}) se indica cuántos y a que grupo pertenecen los registros.

Por ejemplo, se pueden observar subrayadas algunas reglas que sobresalen:

- Si el número de créditos tomado por el estudiante es menor a 11 y la calificación final está entre 23.95 y 27.95 el estudiante APROBARÁ.
- Si el número de créditos tomado por el estudiante esta entre 1 y 11, la calificación final está entre 23.95 y 27.95, la calificación 2 es menor a 7.45 y la calificación 1 es menor a 4.65 el estudiante FALLARÁ.
- Si el número de créditos tomado por el estudiante esta entre 1 y 11, la calificación final está entre 23.95 y 27.95, la calificación 1 es menor a 6,85 y la calificación 2 es menor a 7,25 el estudiante FALLARÁ.
- Si el número de créditos es menor a 7 el estudiante APROBARÁ.

4.1.8. Evaluación del modelo

Por último, se evaluó el modelo analítico-predictivo creado a través del uso de la técnica de validación matriz de confusión, implementada en la herramienta minería de datos RapidMiner mediante el operando validación cruzada, considerando un porcentaje de la data como grupo de entrenamiento y otro como testeo.

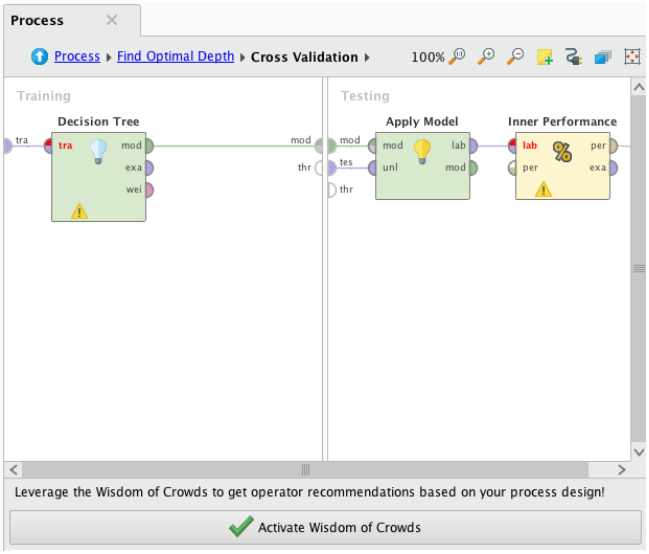


Figura 42. Proceso de validación cruzada

Los porcentajes del grupo de entrenamiento y testeo están determinados por el operando Split Data en 80% de entrenamiento y 20% de testeo lo aconsejado por la herramienta.

