ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE

INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCION DEL TÍTULO DE MAGISTER EN GESTION DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA: ANÁLISIS DE DATOS A TRAVÉS DE DATA MINING DEL PROCESO DE ADMISIÓN A LA EDUCACIÓN SUPERIOR EN ECUADOR

Autora: Ing, Silvana Magally Guala Acuña.

Director: Ing. Jaime Vinueza, MBA

2017

ii

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**ANÁLISIS DE DATOS A TRAVÉS DE DATA MINING DEL PROCESO DE ADMISIÓN A LA EDUCACIÓN SUPERIOR EN ECUADOR**" realizado por la Ing. ***SILVANA MAGALLY GUALA ACUÑA,*** ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar a la Ing. ***SILVANA MAGALLY GUALA ACUÑA*** para que lo sustente públicamente.

Quito, 4 de septiembre del 2017

x ----------

Ing. Jaime Vinueza, MBA

DIRECTOR

iii

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORÍA DE RESPONSABILIDAD

Yo, ***SILVANA MAGALLY GUALA ACUÑA***, con cédula de identidad N° **0503262180**, declaro que este trabajo de titulación "**ANÁLISIS DE DATOS A TRAVÉS DE DATA MINING DEL PROCESO DE ADMISIÓN A LA EDUCACIÓN SUPERIOR EN ECUADOR"** ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Quito, 4 de septiembre del 2017

Ing, Silvana Magally Guala Acuña.

C.C. 0503262180

iv

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORIZACIÓN (PUBLICACIÓN BIBLIOTECA VIRTUAL)

Yo, ***SILVANA MAGALLY GUALA ACUÑA***, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca Virtual de la institución el presente trabajo de titulación **"ANÁLISIS DE DATOS A TRAVÉS DE DATA MINING DEL PROCESO DE ADMISIÓN A LA EDUCACIÓN SUPERIOR EN ECUADOR"** cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Quito, 4 de septiembre del 2017

Ing, Silvana Magally Guala Acuña.

C.C. 0503262180

DEDICATORIA

A mis padres Miguel y Mariana, quienes me han brindado su inmenso cariño durante toda mi vida y con sus palabras de aliento me han motivado he inspirado a superarme y alcanzar esta nueva meta.

A mi hermana Enith, y mi sobrino Esteban, por estar siempre a mi lado apoyándome incondicionalmente.

A mi Abuelita, mis tíos Marco, Ived, Lorena, y a toda mi familia por todos sus consejos, son lo mejor que Dios me ha dado.

AGRADECIMIENTO

Por aquellos buenos momentos que compartimos durante la colegiatura, gracias a todos mis amigos por su apoyo.

Por los conocimientos transmitidos, a todos los catedráticos de la maestría, mi más sincero agradecimiento.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ACRÓNIMOS Y ABREVIATURAS

LOES	Ley Orgánica de Educación Superior
SNNA	Sistema Nacional de Nivelación y Admisión.
IES	Institución de Educación Superior
CES	Consejo de Educación Superior
CEAACES	Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior.
SENESCYT	Secretaria de Educación Superior Ciencia y Tecnología
MINEDUC	Ministerio de Educación
ETL	Extract, Transform and Load.
BI	Business Inteligence
ENES	Examen Nacional de Educación Superior
EBS	Enterprise Service Bus
SEMMA	Metodología aplicada en minería de datos que proviene de las iniciales de Sample, Explore, Modify, Model and Assess (Muestra, Exploración, Modificación, Modelado y Evaluación).
KDD	Knowledge discovery in databases

RESUMEN

El Sistema Nacional de Nivelación y Admisión (SNNA) es una entidad pública del Estado Ecuatoriano encarga de gestionar el acceso a las Instituciones de Educación Superior de los ciudadanos que deseen continuar sus estudios superiores, el mismo se rige bajo los principios de meritocracia, transparencia e igualdad de oportunidades, para ello se apoya en una plataforma informática que automatiza los procesos que realizan los usuarios generando un gran volumen de datos que probablemente contiene información valiosa, y es en este punto donde esta investigación analiza la data del SNNA para encontrar patrones o características de los aspirantes que reprueban el examen de acceso a la educación superior aplicando técnicas de *Data Mining* con modelos supervisados como arboles de decisión y regresiones llegando a determinar varios factores que influyen en el desempeño de las personas que rinden el examen como la edad, pues el 75% del universo que se encuentra entre 17 y 19 años tienen mayor probabilidad de aprobar el examen, reduciéndose drásticamente conforme avanza su edad, el limitado acceso a internet, la dedicación de tiempo parcial a trabajar y la falta de aspiraciones por continuar sus estudios superiores, estos resultados se visualizan gráficamente mediante un Business Intelligence que muestra varios indicadores que servirán para el diseño de políticas públicas que reduzcan las brechas actuales y democraticen el acceso a la Educación Superior.

PALABRAS CLAVE

- **EDUCACIÓN SUPERIOR**
- **DATA MINING**
- **ARBOLES DE DECISIÓN**
- **REGRESIONES**
- **BUSINESS INTELIGENCE**

ABSTRACT

The Sistema Nacional de Nivelación y Admisión is a public entity of the Ecuadorian State in charge of managing access to Higher Education Institutions of citizens who wish to continue their studies, it is governed under the principles of meritocracy, transparency and equal opportunities, It is supported on a computer platform that automates the processes performed by users, generating a large volume of data that probably contains valuable information, and it is at this point that this research analyzes the data of the SNNA to find patterns or characteristics of the candidates who fail the entrance exam to higher education by applying *Data Mining* techniques with supervised models as decision trees and regressions, determining several factors that influence the performance of test takers such as age, since 75 % of the universe between 17 and 19 years have limited access to the internet, part-time dedication to work, and lack of aspirations to continue their studies. These results are visualized graphically through a Business Intelligence that shows several indicators that will serve to design public policies that reduce current gaps and democratize access to Higher Education.

KEYWORDS

- **HIGHER EDUCATION**
- **DATA MINING**
- **DECISION TREES**
- **REGRESSIONS**
- **BUSINESS INTELIGENCE**

CAPÍTULO I

1. MARCO TEÓRICO

1.1. ANTECEDENTES DE LA INVESTIGACIÓN

La evolución que ha existido en los últimos años en las tecnologías de la información y comunicación han dado paso a que las organizaciones tengan la facilidad de procesar grandes cantidades de datos. Esto se refleja por ejemplo en las empresas de servicios públicos o grandes comercios, los cuales almacenan información en bases de datos, las analizan y obtienen información que resulta relevante para mejorar su gestión

Precisamente para identificar la información que se cree más importante se utiliza el *data mining* o minería de datos, porque permite identificar patrones y tendencias que sirven para comprender la información relacionada de una forma sencilla, efectiva y al tiempo que se requiere (PÉREZ & SANTÍN, 2008), lo que beneficia a las mismas organizaciones porque representan la base de los cambios o refuerzos que se requieren para mejorar su actuación. En el caso específico de las instituciones públicas, esta identificación de patrones y tendencias les permite tomar decisiones oportunas que benefician a toda la comunidad.

Así lo demuestra el estudio "Detección de patrones de bajo rendimiento académico y deserción estudiantil con técnicas de minería de datos" realizado por TIMARÁN (2010) del Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño San Juan de Pasto, en el cual el autor concluyó afirmando que "se demostró que Tariy KDD es una herramienta fiable, que puede ser utilizada en cualquier proyecto de Minería de Datos" (pág. 5). Además, de acuerdo a los patrones obtenidos se precisa que la Universidad tome medidas relacionadas con el seguimiento de los estudiantes de acuerdo a los perfiles publicados con la finalidad de prevenir que bajen su rendimiento y disminuir la posibilidad de deserción.

También se considera el estudio "Análisis de satisfacción de universitarios mediante la minería de datos" de los autores SILVA, DOMÍNGUEZ, CORTÉS,

CASTORENA Y VÁSQUEZ (2007) en el cual concluyen que es pertinente que la Universidad de Coahuila establezcan estrategias y realicen modificaciones pertinentes que fortalezcan y mejoren los servicios institucionales que brindan con la finalidad de que brinden un mejor servicio educativo.

Finalmente, se cita la investigación denominada "El *data mining* y su incidencia en la toma de decisiones del catastro de establecimientos y la emisión de los permisos de funcionamiento por parte de la Dirección Provincial de Salud de Cotopaxi" del autor CARVAJAL (2012), en el que le permite concluir que "la solución efectiva para detectar problemas en los pagos de emisión de permisos es un *data mining*. El cual posee todas las características necesarias para administrar, predecir y mejorar los procesos" (pág. 113).

Todos los estudios citados refieren la importancia que tiene la aplicación de la minería de datos para identificar patrones o perfiles que facilitan la toma de decisiones para mejorar o fortalecer los procesos que se realizan en las instituciones y de esta forma mejorar su servicio, lo que concuerda con el objetivo de este estudio.

1.2. FUNDAMENTACIONES

1.2.1. Fundamentación legal

Este estudio se fundamenta legalmente en la Ley Orgánica de Educación Superior (LOES), que en el artículo 2 señala "esta Ley tiene como objeto definir sus principios, garantizar el derecho a la educación superior de calidad que propenda a la excelencia, al acceso universal, permanencia, movilidad y egreso sin discriminación alguna" (ASAMBLEA NACIONAL, 2010). Para lo cual instituye la regulación del ingreso a las instituciones de educación superior por medio del Sistema de Nivelación y Admisión, al que necesariamente deben someterse todos los estudiantes que desean ingresar a universidades estatales.

Por ello, resulta indispensable conocer cuáles son los patrones y tendencias que existen en relación al proceso de admisión a la educación superior que mantiene el SENESCYT con el fin de establecer estrategias que permitan mejorarlo.

1.3. CATEGORÍAS FUNDAMENTALES

Las relaciones que se puedan encontrar entre el proceso de admisión y la aplicación de técnicas de minería de datos brindaran como resultado un conjunto de variables independientes que expliquen la influencia sobre la variable dependiente, que en este caso es la aprobación del examen de acceso a la Educación Superior (Ver Figura 1).

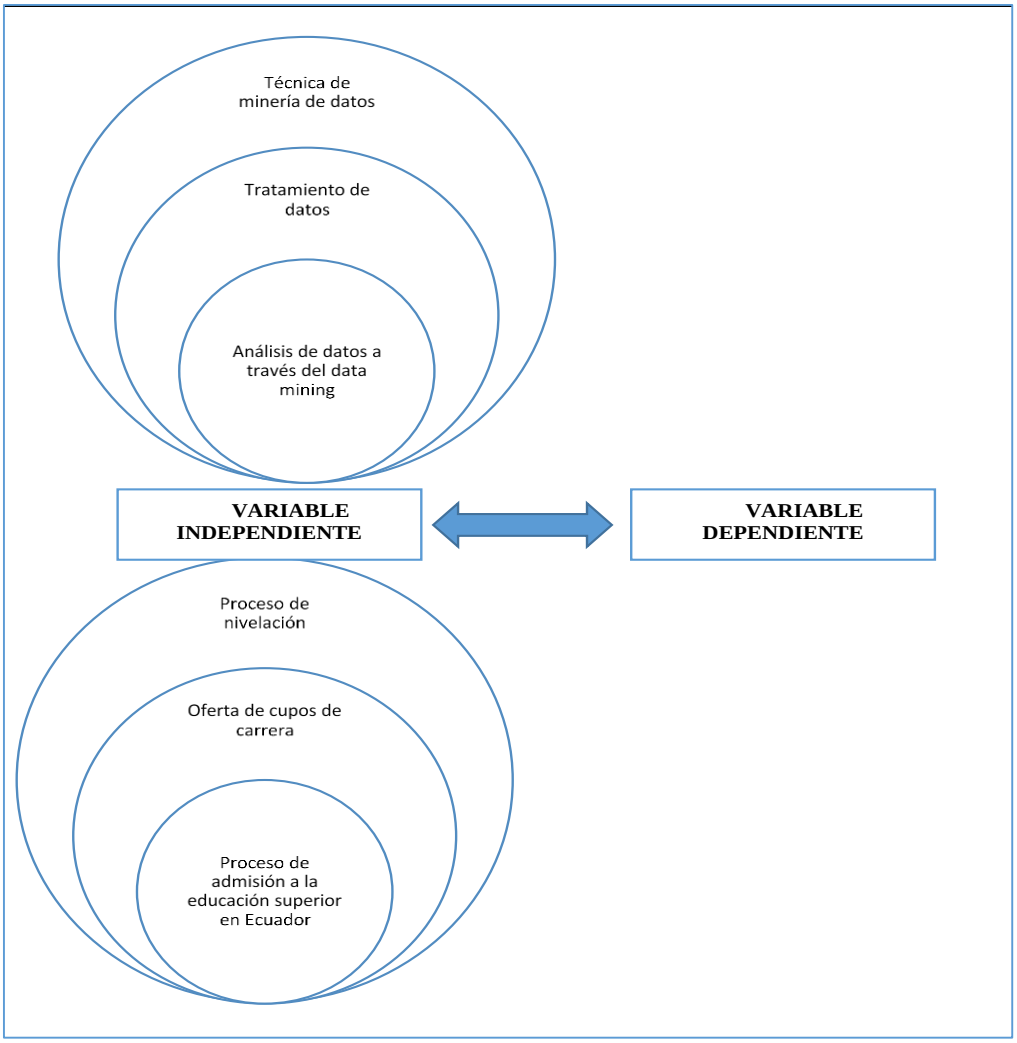


Figura 1 Categorías fundamentales

1.4. SEÑALAMIENTO DE VARIABLES

Variable Independiente: Análisis de datos a través de *data mining*

Variable Dependiente: Proceso de admisión a la educación superior en Ecuador.

1.5. MARCO TEÓRICO DE LAS VARIABLES

1.5.1. Variable Independiente:

Análisis de datos a través del *data mining*

1.5.2. Data mining

Según PÉREZ Y SANTÍN (2008), data mining se define básicamente como: un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o *data mining* (pág. 1).

Lo que sirve para identificar patrones o tendencias que permiten entender de mejor forma la información con la que se cuenta y para predecir el comportamiento de los informantes en el futuro. Para el efecto, las herramientas que utiliza la minería de datos para extraer los patrones o tendencias es preciso que exista previamente un almacén de datos.

El almacén de datos es el sistema de información central en todo este proceso. Un almacén de datos es una colección de datos orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar en la toma de decisiones. Un almacén de datos es un conjunto de datos históricos, internos o externos y descriptivos de un contexto o área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas (PÉREZ & SANTÍN, 2008, pág. 2).

Para realizar la carga y mantenimiento de un almacén de datos existe un sistema compuesto por un procedimiento, hardware y software que se llama sistema ETL, por sus siglas en inglés (Extract, Transform and Load). Este sistema tiene la función de leer los datos, incorporar los datos externos, crear claves, integrarlos, realizar agregaciones, limpieza y transformaciones de los mismos, así como crear y mantener los metadatos, planificar la carga y mantenimiento e indización y realizar pruebas de calidad (VIEIRA & ORTIZ, 2009).

En el siguiente diagrama se puede observar las funciones de este sistema (Ver Figura 2)

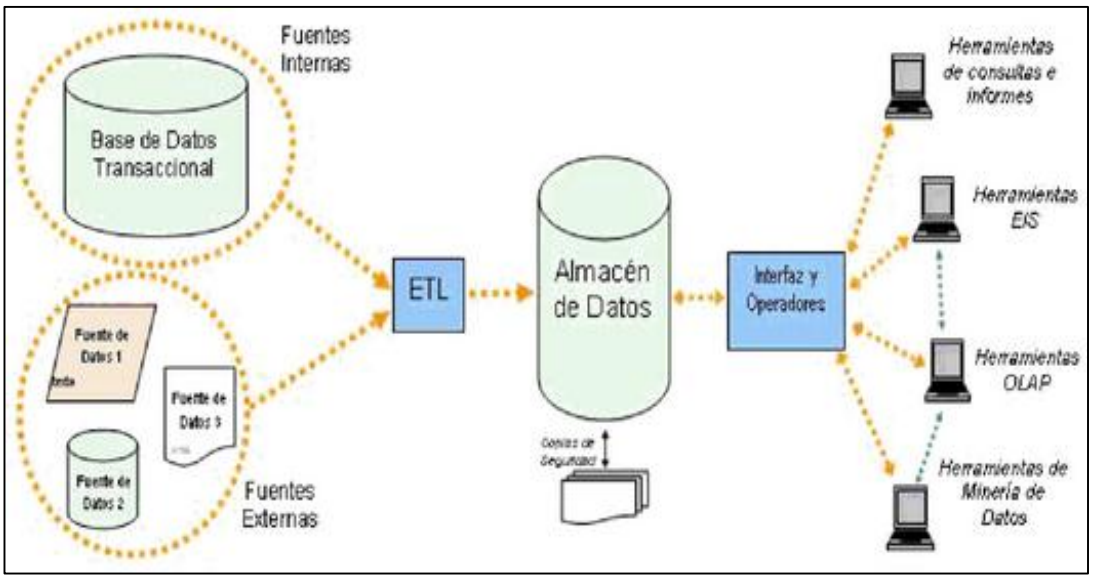


Figura 2. Sistema del almacén de datos
Tomado de: (PÉREZ & SANTÍN, 2008, pág. 3)

Las características que necesariamente deben tomarse en cuenta en el proceso de *data mining* son: "1. Precisión (sin errores de medición o digitación); 2. Consistencia (con coherencia); 3. Completitud (sin atributos faltantes); 4. Relevancia (concerniente al problema); 5. No redundancia (sin duplicación de la misma información)" (VIEIRA & ORTIZ, 2009, pág. 19). A partir de ello, se establecen como principios: la calidad de datos, la información que se encuentran en los datos y el hecho de que el trabajo debe realizarse sobre esos datos.

1.5.3. Tratamiento de los datos

Según VIEIRA Y ORTIZ (2009) el tratamiento de los datos parte de la definición de la población, en este sentido es preciso que se consideren tres categorías: demográficas, comportamentales y psicológicas. A partir de ello se deben destacar tres objetivos comunes: "datos para prospección, evaluación del riesgo y metas respecto a los clientes" (2009, pág. 27).

Una vez que está definida la población se requiere obtener la muestra representativa, compuesta por un número de registros que se van a considerar en el estudio, en vista de que la base de datos es muy grande para considerar a todos.

Al obtener la muestra con la que se va a trabajar se procede a la manipulación de datos, la que consiste en: "tratar los errores, valores aberrantes y valores faltantes" (VIEIRA & ORTIZ, 2009, pág. 31), para el efecto se toman en cuenta la clasificación de los datos cuantitativos y cualitativos.

Los datos cualitativos son expresados en categorías y básicamente son utilizados en segmentación y clasificación. Los datos cuantitativos son expresados numéricamente y se presentan en cuatro escalas diferentes: nominal, ordinal, por intervalos y proporcional. Para un dato cualitativo es más fácil detectar un error, o un *outlier*, basta verificar si los valores encontrados en la muestra corresponden a los valores posibles, en caso de que esto no ocurra para algún elemento, se puede descartarlo, o alternativamente sustituirlo por la "moda". En el caso de los datos expresados en escalas nominales u ordinales se puede aplicar el mismo procedimiento. (VIEIRA & ORTIZ, 2009, pág. 32)

A paso seguido se procede a transformar los datos, para lo cual se utiliza:

Agregación (reducir el número de valores mediante alguna agregación, por ejemplo, substituir datos diarios por medias semanales). **Razones** (generar una nueva variable a partir de la razón de dos variables). **Codificación** (transformar datos cualitativos en cuantitativos, por ejemplo fechas en formato dd.mm.aa inviabiliza operaciones matemáticas; se establece, por tanto, una fecha de referencia a partir de la

cual los días son contados. **Codificación simbólica** (transformar datos cuantitativos en cualitativos, no deja de ser una forma de agregación. Intervalos de variación pasan a ser asociados a una categoría); **reducción de variables** (eliminar variables redundantes o con escaso poder predictivo). **Parametrización** (transformar una variable en otra cuyo dominio de variación sea más adecuado. Por ejemplo, la tipificación. Transformaciones matemáticas (calcula una función de la variable obteniéndose una nueva variable con propiedades más convenientes, por ejemplo, simetrización por medio de la transformación logarítmica. (VIEIRA & ORTIZ, 2009, págs. 32,33)

En resumen, según SANTACRUZ (2015), la preparación de los datos requiere:

- **Selección** que está dada por la recopilación e integración de las fuentes de datos, así como la integración y selección de las variables importantes, y la aplicación de la fórmula para determinar la muestra.
- **Exploración** por medio del uso de técnicas de análisis exploratorio que permiten deducir la distribución de los datos, su proporción y regularidad, con la finalidad de analizar las correlaciones existentes.
- **Limpieza**, a través de detectar valores que resulten atípicos, además, de imputar valores faltantes o perdidos, y eliminar datos que resulten errados o irrelevantes.
- **Transformación** con el uso de técnicas apropiadas de reducción o incremento de las dimensiones, así como de técnicas de discretización y numeración, y escalado simple y multidimensional.