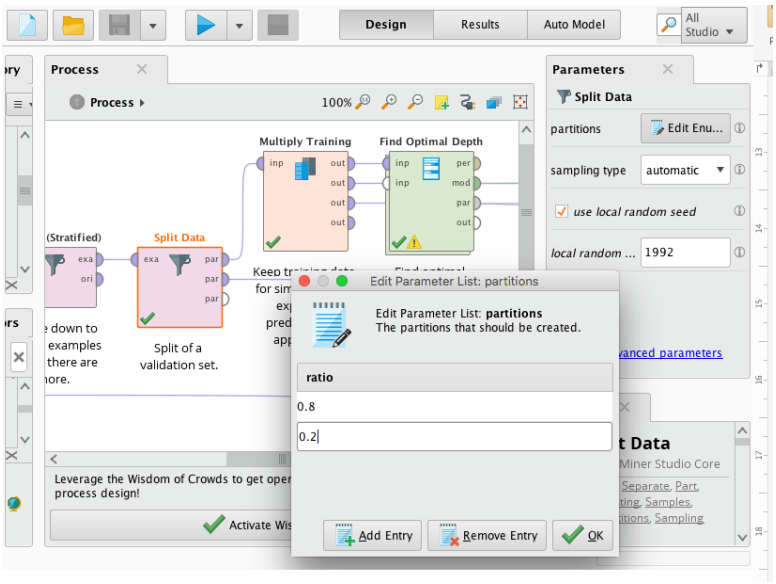


Figura 43. División de la data

La técnica de validación matriz de confusión, permite determinar el número de aciertos al predecir cada una de las opciones mediante la siguiente matriz.

```
Performance:
PerformanceVector [
*****accuracy: 96.88% +/- 0.07% (micro average: 96.88%)
ConfusionMatrix:
True:   Aprobado        Fallido Exonerado       Ninguno
Aprobado:       21660   373     22      29
Fallido:        51      15887   3       0
Exonerado:      1774    6       36273   0
Ninguno:        156     0       0       1221
-----classification_error: 3.12% +/- 0.07% (micro average: 3.12%)
```

Figura 44. Matriz de confusión del modelo

Toda la información contenida en la diagonal principal de la matriz representa aquellos elementos que han sido clasificados correctamente.

Analizando los resultados obtenidos verticalmente del total como Aprobado, 21660 fueron clasificados como Aprobados, 51 como fallidos, 1774 como Exonerados y 156 como Ninguno. Lo clasificado como Fallido, Exonerado y Ninguno se le denomina FALSOS NEGATIVOS, es decir, dentro de los Aprobados reales, los que han sido clasificados incorrectamente por el modelo como

Fallido, Exonerado y Ninguno. El porcentaje que se obtiene a partir de la relación entre los correctamente clasificados y los falsos negativos se denomina cobertura.

$$Cobertura_{clase}(\%) = \frac{Acierto_{clase}}{FalsosNegativos_{clase}} * 100$$

Por lo tanto, las coberturas de las diferentes clases se calculan:

$$Cobertura_{aprobado}(\%) = \frac{21660}{23641} * 100 = 91{,}62\%$$

$$Cobertura_{fallido}(\%) = \frac{15887}{16266} * 100 = 97{,}67\%$$

$$Cobertura_{exonerado}(\%) = \frac{36273}{36298} * 100 = 99{,}93\%$$

$$Cobertura_{ninguno}(\%) = \frac{1221}{1250} * 100 = 97{,}68\%$$

También se puede leer la matriz de manera horizontal, del total predicho como Aprobado, el modelo marcó 21660 que eran verdaderos Aprobados, 373 que eran verdaderos fallidos, 22 que eran verdaderos Exonerados y 29 que eran verdaderos Ninguno. Y de igual manera el resto de clases. Los Fallidos, Exonerados y Ningunos son FALSOS NEGATIVOS, es decir, que han sido predichos como Aprobados, cuando realmente no lo eran. El porcentaje que se obtiene a partir de la relación entre los correctamente clasificados y los falsos negativos se denomina precisión.

$$Precisión_{clase}(\%) = \frac{Acierto_{clase}}{FalsosNegativos_{clase}} * 100$$

Los resultados en la precisión de ceda clase son:

$$Presición_{aprobado}(\%) = \frac{21660}{22084} * 100 = 98{,}08\%$$

$$Presición_{fallido}(\%) = \frac{15887}{15941} * 100 = 99{,}66\%$$

$$Presición_{exonerado}(\%) = \frac{36273}{38053} * 100 = 95{,}32\%$$

$Presición_{ninguno}(\%) = \frac{1221}{1377} * 100$=88,67%

Ambos cálculos son importantes al momento de analizar el modelo creado. Puesto que la cobertura indica el porcentaje de clasificación del modelo y la precisión el grado de predicción.