El análisis de los datos involucra:

- Uso de **técnicas predictivas** como: regresión y series temporales, análisis discriminatorios, análisis de la varianza, árboles de decisión, redes neuronales.

- Uso de **técnicas descriptivas** como: *clustering* y segmentación, asociación, dependencia, análisis exploratorio y reducción de la dimensión.

Dentro de la evaluación e interpretación de datos se considera: intervalos de confianza, análisis curva ROC y evaluación de modelos. Y la difusión y uso de modelos incluye visualización y simulación.

1.5.4. Técnica de minería de datos

Según SANTACRUZ (2015) las técnicas de minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas no son más que algoritmos que se aplican sobre un conjunto de datos para obtener determinados resultados, las técnicas más representativas son:

- **Redes neuronales,** el que se trata de un "sistema de interconexión de neuronas en una red" que permite originar un estímulo de salida.
- **Árboles de decisión,** es un modelo que predice, en razón de que a partir de la base de datos con la que se cuenta se construyen diagramas de construcciones lógicas que facilitan la representación y categorización de una serie de condiciones que aparecen sucesivamente, lo que permite resolver un problema.
- **Modelos estadísticos,** están dados por una expresión simbólica a manera de ecuación que se utilizan en todos los diseños experimentales y en la regresión con la finalidad de mostrar los distintos factores que cambian la variable de respuesta.
- ***Clustering,*** denominado también agrupamiento, que es un proceso de agrupación de una serie de vectores de acuerdo a criterios regularmente de distancia. Para lo cual se procura ubicar los vectores de entrada de tal forma que estén más cerca de los que tengan características similares.
- **Algoritmos predictivos o supervisados,** los que predicen un dato o conjunto de datos desconocidos a priori, al considerar otros que son conocidos.

- **Algoritmos no supervisados,** conocidos también como de descubrimiento del conocimiento, lo que hacen es descubrir patrones y tendencias en los datos dispuestos.

En razón de que las bases de datos objeto del análisis son el resultados de los registros obtenidos por el Sistema Nacional de Nivelación y Admisión a lo largo de los procesos aplicados desde el año 2012; en relación a la amplia diversidad de técnicas de *Data Mining* disponibles, es preciso tener claro que no existe una sola técnica para resolver un determinado problema, sino más bien existen formas inteligentes de utilizar una técnica, sin embargo, las técnicas de árboles de decisión, clústeres y reglas de asociación son las más indicadas en el proceso de los datos.

Al respecto, en el capítulo II se presenta un detalle pormenorizado de los métodos que se utilizan en la minería de datos.

1.5.5. Variable Dependiente: Proceso de admisión a la educación superior en Ecuador

El proceso de admisión a la educación superior en Ecuador está concebido según el siguiente conjunto de normas organizadas de forma escalonada tal como se visualiza a continuación (Ver Figura 3**)**

Figura 3 Normativa del SNNA.

El ordenamiento jurídico como un sistema de normas que se encuentran organizadas de forma escalonada según el artículo 425 de la Constitución Ecuatoriana

Constitución de la República del Ecuador.

- **Art. 356:** Segundo inciso señala que, "El ingreso a las instituciones públicas de educación superior se regulará a través de un sistema de nivelación y admisión, definido en la ley. La gratuidad se vinculará a la responsabilidad académica de las y los estudiantes"

Ley Orgánica de Educación Superior (LOES)

- **Art. 3: "**La Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación, SENESCYT, implementará el **Sistema de Nivelación y Admisión** para el ingreso a las instituciones de educación superior públicas. El Sistema de Nivelación y Admisión tendrá dos componentes.

    - **El de admisión** tendrá el carácter de permanente y establecerá un sistema nacional unificado de inscripciones, evaluación y asignación de cupos en función al mérito de cada estudiante.

    - **El componente de nivelación** tomará en cuenta la heterogeneidad en la formación del bachillerato y/o las características de las carreras universitarias."

- **Art. 81:** "El ingreso a las Instituciones Educativas Superiores públicas, estará regulado a través del **Sistema de Nivelación y Admisión**, al que se someterán todos los y las estudiantes aspirantes. Para el diseño de este Sistema, la SENESCYT coordinará con el Ministerio de Educación lo relativo a la articulación entre el nivel bachiller o su equivalente y la educación superior pública"

- **Art. 183:** Funciones SENESCYT, e) "Diseñar, implementar, administrar y coordinar el Sistema Nacional de Información de la Educación Superior en el Ecuador y el **Sistema de Nivelación y Admisión."**

La SENESCYT en coordinación con el Ministerio de Educación establece el Sistema Nacional de Nivelación y Admisión (SNNA), que será el encargado de establecer los procesos y requisitos que esto demanda, para ello la SENESCYT según acuerdo Ministerial Nro. 76, expide el Reglamento del Sistema Nacional de Nivelación y Admisión.

El Reglamento SNNA, tiene por objeto: "…establecer las normas que regulan el Sistema Nacional de Nivelación y Admisión SNNA, a través del cual se establece el proceso que el aspirante debe seguir para conseguir su ingreso en las instituciones de educación superior públicas, una vez concluido el bachillerato, a fin de realizar los estudios correspondientes en los niveles de formación técnica, tecnológica superior y de grado hasta el tercer nivel, mediante la realización de un examen de aptitud y la superación de las distintas modalidades de los cursos de nivelación." (Reglamento del Sistema Nacional de Nivelación y Admisión SNNA, 2013, Art. 1)

El reglamento del SNNA, regula la participación de todos los estamentos que forman parte del Sistema Nacional de Nivelación y Admisión (SNNA), esto es: aspirantes, Instituciones de Educación Superior (IES) y personal académico de nivelación, con la finalidad de realizar los estudios correspondientes en los niveles de formación técnica, tecnológica superior y de grado hasta el tercer nivel, mediante la realización de un examen estandarizado de aptitud y la superación de las distintas modalidades de los cursos de nivelación (SENESCYT, 2013).

El Sistema Nacional de Nivelación y Admisión (SNNA) aplicó del Examen Nacional de Educación Superior (ENES) en el Ecuador por primera vez, en febrero del 2012 en un Plan Piloto con el fin de probar la técnica y definir la forma más adecuada para implementarlo definitivamente. Este plan sirvió para dar paso a que se oficialice el proceso de admisión a partir del segundo semestre del 2012, aplicando este examen el 19 de mayo del mismo año. A partir de esta fecha se ha venido realizado de forma permanente dos veces al año, como se expone a continuación (Ver Tabla 1).

Tabla 1
Periodos en los que se ha realizado el Examen Nacional de Educación Superior

AÑOS	MES	SEMESTRE
2012	Febrero	I Semestre 2012
	Mayo	II Semestre 2012
	Noviembre	I Semestre 2013
2013	Abril	II Semestre 2013
	Septiembre	I Semestre 2014
2014	Marzo	II Semestre 2014
	Septiembre	I Semestre 2015
2015	Marzo	II Semestre 2015
	Septiembre	I Semestre 2016
2016	Junio	II Semestre 2016
	Febrero	I Semestre 2017

Tomado de: (SENESCYT, 2016)

1.1. Proceso de admisión – componentes y características

Es importante señalar que, el SNNA está compuesto por dos grandes componentes: nivelación y admisión y cada uno de ellos contiene varios módulos que permiten gestionar el proceso de acceso a la educación superior tal como describe se describe a continuación (Ver Figura 4).

Figura 4 Macro procesos del sistema de admisión

1.1.1. Oferta de cupos de Carrera

Para dar inicio a este proceso, la SENESCYT realiza el pedido al IES, con el fin de que se proporcione información por escrito sobre el número de cupos por carrera y sobre todo de la nivelación de carrera. La SENESCYT debe cumplir con este requisito mínimo con 60 días de antelación a cada convocatoria del ENES.

Una vez que la SENESCYT ha solicitado el pedido formal de cupos, las IES trascurrido un máximo 15 días, deberá realizar el ingreso de los datos a la plataforma de admisión, esta se hará utilizando y respetando uso de usuario y contraseña de acuerdo a lo establecido por la LOES, en el caso de que se dé un uso indebido y sobre todo que ponga en riesgo la información de la institución, se aplicará las sanciones que establece la misma ley.

1.1.2. Inscripción

Para continuar con el proceso de inscripción, el aspirante ingresa al siguiente link www.snna,gob.ec, el mismo que se habilita 60 días antes del inicio del periodo académico y debe ingresar su información personal y académica que le será requerida. Toda la información ingresada se convierte en el expediente del aspirante, la misma que es utilizada por la SENESCYT para gestionar diversos programas, finalmente se emite el certificado de inscripción que es utilizado por el estudiante para presentarse a rendir el mismo (Ver Figura 5).

Figura 5 Requisitos para la inscripción

1.1.3. Examen Nacional de Educación Superior (ENES)

El examen está dirigido a todos los aspirantes a la educación superior, los mismos que ya pueden contar con su título de bachiller o deben estar cursando el último año de bachillerato en el Sistema Nacional de Educación, este examen fue creado con el objetivo de garantizar la igualdad de oportunidades.

Este examen explora habilidades del pensamiento, en relación al razonamiento abstracto, verbal y numérico. "El Sistema Nacional de Nivelación y Admisión (SNNA) de la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación, en articulación con Ineval, se encarga del diseño, y definición de características técnicas. Los resultados de la calificación podrán ser conocidos a través de la cuenta personal del aspirante en la página web del SNNA (2016)"

Posteriormente en conjunto con el expediente del aspirante, se genera una hoja de respuestas personalizada, con las 120 preguntas, cada una cuenta con 4 iniciativas de respuesta. Además a la par se crean varios cuadernillos de preguntas, el nombre de estos es "formas", cada forma contiene 12 variedades con el fin de que no existan casualidades entre los aspirantes de un mismo paralelo.