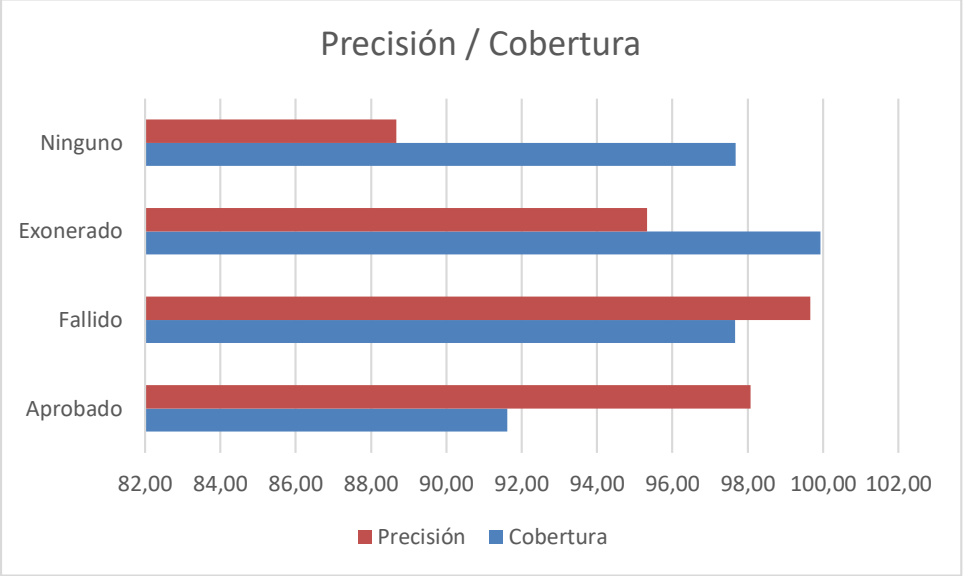


Figura 45. Porcentaje de precisión y cobertura del modelo

La exactitud del modelo (*accuracy*) se calcula a partir del número de elementos que han sido clasificados correctamente dividido para el número total de registros.

$$Exactitud_{modelo}(\%) = \frac{Aciertos}{Total\ elementos} * 100$$

$Exactitud_{modelo}(\%) = \frac{75041}{77455} * 100$ = 96.88%

Donde se observa que el nivel de confianza ofrecido por el modelo creado de acuerdo a RapidMiner es del 96,88% similar al obtenido mediante el cálculo, considerándolo bastante aceptable.

Además, Auto Model de RapidMiner mediante el operando Performance, permite conocer y comparar la eficiencia del modelo construido con respecto a otros modelos predictivos.

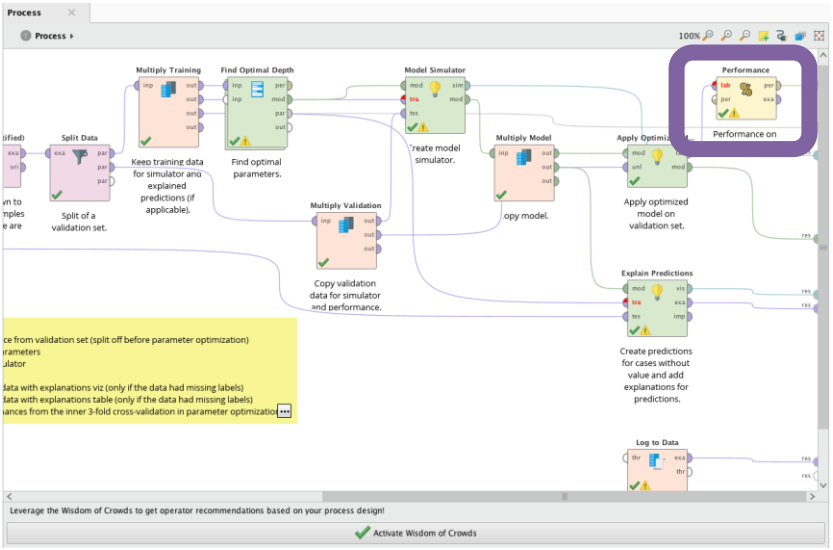
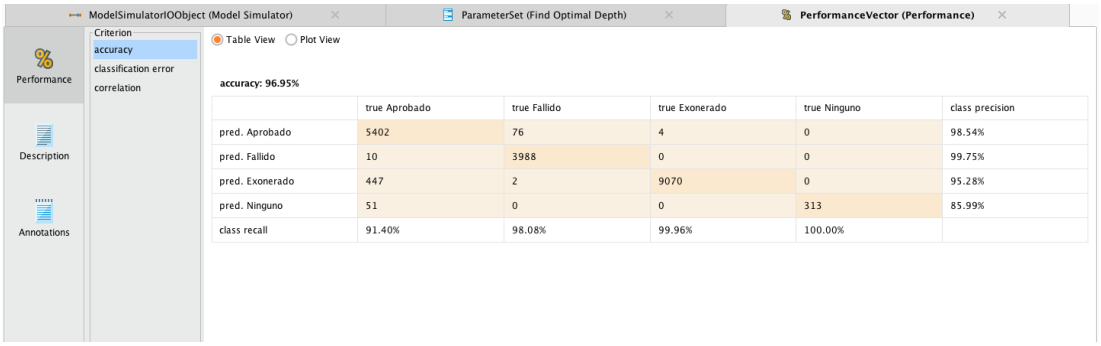


Figura 46. Proceso de evaluación del modelo

A través de esta técnica de evaluación del modelo, se compara los valores predichos por el modelo con los valores reales y así determinar la precisión del mismo.



	true Aprobado	true Fallido	true Exonerado	true Ninguno	class precision
pred. Aprobado	5402	76	4	0	98.54%
pred. Fallido	10	3988	0	0	99.75%
pred. Exonerado	447	2	9070	0	95.28%
pred. Ninguno	51	0	0	313	85.99%
class recall	91.40%	98.08%	99.96%	100.00%	

Figura 47. Matriz de precisión

Según la matriz de precisión se observa un nivel de confianza del 96,95%, considerándolo bastante aceptable.

Además, se puede visualizar en la siguiente figura que se tiene un porcentaje de error del 3,05%, considerado mínimo.

classification_error: 3.05%

	true Aprobado	true Fallido	true Exonerado	true Ninguno	class precision
pred. Aprobado	5402	78	4	1	98.49%
pred. Fallido	10	3988	0	0	99.75%
pred. Exonerado	447	0	9070	0	95.30%
pred. Ninguno	51	0	0	312	85.95%
class recall	91.40%	98.08%	99.96%	99.68%	

Figura 48. Matriz de error de clasificación

4.2 Metodología para ejecutar la propuesta

Un modelo de minería de datos permite descubrir conocimiento oculto dentro de grandes volúmenes de datos, mediante la determinación de patrones de comportamiento de estos. Mediante el proceso de extracción de la información de un conjunto de datos y transformación se consiguió una estructura comprensible de los datos aplicada en la minería de datos.

Existen diversas técnicas y algoritmos dentro de la minería de datos que permiten explorar los datos, de manera automática o semiautomática, para encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos de una manera comprensible para el usuario y en algunos casos que de soporte para la toma de decisiones.

El algoritmo DECISION TREE facilita la interpretación de la decisión adoptaba permitiendo comprender el conocimiento adquirido luego de aplicar el modelo, principalmente por su representación gráfica. Además, reduce el número de variables independientes para tener respuestas más concisas y entendibles.

Dentro del campo educativo, un modelo de minería de datos permitiría crear un sistema de apoyo al sistema de tutorías académicas principalmente.

Es así, que se considera que el modelo creado mediante este proyecto debe ser implementado, creando un sistema de apoyo al proceso de tutorías académicas, donde un profesor tutor pueda ingresar los factores de cada estudiante que asista a una tutoría académica y en base a los patrones

de comportamiento definidos en este modelo, éste pueda predecir cuál va a ser el rendimiento académico del estudiante en una determinada materia.

RapidMiner permite crear un simulador de modelos, parecido al sugerido para un sistema de apoyo a tutoría académicas. Aquí, se puede observar cómo se pueden ir variando los factores escogidos en el análisis y el simulador predice el rendimiento académico del estudiante.