Para que el aspirante se presente al examen debe contar con el certificado de inscripción y la cédula de ciudadanía, es muy importante que este en el lugar y hora indicada. Una vez finalizado el examen el tutor del paralelo retira los cuadernillos y

hojas de respuestas, estas son trasportadas con un custodio militar hasta las instalaciones del CTT-ESPE en la ciudad de Sangolquí, en este lugar, se inicia con la calificación automatizada de los exámenes, los resultados obtenidos son subidos en el portal, con el fin de que cada aspirante pueda revisar la nota obtenida en su cuenta personal del SNNA.

En procesos anteriores de postulación a universidades públicas existía un valor mínimo y máximo fijo para poder aplicar a un cupo luego del examen ENES. Actualmente ya no hay mínimos ni máximos. Los cupos se van otorgando de acuerdo con los mejores puntajes de los aspirantes. Entre más alto sea el puntaje del aspirante más posibilidades tiene de obtener un cupo

Este proceso tiene como datos de entrada el expediente del aspirante y entrega como resultado las notas del examen.

1.1.4. Postulación

Los aspirantes que obtuvieron una calificación igual o mayor a 601 /1000 en el ENES, podrán iniciar con el proceso de postulación a la carrera universitaria de su interés. Sin embargo es importante el mencionar que para la asignación de cupos la SENESCYT tomara en consideración el puntaje obtenido en el ENES, además la postulación realizada y la cantidad de cupos ofertado en las Instituciones de Educación Superior (Ver Figura 6)

La postulación consiste en la selección de cinco opciones de carrera de acuerdo a la oferta académica, las opciones numeradas del 1 al 5 se presentan en orden de prioridad, siendo la prioridad 1 la de más alto interés para el postulante, y la 5 la de menos interés.

El dato de entrada en el proceso de postulación es la nota obtenida en el ENES, y como dato de salida el registro de solicitud de los postulantes.

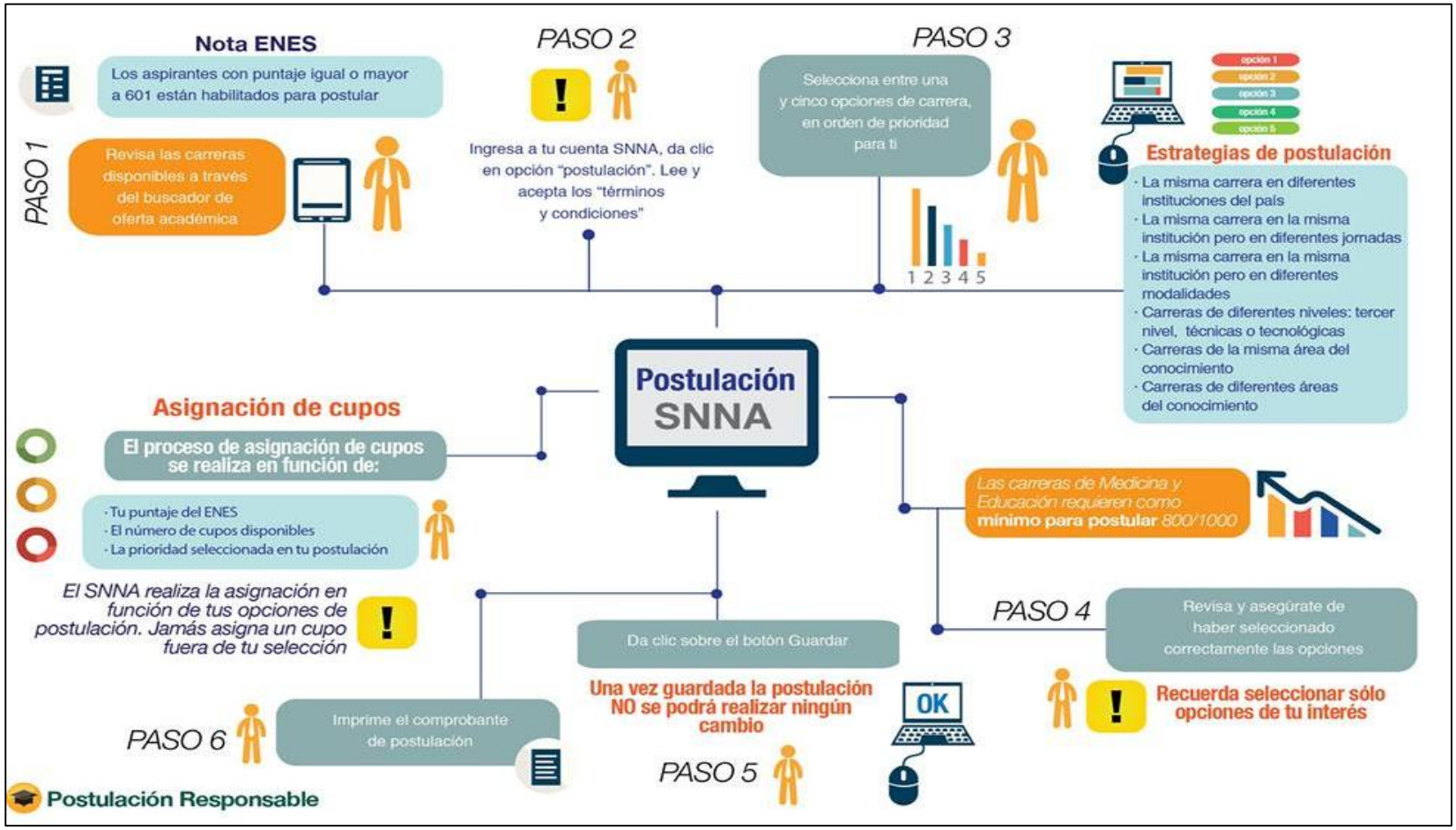


Figura 6 Proceso de postulación

1.1.5. Asignación de cupos

La Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) realiza la asignación de cupos para el ingreso a las Instituciones de Educación Superior (IES) públicas del país mediante el Sistema Nacional de Nivelación y Admisión (SNNA). Los cupos para las carreras se asignan de forma descendente, es decir, desde el puntaje más alto hacia el puntaje más bajo tomando en cuenta los siguientes parámetros:

- Puntaje obtenido en el Examen Nacional para la Educación Superior (ENES)
- Cupos ofertados por las Instituciones de Educación Superior (IES)
- Orden de prioridad de las carreras seleccionadas libremente por cada postulante

Como se mencionó en párrafos anteriores, los estudiantes pueden postularse en cinco opciones de carrera, pero solo los estudiantes que alcanzaron los puntajes más altos, tienen mayores probabilidades de que un cupo sea asignado a la primera prioridad, si ese no es el caso, se podrá optar por la segunda opción y así sucesivamente.

El sistema de encarga de informar vía correo electrónico al postulante sobre el cupo que le ha sido asignado, el postulante deberá responder sobre si acepto o no dicho cupo, la respuesta debe ser expuesta en un plazo mostrado en el cronograma, una vez culminado el plazo, el sistema no tomará en cuenta el registros realizados fuera de las fechas expuestas en el cronograma.

El dato de entrada en el proceso de asignación de cupo es la nota obtenida en el ENES, las opciones escogidas por el postulante y de los cupos asignados por las IES de acuerdo a cada carrera ofertada. Los datos de salida son el listado de estudiantes con el respectivo cupo asignado indicando carrera e IES, listado de estudiantes que no obtuvieron cupo, listado de cupos remanentes.

1.1.6. Aceptación de cupos de Carrera

Como se mencionó anteriormente, es el postulante quien debe aceptar el cupo por medio de su cuenta individual creada en el SNNA. Su aceptación es el requisito necesario para participar en los diferentes procesos de nivelación del sistema de educación superior.

Una vez aceptado el cupo, el SNNA da inicio al proceso de pre-matrícula, este es un requisito primordial que todo aspirante debe cumplir previo a la matriculación en el curso de nivelación de la carrera, este curso se realiza directamente en las Instituciones de Educación Superior.

La pre-matricula es un proceso que radica en confirmar la aceptación del cupo obtenido, este proceso tiene como objetivo primordial, el incrementar la eficiencia del sistema, tramitando el uso máximo de los cupos que ha sido asentidos por los aspirantes. El sistema cederá un certificado para legalizar y respaldar la decisión del postulante.

Los estudiantes que hayan alcanzado una nota sobre el máximo determinado por la SENESCYT con el término de nivel de alto rendimiento, deberán aceptar el cupo y pertenecerán al GAR, ellos además tienen un beneficio muy representativo, que es el aplicar a una beca sea dentro o fuera del país.

El Programa de becas Reconocimiento al Mérito Académico 2015 tiene por objetivo conceder financiamiento total o parcial a las personas naturales en goce de los derechos de ciudadanía ecuatoriana que hayan obtenido un galardón otorgado o avalado por la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación para que puedan cubrir el costo de rubros inherentes a sus actividades de investigación, capacitación, perfeccionamiento, entrenamiento profesional, encaminadas a fortalecer sus capacidades estudiantiles, profesionales e investigativas (SENESCYT, 2016).

Es importante el señalar que, si el postulante no acepta el cupo asignado, su nota ENES se mantendrá vigente durante dos convocatorias posteriores a la realización del mismo. Sin embargo. Si postulante se presentare a rendir un nuevo ENES, su nota más alta obtenida será la que mantenga vigente.

Además, es importante el mencionar que si el postulante no realiza la pre-matrícula, puede perder la opción de primera matrícula en la nivelación de carrera, pues los cupos liberados serán otorgados a los estudiantes que se encuentren en lista de espera.

1.2. Proceso de nivelación - componentes y características

Según la SENESCYT (2013) el proceso de nivelación es una técnica amparada en la ley con el fin de identificar y erradicar las diferencias existentes entre la educación que se provee a los estudiantes del bachillerato y la que de educación superior. Para ello, se han establecido estrategias dirigidas a encontrar una coherencia entre estos dos sistemas educativos, de tal forma que la nivelación sea innecesaria en el futuro.