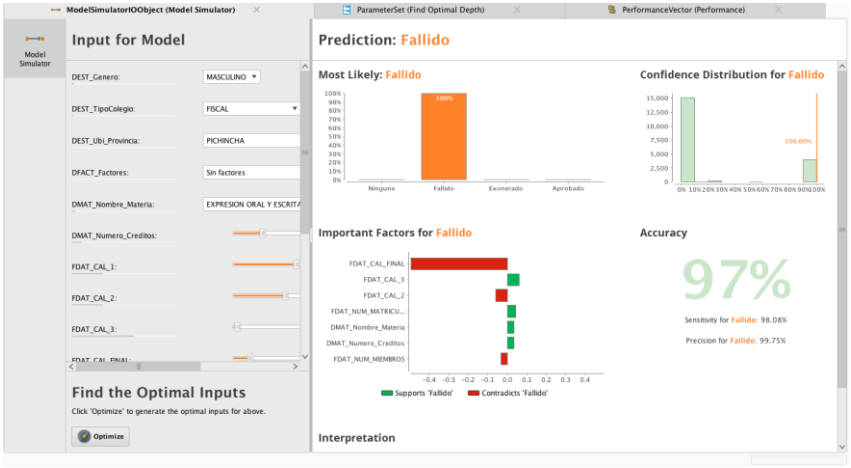


Figura 49. Simulador del modelo con parámetros óptimos

Por ejemplo, con los parámetros óptimos del modelo se tendría como resultado lo mostrado en la figura anterior, que indica que de acuerdo a los parámetros seleccionados en la izquierda se tendría la reacción del modelo a la derecha. El modelo indica que tiene mucha confianza en que la predicción correcta es Fallido. La confianza para esta decisión es alta con 100.00%. Sin embargo, el valor de FDAT_CAL_FINAL no respalda esta decisión. Tenga en cuenta que el 96.95% de todas las predicciones hechas por este modelo son correctas. Cuando el modelo dice Fallido, cubre el 98.08% de esos casos. Y es correcto con el 99.75% de todas las predicciones para la clase Fallido.

Si se cambiara los valores de los parámetros de entrada con el mismo modelo se consigue respuesta como la mostrada a continuación.

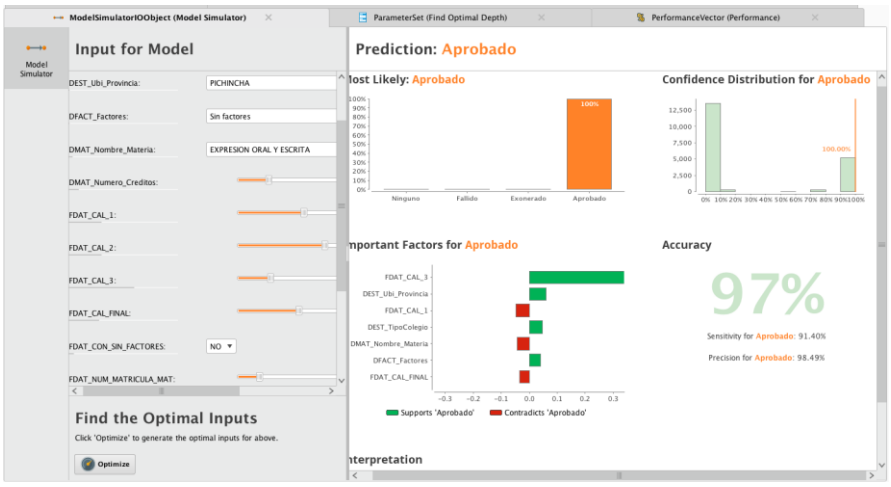


Figura 50. Ejemplo de simulador de modelo

En este caso el modelo tiene mucha confianza en que la predicción correcta es Aprobado. La confianza para esta decisión es alta con 100.00%. El mayor respaldo para esta decisión proviene de FDAT_CAL_3. Tenga en cuenta que el 96.95% de todas las predicciones hechas por este modelo son correctas. Cuando el modelo dice Aprobado, cubre el 91.40% de esos casos. Y es correcto con 98.49% de todas las predicciones para la clase Aprobado.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

El análisis de grandes volúmenes de datos es un elemento fundamental en la toma de decisiones dentro de las empresas, pero gracias a sus ventajas se ha extendido al campo de la educación. Mediante este proyecto se ha logrado aprovechar la información almacenada en bases de datos de la Escuela Politécnica Nacional (EPN) para determinar patrones de comportamiento, en busca de beneficios para los estudiantes de la Escuela de Formación de Tecnólogos (ESFOT).

La integración de diferentes fuentes de datos, se realizó diseñando un modelo multidimensional, que brindó la posibilidad de establecer relaciones entre los datos recibidos. Así, se obtuvo un esquema de la data a cargar en la herramienta ETL. Además, los datos analizados de los estudiantes de la ESFOT, permitieron determinar los requerimientos necesarios para realizar la limpieza y transformación, y así determinar la herramienta ETL con mejores prestaciones, para la información proporcionada.

En la selección de la herramienta ETL se encontró una amplia cantidad de software que permiten realizar este proceso, pero mediante una comparación se determinó que la más factible para el diseño y creación de una bodega de datos en el proyecto es Data Integration Pentaho. Esta herramienta dio la posibilidad de tomar las fuentes de información otorgadas por Escuela Politécnica Nacional, para armar el modelo multidimensional planteado, consiguiendo obtener datos validados, organizados y optimizados, con el propósito de minimizar los fallos en etapas posteriores del proceso, como: existencia de campos o valores nulos, tablas de referencia inexistentes, etc.

Existen diversas técnicas y algoritmos dentro de la minería de datos que permiten explorar los datos, de manera automática o semiautomática, para encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos educativos. En el proyecto se escogió al algoritmo DECISION TREE, por su manera comprensible de presentar los resultados al usuario, por la posibilidad de manejar datos numéricos y categóricos, como múltiples respuestas, facilitando la interpretación de los datos para dar soporte a la toma de decisiones. Además, DECISION TREE es un algoritmo disponible en la herramienta RapidMiner, permitiendo visualizar a través de un simulador los resultados del modelo, que hace aún más simple la interpretación.

El modelo de minería de datos creado mediante el algoritmo DECISION TREE permitió determinar los patrones que afectan el desempeño de los alumnos, en base a factores establecidos en datos históricos. El modelo de minería de datos es el resultado del análisis de información, en este caso, de los estudiantes de la Escuela de Formación de Tecnólogos, que se encuentra almacenada en la base de datos de la Escuela Politécnica Nacional. En estos datos se identificó patrones de comportamiento del alumnado que influyen en su aprovechamiento académico como: calificaciones obtenidas en cada uno de los parciales, el número de créditos de la asignatura, el género, el total de ingresos del estudiante, el número de matrícula de la asignatura, entre otros.

La evaluación del modelo analítico-predictivo creado se realizó a través del uso de la técnica de validación matriz de confusión, implementada en la herramienta minería de datos RapidMiner mediante el operando validación cruzada, considerando un porcentaje de la data como grupo de entrenamiento y otro como testeo. A través, de esta matriz se determinó un nivel de confianza del 96,9 % en el resultado predictivo.

Además, se pudo determinar la cobertura de los atributos del modelo, mostrando el porcentaje del total de aciertos, de acuerdo al total de datos reales perteneciente a dicho atributo, es decir, que

se indica el grado de clasificación del modelo por atributos. De acuerdo a los resultados obtenidos, el atributo mejor clasificado es el “Exonerado”, indicando que del total de estudiantes que se han exonerado una materia el 99,93% han sido clasificados en el mismo atributo, según el modelo planteado. Por otro lado, la clase “Aprobado” presenta el menor porcentaje de clasificación, indicando que del total de estudiantes que han aprobado una materia el 91,62% han sido clasificados en la misma clase.

También, se encontró la precisión de los atributos, determinado por la relación entre el total de aciertos con respecto al total de datos predichos en ese atributo, es decir, mide el grado de predicción de un atributo determinado. Con respecto a los resultados obtenidos, la predicción de que un estudiante “Falle” una asignatura tiene una precisión del 99.66%, que se exonere será del 95.32% y de que apruebe 98,08%, siendo valores aceptables para un modelo.