No obstante, el tiempo que se mantenga vigente el sistema de nivelación depende de la evaluación que al respecto realice la SENECSYT basado en lo que estipula al respecto la LOES.

Una vez que el proceso de admisión ha culminado, inicia el proceso de nivelación. Para el efecto, se consideran los datos de salida, en razón que éstos serán los de entrada del segundo proceso. En otras palabras, el proceso de admisión se basa en los datos registrados de los postulantes que obtuvieron un cupo y aquellos que no lo obtuvieron.

El proceso da como resultado una lista de datos de estudiantes que permite a las IES ingresarlos a sus sistemas para distribuirlos en las diferentes carreras ofertadas, de esta forma al concluir proceso de nivelación se ha logrado que los bachilleres sean integrados en el sistema de educación superior una vez que han obtenido un cupo y se han nivelado académicamente.

1.2.1. Inscripción

Una vez que el estudiante acepta el cupo asignado por el sistema, está automáticamente matriculado en el curso de nivelación que corresponda. No obstante, la matrícula se legaliza en la institución superior a la que ha sido asignado, para lo cual necesariamente debe presentar su copia de cédula y una copia del acta de grado legalizada. Los documentos en digital se suben al sistema de cada institución educativa superior, lo

que da paso a la entrega de la certificación de la matrícula del postulante como documento habilitante.

Inmediatamente terminado este proceso, máximo en 8 días laborables cada IES debe hacer llegar a la SENESCYT por medio del portal del SNNA el listado de estudiantes que legalizaron su matrícula y el listado de cupos disponibles.

1.2.2. Capacitación

El personal encargado de preparar y ejecutar el proceso de matriculación estudiantil, así como registrar la aprobación o reprobación de la nivelación es capacitado y administrado por las IES. Cada funcionario que trabaja en estos procesos cuenta con la asignación de un usuario y contraseña para garantizar la seguridad del sistema. La SENESCYT también asigna un usuario y contraseña a cada IES lo que le permite el manejo de datos en el sistema en un ambiente de producción (SENESCYT, 2013).

1.2.3. Matrícula

Según la SENESCYT (2013) los postulantes que han legalizado su matrícula aparecen en el sistema de nivelación en el archivo plano que es subido al sistema de Gestión Académica UXXI. En el caso de los postulantes que no se presenten a las IES a legalizar su matrícula aparecerán en el sistema como nivelación no reprobada.

Para llevar a cabo la nivelación las IES cuentan con el permiso necesario por parte de la SENESCYT para crear paralelos de acuerdo a su capacidad. El proceso de legalización de matrícula inicia con el ingreso del usuario al sistema, quien ingresa el número de cédula del postulante y se despliegan sus datos de acuerdo al cupo asignado. De esta forma el usuario del sistema selecciona el grupo al que ingresa el estudiante y el UXXI detalla las materias a las que se matricula de acuerdo a la carrera seleccionada. La legalización de la matrícula estará dada en cuanto se suban al sistema los documentos (cédula y copia del acta de grado) en digital, lo que da paso a la entrega del certificado de matrícula al estudiante.

1.1. Plataforma informática del SNNA

El SNNA cuenta con una plataforma informática compuesta por dos grandes componentes: Admisión y Nivelación, cada uno de ellos está organizado por diferentes módulos que gestionan los procesos de acceso a la Educación Superior (Ver Figura 7).

Figura 7 Módulos y componentes del Sistema Nacional de Nivelación y Admisión.

La implementación utiliza tecnología Angular 2.0 sobre Apache para la presentación, Java sobre Wildfly 10.1 para el core y ORACLE 11 g para la base de datos, la infraestructura es virtual y soporta 300.000 inscripciones aproximadamente por cada semestre.

1.1.1. Sistema de Admisión

La SENESCYT (2013) para integrar estas aplicaciones realizó un convenio con el Centro de Transferencia Tecnológica de la Escuela Politécnica del Ejército (ESPE), a través del cual se implementó un sistema informático que da soporte al proceso de admisión con todos los subprocesos señalados, con el objetivo de que se cumplan los criterios esenciales de información, esto es: "disponibilidad, integridad, confidencialidad, confiabilidad, cumplimiento, eficiencia y efectividad" (FIGUEROA, 2007). Este sistema está a cargo de los procesos relacionados con el registro y calificación de postulantes, así como de los docentes para nivelación.

1.1.2. Sistema de Gestión Académica de Nivelación

Este sistema es el resultado de la unión de esfuerzos del Ministerio Coordinador de Conocimiento y Talento Humano, entidad a la que está suscrita la SENESCYT y la Oficina de Cooperación Universitaria de España, quien facilitó la implementación del Sistema de Gestión Académica Universitas XXI, que cuenta con una aplicación a través de la web (Ver Figura 8) que permite el acceso de varios usuarios (ESCUELA POLITÉCNICA DEL CHIMBORAZO, 2013).

Este sistema tiene a cargo el proceso de nivelación académica general y de carrera. Para lo cual recibe el archivo plano que entrega el sistema de admisión y entrega como resultado la lista de postulantes que aprobaron la nivelación y que pueden acceder al cupo obtenido e ingresar a primer nivel de carrera.

Figura 8 Programas de nivelación

1.2. HIPÓTESIS

El proceso de admisión a la educación superior podría mejorar mediante la aplicación de *data mining* para explorar su base de datos.

CAPÍTULO II

2. ANÁLISIS DE DATA MINING

2.1. INTRODUCCIÓN A LA MINERÍA DE DATOS

Según VIERA Y ORTIZ (2009) la minería de datos forma parte de un proceso mayor llamado en inglés *Knowledge Discovery in Database* (KDD), en español, Descubrimiento de Conocimientos en Base de Datos. El *data mining* (DM) se limita de forma rigurosa a "la obtención de modelos, restando las etapas anteriores y el propio DM como instancias del KDD" (pág. 15).

Para entender de mejor manera su funcionamiento se detalla el procedimiento (Ver Figura 9).

Figura 9. Esquema para la generación de conocimiento en bases de datos KDD
Fuente: (VIEIRA & ORTIZ, 2009, pág. 15)

Los sistemas de minería de datos que existen en el mercado tienen la función de colectar los datos, depurarlos y analizarlos, "resultando en un 'modelo descriptivo', y, caso se desee, los resultados serán utilizados para la construcción de un 'modelo predictivo" (VIEIRA & ORTIZ, 2009, pág. 15).

La minería de datos permite detectar patrones relevantes en los datos. La minería de datos "es un complemento natural al proceso de explorar y entender los datos a través de BI tradicional. Los algoritmos automáticos pueden procesar cantidades de datos muy grandes y detectar patrones y tendencias que, de lo contrario, estarían ocultos" (MICROSOFT, 2016).

Para llevar a cabo la minería de datos, es preciso recopilar los datos adecuados que respondan a un cuestionamiento específico, en este caso puede ser ¿Qué características tienen los estudiantes que acceden a las universidades en el Ecuador en función del puntaje con el que aprueban el ENES? y a paso seguido "aplicar un algoritmo para encontrar las correlaciones estadísticas en los datos. Los patrones y las tendencias que se detectan en el análisis se almacenan en forma de modelo de minería de datos" (MICROSOFT, 2016).

Cuando se ha aplicado el modelo de minería de datos a nuevos datos se pueden:

- Utilizar las tendencias anteriores con el fin de predecir los resultados del siguiente periodo de postulación.
- Considerar qué características tienen los estudiantes que alcanzan un cupo en las universidades con el fin de utilizar la información para hacer las debidas recomendaciones.
- "Buscar correlaciones entre procesos anteriores con el fin de predecir los errores o el tiempo de inactividad del servidor" (MICROSOFT, 2016).

El analista, deberá escoger la metodología de análisis que mejor convenga a sus intereses, para lo cual es fundamental que establezca los objetivos que desea alcanzar. Por ello, es necesario que antes de iniciar se conozca a qué pregunta se desea dar respuesta.

2.2. EXPLORACIÓN Y SELECCIÓN DE DATOS

Según PINTADO Y SÁNCHEZ (2014) una vez que se han definido los objetivos que se buscan alcanzar, que en este caso de estudio están dados por la identificación de las características de los estudiantes que logran obtener un cupo en las universidades del país, por las calificaciones que sacan en el ENES, lo que incluye la definición de los aspectos socio-demográficos y económicos que los caracterizan.

La segunda fase del proceso de la minería de datos es la selección de datos, en la cual es preciso tomar en cuenta que exista por lo "menos un atributo de 'resultado' que se pueda usar para aprendizaje y predicción" (MICROSOFT, 2016). Por ejemplo, la calificación del aspirante en el examen de admisión ENES relacionado con el tipo de vivienda en el que vive, el nivel educativo de los padres, el tipo de establecimiento en el que cursó los estudios de bachillerato, entre otros.

La exploración y transformación de datos es la tercera y cuarta fase del proceso. La exploración requiere dedicación para generar los perfiles de los datos que se van analizar de acuerdo al modelado. En este punto es necesario "ver la distribución de valores e identificar posibles problemas como valores o marcadores de posición ausentes. (…) considere la posibilidad de muestrear o reducir los datos" (MICROSOFT, 2016).

Para el efecto se requiere saber la forma de distribución de los datos, la relación que existe entre las columnas o tablas, si falta algún valor o requieren ser procesados o convertidos, si los datos son sobre todo un texto o principalmente sobre números o una combinación de los dos, si existen los suficientes datos para que sean analizados los resultados de destino (PINTADO & SÁNCHEZ, 2014).

Una vez completado el modelo, es preciso que se revisen detenidamente los resultados y se identifiquen las formas de modificar los datos o de obtener mejores resultados, en vista de que es extremadamente excepcional que el primer modelo provea todas las respuestas (PINTADO & SÁNCHEZ, 2014).

A medida que se ejecuta cada asistente o herramienta, el algoritmo analiza el contenido de los datos y determina si existe un patrón estadísticamente válido. Si el algoritmo no puede encontrar patrones válidos, obtendrá un mensaje de error. Sin embargo, aunque un modelo se creara correctamente, es aconsejable probarlo para ver si valida sus suposiciones (MICROSOFT, 2016).