5.2.Recomendaciones

El análisis de datos educativos tiene un amplio campo de aplicación puesto que con el mismo conjunto de datos se puede observar patrones de comportamiento con respecto a otros atributos predecibles, como situación económica del estudiante, materias con más rezago, entre otros.

El modelo de minería de datos diseñado, se recomienda utilizarlo en el diseño de un sistema para mejorar el proceso de tutorías académicas, brindando al profesor tutor una herramienta de apoyo para guiar de una mejor manera al estudiante de acuerdo a los factores que presente. Este sistema, podría utilizar una estructura similar al simulador de modelos de RapidMiner Auto Model, donde el profesor tutor ingresa los factores que posee un estudiante y el sistema predice cuál sería su rendimiento académico.

Además, mediante este modelo se puede establecer una guía para el proceso de acreditación de carreras en las instituciones educativas y la toma de decisiones de las autoridades con el propósito de tener mejores procesos de contratación docente.

Analizar datos como encuestas a docentes, podría completar el modelo planteado ayudando al alumnado desde el punto de vista docente a la vez. También, es importante considerar que, si la cantidad de datos aumenta, el modelo predecible tendrá una mejor precisión, presentando resultados con un nivel de confianza mayor, y con mejores prestaciones.

RapidMiner Auto Model es una opción que ahorra tiempos de ejecución y análisis, permitiendo al investiga dedicar más tiempo al análisis de los resultados obtenidos y a obtener mejores modelos.

BIBLIOGRAFÍA

Acosta, C. (2016, noviembre 10). La deserción universitaria bordea el 40% [Articulo de Periódico digital]. Recuperado 16 de octubre de 2017, de http://www.eltelegrafo.com.ec/noticias/sociedad/4/la-desercion-universitaria-bordea-el-40

Alharbi, Z., Cornford, J., Dolder, L., & De, L. I. (2016). Using data mining techniques to predict students at risk of poor performance (pp. 523-531). Presentado en Proceedings of 2016 SAI Computing Conference, SAI 2016. https://doi.org/10.1109/SAI.2016.7556030

Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study (Vol. 82, pp. 65-71). Presentado en Procedia Computer Science. https://doi.org/10.1016/j.procs.2016.04.010

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

Badr, G., Algobail, A., Almutairi, H., & Almutery, M. (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department (Vol. 82, pp. 80-89). Presentado en Procedia Computer Science. https://doi.org/10.1016/j.procs.2016.04.012

Blog, A. (2016). Los 8 algoritmos de Aprendizaje Automático o Data Mining más usados | Parte 1 | Arame Blog. Recuperado 1 de noviembre de 2017, de http://www.arame.com.mx/blog/8-algoritmos-aprendizaje-automatico-data-mining-mas-usados-arame-blog/

Bravo, D. (2016, noviembre 21). Los estudiantes aún desertan de las carreras universitarias [Articulo de Periódico digital]. Recuperado 16 de octubre de 2017, de

http://www.elcomercio.com/tendencias/estudiantes-desercion-carreras-universidad-educacion.html

Camana, R. (2016). Aplicación de técnicas de minería de datos educacionales, para predecir la deserción académica.

Cedano, I. J. Á. H., & Castro, M. J. A. (2015). *Modelo de minería de datos para identificación de patrones que influyen en el aprovechamiento académico*. Tecnológico Nacional de México Instituto Tecnológico de La Paz, México.

EPN. (2018, marzo 26). Historia Escuela Politécnica Nacional. Recuperado 26 de marzo de 2018, de http://www.epn.edu.ec/

EPN. (2012). Informe-de-Gestión-2011.pdf. Recuperado 25 de octubre de 2017, de http://www.epn.edu.ec/wp-content/uploads/2008/02/Informe-de-Gesti%C3%B3n-2011.pdf

EPN. (2013). Informe-de-Gestión-2012.pdf. Recuperado 25 de octubre de 2017, de http://www.epn.edu.ec/wp-content/uploads/2013/08/Informe-de-Gesti%C3%B3n-2012.pdf

EPN. (2016). Informe-Completo-de-Rendición-de-Cuentas-2015.pdf. Recuperado 25 de octubre de 2017, de http://www.epn.edu.ec/wp-content/uploads/2016/03/Informe-Completo-de-Rendici%C3%B3n-de-Cuentas-2015.pdf

EPN. (2017). Rendicion-EPN-2016-VF-152-pags.pdf. Recuperado 25 de octubre de 2017, de http://www.epn.edu.ec/wp-content/uploads/2017/05/Rendicion-EPN-2016-VF-152-pags.pdf

Garcia, F. (2013). *Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA)*. Universidad de Granada, Granada.