La evaluación de los resultados del primer modelo parte de las respuestas que se obtengan de acuerdo a los objetivos planteados. Así se deberá saber los tipos de patrones que han sido encontrados, las probabilidades que existen y los valores de coincidencia, si las suposiciones fueron correctas de acuerdo a las tendencias encontradas o si existieron correlaciones que no se esperaban, si existió la suficiente recopilación de datos, entre otros (CASTAÑEDA & RODRÍGUEZ, 2003).

Una vez que se establece que el modelo es válido, éste puede ser usado para realizar las predicciones y recomendaciones que se requieren, de acuerdo al caso.

2.3. MÉTODOS DE MINERÍA DE DATOS

Según PÉREZ (2007) la minería de datos al tener la capacidad de combinar un amplio número de datos que se encuentran almacenados en uno o varios repositorios, es considerado un proceso de descubrimiento de nuevas relaciones existentes entre patrones y tendencias. Para lo cual es indispensable el uso de tecnologías de reconocimiento, así como de técnicas de estadística y matemática.

PÉREZ (2007) señala que dentro de la clasificación primaria de los métodos existentes de minería de datos se encuentran: los descriptivos o no supervisados y las predictivos o supervisados.

Los **métodos descriptivos o no supervisados** generalmente se usan cuando se trata de patrones de datos que no se conocen, con el objetivo de clasificar las variables en estudio previamente a que se apliquen otras teorías. Este tipo de método facilita la formación de grupos de datos de forma rápida.

Las observaciones son generalmente clasificadas en grupos que no son conocidos con anterioridad, los elementos de las variables pueden estar conectadas entre sí de acuerdo a vínculos desconocidos de antemano, de esta manera, todas las variables disponibles son tratadas en el mismo nivel y no hay hipótesis de causalidad (UNIVERSIDAD DE TLAXCALA, 2014).

Clasificación. Es una técnica que facilita la identificación de propiedades comunes en un conjunto de datos y a la vez agruparlos en diferentes tipos. Esta técnica permite describir cada tipo usando para el efecto las características disponibles en los datos. Después, las descripciones realizadas se usan para clasificar nuevos datos (PÉREZ C. , 2007).

Predicción. Esta técnica permite "predecir los valores de una variable continua a partir del cambio o evolución de otra variable continua que generalmente puede ser el tiempo" (PÉREZ C. , 2007, pág. 45), en este caso, por ejemplo, se puede predecir el número de estudiantes que alcanzaran puntajes de alto rendimiento a partir de los resultados de las pruebas anteriores.

Los **métodos predictivos o supervisados** se basan en "entrenar a un modelo o método por medio de diferentes datos para poder predecir una variable partiendo de estos mismos datos" (UNIVERSIDAD DE TLAXCALA, 2014). El método aprende de los datos anteriores y es capaz de emitir una respuesta.

El objetivo de este tipo de métodos está en describir una o más variables que se encuentren en relación con todas las otras. Por ello, utilizan la búsqueda de normas de predicción basada en los datos existentes, esto permite clasificar o "predecir un resultado futuro de una o más variables de respuesta o de destino en relación a lo que ocurre en la práctica con los motivos que la causan o bien en relación con las variables de entrada" (UNIVERSIDAD DE TLAXCALA, 2014). Las redes neuronales, modelos estadísticos clásicos, de regresión lineal y logística son considerados los principales métodos predictivos.

Clustering o segmentación/agrupamiento. Es una técnica que facilita la agrupación de datos similares entre sí y desiguales en relación con otros grupos (Ver Figura 10) . Es decir que esta técnica permite dividir los datos en grupos similares y a la vez facilita la agrupación organizada de un conjunto de datos (CAMPOS, 2012).

En este estudio este tipo de técnica puede aplicarse a la segmentación de estudiantes que obtienen un cupo en la universidad de acuerdo a la región a la que pertenece la institución educativa en la que se graduaron.

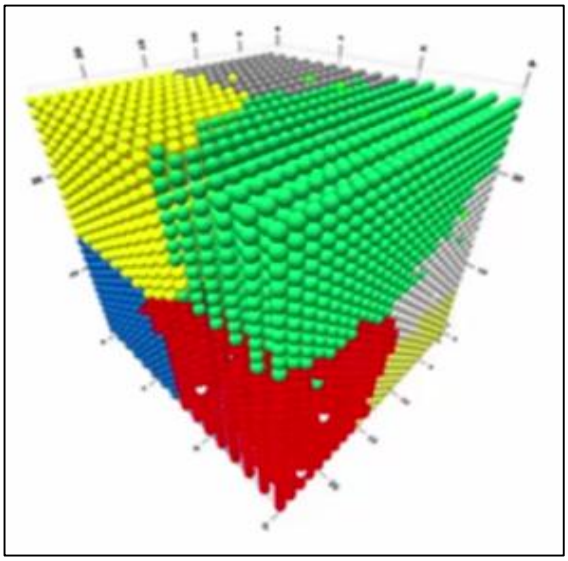


Figura 10. Clustering
Fuente: (CAMPOS, 2012)

Asociación. Es la técnica que detecta de forma automática las reglas que relacionan dos o más características, para lo cual observa si la frecuencia de aparición de los valores definidos para los atributos previamente escogidos es comparativamente elevada. Este tipo de modelo se usa fundamentalmente para establecer recomendaciones (PÉREZ C. , 2007).

2.3.1. Principales algoritmos relacionados con los métodos de aprendizaje automático

PÉREZ (2007) señala que los métodos descriptivos por asociación utilizan algoritmos A priori, y por clustering, k-medias. En el caso del método predictivo por clasificación utiliza árboles de decisión, Naive Bayes y Redes Neuronales, y en el caso de la predicción la regresión lineal.

Algoritmo A-priori. Es usa en la minería de datos para generar reglas de asociación en un conjunto de datos. Se trata de un algoritmo que se sustenta en el

conocimiento a priori o previo de los conjuntos de datos frecuentes, con el fin de minimizar el espacio de búsqueda e incrementar la eficacia (PÉREZ C. , 2007).

Es un algoritmo que, según MARTÍN, citada por la UNIVERSITAT JAUME I (2013) "permite la búsqueda automática, mediante el entrenamiento de los datos, de reglas que relacionan conjuntos de pares atributo-valor entre sí. Este tipo de técnicas estadísticas se emplea para establecer las posibles correlaciones entre distintos sucesos aparentemente independientes" (pág. 66).

En este sentido, este algoritmo reconoce como la ocurrencia de una acción influye en la aparición de otras. Por ello, las asociaciones que se identifican bien pueden utilizarse para predecir comportamientos.

Algoritmo k-medias. Para el uso de este algoritmo es preciso que se especifique el número de clusters que se desean obtener (k). Es un método de agrupación, su función es realizar la partición de un conjunto en *n* número de observaciones en k grupos en el que cada observación corresponde al grupo más próximo a la media (PÉREZ & SANTÍN, 2008), (Ver Figura 11).

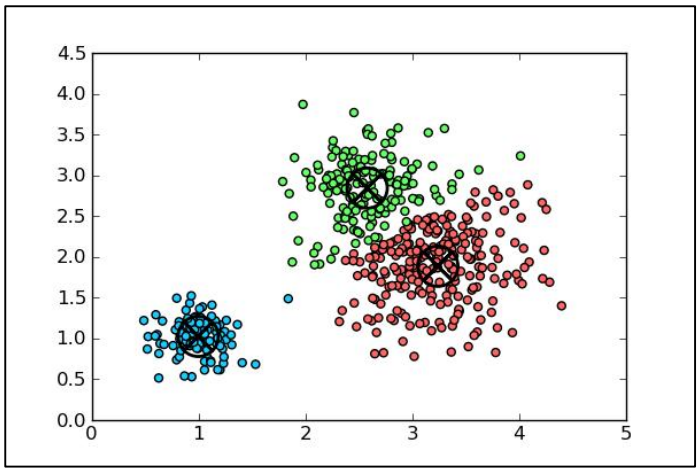


Figura 11. k-media. Clustering
Fuente: (Mpacula, 2011)

Árboles de decisión. Es una herramienta que permite realizar una clasificación adecuada, los resultados que emite son fácilmente comprensibles por los usuarios. Se trata de una técnica que permite:

- Segmentación: establecer que grupos son importantes para clasificar un cierto ítem.
- Clasificación: asignar ítems a uno de los grupos en que está particionada una población.
- Predicción: establecer reglas para hacer predicciones de ciertos eventos.
- Reducción de la dimensión de los datos: Identificar qué datos son los importantes para hacer modelos de un fenómeno.
- Identificación-interrelación: identificar que variables y relaciones son importantes para ciertos grupos identificados a partir de analizar los datos.
- Recodificación: discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante (BOUZA, 2012, pág. 65).

Este tipo de algoritmo combina técnicas matemáticas y computacionales que sirven para describir la clasificación y la generalidad de un conjunto de datos tal como se muestra a continuación (Ver Figura 12).

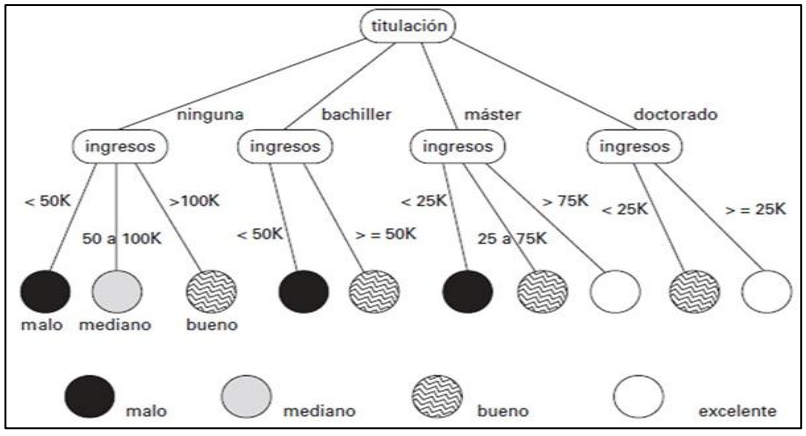


Figura 12. Ejemplo de árbol de decisión
Fuente: (WIKISPACES, 2012)

Naive Bayes. Se trata de una técnica que clasifica y predice al construir modelos que pronostican la posibilidad de posibles resultados. Este algoritmo usa datos históricos con el fin de hallar asociaciones y relaciones, con el fin de hacer predicciones, por ello es considerado de clasificación probalístico, detecta las relaciones entre las columnas de entrada y las columnas de predicción (Ver Figura 13). Puede utilizar este algoritmo para realizar la exploración inicial de los datos y, más adelante, aplicar los resultados para crear modelos de minería de datos adicionales con otros algoritmos más complejos y precisos desde el punto de vista computacional (MICROSOFT, 2016).

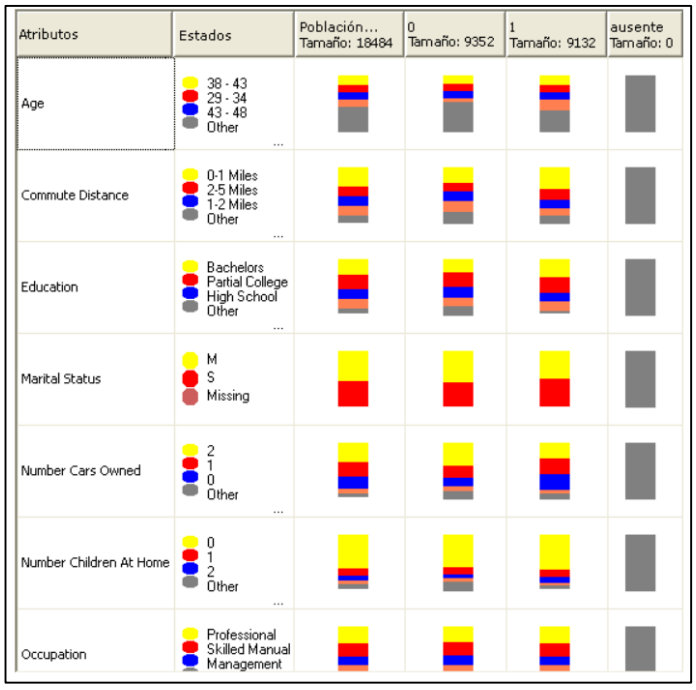


Figura 13. Naive Bayes de Microsoft
Fuente: (MICROSOFT, 2016)

Redes Neuronales. "Son modelos no lineales, inspirados en el funcionamiento del cerebro, que fueron diseñados para resolver una gran variedad de problemas" (GRUPO DE METEREOLOGÍA SANTANDER, 2015). Las conexiones que existen entre las neuronas facilitan el aprendizaje de patrones y la interrelación de datos (Ver Figura 14).

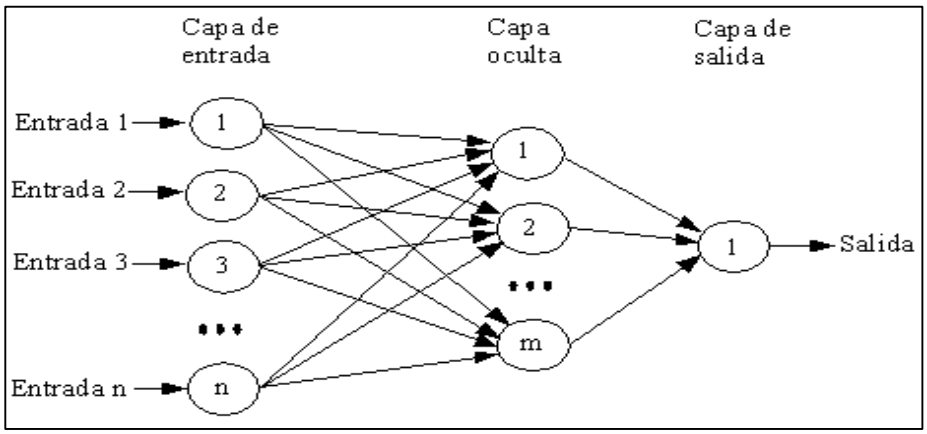


Figura 14. Redes neuronales
Fuente: (GRUPO DE METEREOLOGÍA SANTANDER, 2015)

Regresión lineal. **Este algoritmo se utiliza para predecir una variable cuantitativa a partir cuantitativa a partir de predictores que también son cuantitativos, sin embargo, de que las que las variables cualitativas también puedan incluirse generando variables falsas en las en las bases de datos. Este método es el más usado para formar relaciones entre datos, se datos, se caracteriza por ser rápido y efectivo, no obstante, resulta insuficiente en espacios espacios multidimensionales en donde se requiera relacionar más de dos variables (PÉREZ (PÉREZ C. , 2007), (Ver**

Figura 15).

Se trata de un método que "ayuda a calcular una relación lineal entre una variable independiente y otra dependiente y, a continuación, utilizar esa relación para la predicción" (MICROSOFT, 2016).

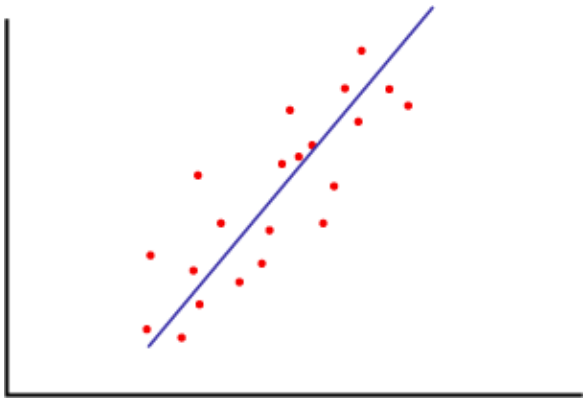


Figura 15. Regresión lineal
Fuente: (MICROSOFT, 2016)

2.4. RECONOCIMIENTO DEL DOMINIO Y DE LOS USUARIOS

Es el proceso que permite obtener la documentación y organizarla esquemáticamente, con el fin de definir las prioridades de análisis. Para el efecto es indispensable que se tenga claro el objetivo que se desea alcanzar y conocer las variables que intervienen en el análisis que se realizará.

Por ello, es importante que el analista reconozca y explore los datos para determinar aquellos que resulten de interés y respondan a la pregunta que se ha planteado. Para lograrlo, es necesario que se exploren los datos a través de la visualización, lo que permitirá encontrar patrones o tendencias que faciliten la comprensión de los datos registrados (BOUZA, 2012).

Según GONZÁLEZ (2012), los usuarios pueden ser negocios, consumidores e investigadores. En el primer caso la minería de datos permite construir modelos a partir de extensas bases de datos que contienen información transaccional para identificar patrones y predecir comportamientos en los consumidores. En el segundo, para filtrar información de grandes bases de datos, como el caso de la Web y en el tercero para analizar grandes bases de datos.

2.5. RECONOCIMIENTO DE PATRONES DE COMPORTAMIENTO

ESPINOZA Y GUTIÉRREZ (2010) indican que el reconocimiento de patrones de comportamiento se relaciona con "la clasificación de objetos y fenómenos y con la determinación de los factores que inciden en los mismos" (pág. 20). Es decir que se trata de la selección y clasificación de características que intervienen en las variables de estudio.

El reconocimiento de patrones desde la perspectiva estadística consiste en representar cada patrón mediante un vector de números resultantes del muestreo y cuantificación de las señales externas, y cada clase por uno o varios patrones prototipo. Un patrón no es más que un punto del espacio de representación de los patrones, que es un espacio dimensional por el número de variables consideradas BISQUERRA, citado por (ESPINOZA & GUTIÉRREZ, 2010, pág. 48).

En este sentido el reconocimiento estadístico de patrones está dado por las siguientes características: se sustenta en descripciones de variables numéricas, a éstas variables se les presupone características, como, por ejemplo, estar determinadas sobre un espacio métrico, se usa la probabilidad. En el proceso de reconocimiento de patrones mediante la modelación matemática está constituido por siete fases (Ver Figura 16):

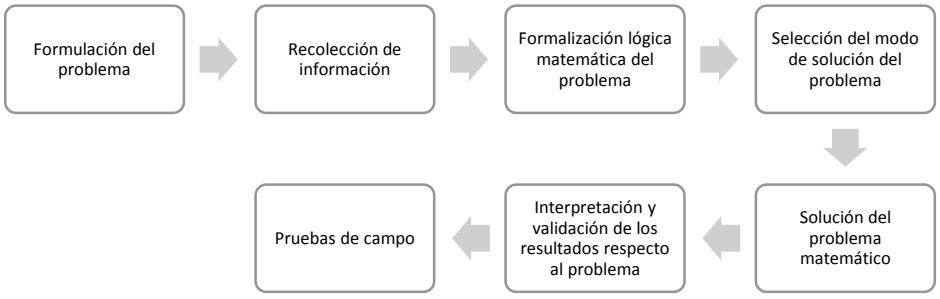


Figura 16. Fases del proceso de reconocimiento de patrones
Fuente: (ESPINOZA & GUTIÉRREZ, 2010)

En la formulación del problema el analista debe establecer el objetivo del análisis, los objetos que se analizaran, las propiedades de éstos, las características de las propiedades, las relaciones entre los objetos y sus propiedades, las hipótesis, las fuentes de información, la información que resulta más importante, la forma de recolección de la información, la forma de interpretación y manipulación de ésta, la manera en que se desea se presenten los resultados, la identificación de interferencias en la información, la valoración de errores de la información entrante, su procesamiento y salida. Para todo lo que implica esta fase es preciso contar con el apoyo informático estadístico necesario (ESPINOZA & GUTIÉRREZ, 2010).

La formalización del problema implica que el analista mentalmente formule el problema sobre el que va a trabajar. Para ello, debe seleccionar el espacio de representación de los objetos sujetos de investigación; determinar las funciones que modelarán los criterios de comparación de valores de cada variable, así como de las descripciones de los objetos; analizar los requisitos de solución se establecen a los resultados; la forma de interpretación. Esto permitirá escoger los algoritmos correctos para el análisis (ESPINOZA & GUTIÉRREZ, 2010).

- **La recolección de datos** está dada por la forma en que el especialista obtiene la información.
- **La selección del modo de solución** del problema está dada por la búsqueda de las técnicas de la minería de datos que se van a utilizar para el efecto.
- **La solución del problema** expresado en términos matemáticos involucra determinar la pertinencia del uso del algoritmo correcto de acuerdo al tipo de datos con los que se cuenta.
- **El análisis e interpretación** de los resultados precisamente busca encontrar la pertinencia de la correlación de las variables que se pretenden analizar.
- **Las pruebas de campo** permiten validar los resultados para conocer si éstos resultan lógicos y van acorde con la realidad.