González, M. S. (2016). Patrones de comportamiento. *El campamento de Dios*.

Gonzalez, R., & Pomares Quimbaya, A. (2012). *La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería*.

Iam-On, N., & Boongoen, T. (2017a). Generating descriptive model for student dropout: a review of clustering approach. *Human-centric Computing and Information Sciences*, *7*(1). https://doi.org/10.1186/s13673-016-0083-0

Iam-On, N., & Boongoen, T. (2017b). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, *8*(2), 497-510. https://doi.org/10.1007/s13042-015-0341-x

Jorge Bustillos. (18:53:28 UTC). *Comparativa herramientas ETL*. Software. Recuperado de https://es.slideshare.net/JorgeCarlos3/comparativa-herramientas-etl

Latino BI. (2013). Latino BI Consulting - Inteligencia de Negocios, Business Intelligence, Bodega de datos, Planeacion y Presupuestacion EP, DataQuality, Bases de Datos - Latino BI Consulting - The Business Intelligence Company. Recuperado 1 de noviembre de 2017, de http://www.latino-bi.com/espanol/fundamentos-bi/introduccion-al-bi.php

López, D., & Galindo, Y. (2013). *Estudio del Pentaho Data Integration en los procesos de integración de datos (ETL)*. Universidad Central "Marta Abreu" de Las Villas, Santa Clara.

Montoya, N. P. M. (2005). ¿Qué es el estado del arte? *Ciencia & Tecnología para la Salud Visual y Ocular*, (5), 73-75. https://doi.org/10.19052/sv.1666

Ordoñez, K. (2013). *Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL*. Universidad Técnica Particular de Loja, Loja.

Parraga, V. (2015). *Diseño de un sistema de tracking y mapeo de dispositivos GTB gestionada a*

través de una aplicación web. Escuela Politecnica Nacional, Quito. Recuperado de http://bibdigital.epn.edu.ec/handle/15000/11305

Popoola, S. I., Atayero, A. A., Badejo, J. A., John, T. M., Odukoya, J. A., & Omole, D. O. (2018). Learning analytics for smart campus: Data on academic performances of engineering undergraduates in Nigerian private university. *Data in Brief*, *17*, 76-94. https://doi.org/10.1016/j.dib.2017.12.059

Rivadera, G. R. (2012). La metodología de Kimball para el diseño de almacenes de datos (Bodega de datoss). *Universidad Catolica de Salta*, 56–71.

Sarmiento, J. C. C. (2012). Construcción y poblamiento de una bodega de datos basado en el paradigma de bases de datos objeto relacional. *Prospectiva*, *9*(1), 69–77.

Siraj, F. (2016). Modeling Academic Achievement of UUM Graduate Using Descriptive and Predictive Data Mining. En *Advanced Computer and Communication Engineering Technology* (pp. 609-620). Springer, Cham. https://doi.org/10.1007/978-3-319-24584-3_52

Stark, K. (2016, julio 13). Técnicas de Data Mining. Recuperado 1 de noviembre de 2017, de http://www.evaluandosoftware.com/tecnicas-data-mining/

Villena, J. (2016, agosto 8). CRISP-DM: La metodología para poner orden en los proyectos de Data Science. Recuperado 31 de octubre de 2017, de https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science

Winter, R., Zhao, J. L. & Aiero, S. (2012). *Global perspectives on design science research: 5th international conference, Desrist 2010, St. Gallen, Switzerland, June 4-5, 2012 ; proceedings*. Berlin: Springer.