- En definitiva, la metodología que se utilice para el efecto permitirá al especialista dirigir su trabajo de forma ordenada, para que encuentre la solución definitiva al problema que se ha planteado.

2.6. METODOLOGÍAS UTILIZADAS EN PROYECTOS DE MINERÍA DE DATOS

Según BELTRÁN (2012) entre las metodologías que más se utilizan para proyectos de minería de datos se encuentran: Cross Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Databases (KDD), Sample, Explorer, Modify, Model, Assess (SEMMA).

2.6.1. Cross Industry Standard Process for Data Mining (CRISP-DM)

Esta metodología nació en 1996 y en la actualidad es una de las más utilizadas en proyectos de minería de datos (BELTRÁN, 2012). El proceso que considera está compuesto por 6 etapas (Ver Figura 17):

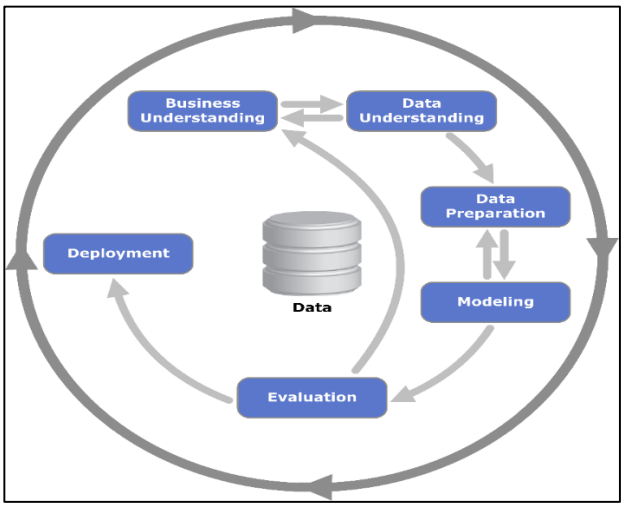


Figura 17. Fases de la metodología Cross Industry Standard Process for Data Mining
Fuente: (COMPUTER SCIENCE, 2013)

- **Comprensión del negocio,** en la cual se debe entender claramente el negocio para proponer los objetivos que se desean alcanzar de la minería de datos.

- **Comprensión de los datos**, para lo cual es preciso formularse hipótesis en base a los objetivos del negocio relacionados con la información oculta en los datos.
- **Preparación de los datos**, lo que requiere que éstos sean seleccionados, limpiados y transformados.
- **Modelado,** en esta etapa se selecciona la técnica del modelado y se calibra los parámetros.
- **Evaluación,** en este punto se evalúa el modelo que va a ser usado para constatar que permite cumplir los objetivos propuestos en el proyecto.
- **Implantación,** correspondiente a la producción del conocimiento alcanzado del proceso de minería de datos.

2.6.2. Knowledge Discovery in Databases (KDD)

Según VIEIRA Y ORTIZ (2009) este método fue creado en 1995 con el fin de "designar el conjunto de procesos, técnicas y abordajes que propician el contexto en el cual la minería de datos tendrá lugar" (pág. 23). Se trata de una metodología que revela el conocimiento a través de la identificación de patrones válidos dentro de una extensa cantidad de datos. Este proceso se lleva a cabo en 8 etapas (Ver Figura 18)

- **Entendimiento del dominio de aplicación.** Fase en la cual se identifican los objetivos organizacionales del proceso de minería de datos.
- **Instauración del conjunto de destino de datos.** Fase de descubrimiento, para lo cual se selecciona el conjunto de datos.
- **Limpieza y pre-procesamiento.** Para lo cual se usa técnicas de tratamiento de campos de datos faltantes.
- **Disminución y protección de datos.** Se trata de acciones operativas de transformación de datos en relación a las metas propuestas en el proceso de minería de datos.

- **Definición de la tarea del proceso de minería de datos.** A través de la cual se establece el algoritmo que se usará para alcanzar las metas propuestas.
- **Data mining.** Se considera a la búsqueda de patrones válidos de datos.
- **Interpretación.** Dada por la visualización de los patrones hallados.
- **Utilización del conocimiento descubierto.** A través de la documentación y reporte hacia los interesados.

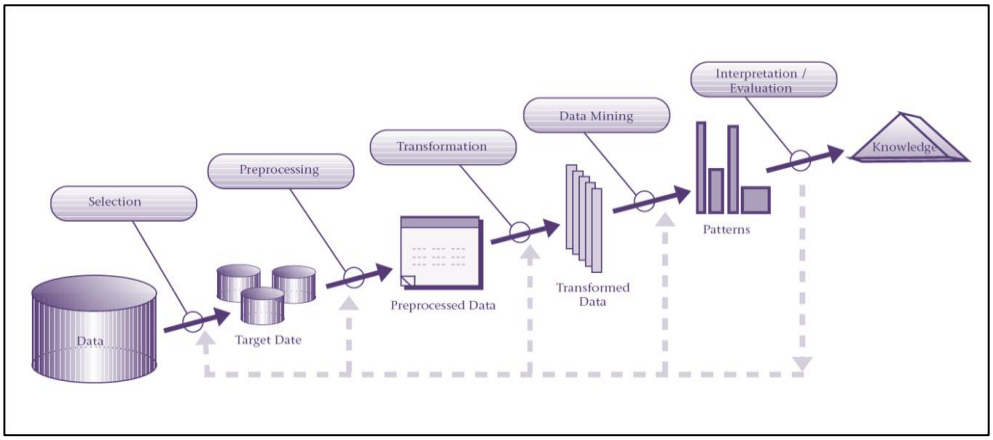


Figura 18. Fases de la metodología KKD
Fuente: (UNIVERSIDAD TECNOLÓGICA METROPOLITANA, 2013)

2.6.3. Sample, Explore, Modify, Model, Assess (SEMMA)

Según CAMARGO Y SILVA (2012) esta metodología está dada por el significado de su acrónimo SEMMA, que identifica Sample (muestra), Explore (explorar), Modify (modificar), Model (modelar) y Assess (evaluar). Es decir que se trata de un proceso que de forma lógica organiza la minería de datos, "intenta hacer fácil de aplicar la exploración estadística y la visualización de técnicas, seleccionando y transformando las variables predictivas más relevantes, modelándolas para obtener resultados, y finalmente confirmar la precisión del modelo" (pág. 15), (Ver Figura 19).

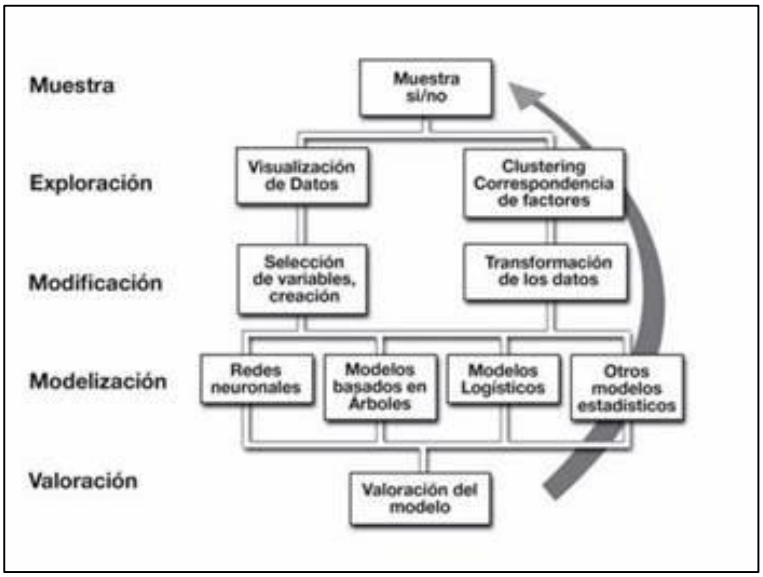


Figura 19. Fases de la metodología SEMMA
Fuente: (COMPUTER SCIENCE, 2013)

- **Muestreo.** Etapa en la que se extrae una porción significativa de datos para contener la información requerida, la que será reducida para facilitar su manipulación.
- **Explorar.** A través de buscar tendencias y anomalías imprevistas que permitan comprender totalmente los datos.
- **Modificar.** A través de la creación, clasificación y transformación de variables con el fin de puntualizar el proceso de selección del modelo.
- **Modelar.** Tarea realizada por el software, el cual busca automáticamente una combinación de datos que prediga con cierta seguridad un resultado deseado.
- **Evaluar.** Etapa en la cual se clasifican los datos de acuerdo a la evaluación de utilidad y fiabilidad.

2.7. ANÁLISIS DE LA PROBLEMÁTICA

Como se mencionó en el capítulo I, el principal partícipe en el proceso de admisión a la educación superior son los estudiantes que aspiran a obtener un cupo en alguna Institución de Educación Superior Publica, dentro del mencionado proceso se realiza un levantamiento de información del aspirante así como también del entorno en el que se desarrolla, esta información es muy importante a fin de entender y determinar cuáles son los principales patrones y tendencias que describen a los aspirantes que aprueban o no el Examen Nacional de Educación Superior, la determinación de estos perfiles de aspirante puede ayudar a encontrar factores que expliquen el resultado obtenido.

Una de las dificultades que se presenta al momento de analizar la información recolectada a través de los aplicativos informáticos, es el volumen de datos levantados, tanto en número de aspirantes como en el total de variables recogidas al momento de la inscripción, por ejemplo, el número de aspirantes que se presentaron a rendir el Examen Nacional de Educación Superior del régimen sierra para el periodo 2016 fue de 242.324, de los cuales se generaron 310 variables, las características de estas variables, revelan que es necesario identificar y utilizar herramientas de análisis más sofisticadas que permitan encontrar los patrones de comportamiento de los aspirantes dentro de los datos y expliquen de manera estadística su relación con la aprobación o no del examen.

En la actualidad una de las principales soluciones analíticas para el manejo de grandes volúmenes de información y extracción de conocimiento inmerso en los mismos es la minería de datos, la cual nos presenta un conjunto de técnicas de manipulación y transformación de datos, así como también de algoritmos de carácter estadístico matemático que en conjunto nos permite presentar u obtener una solución con una estructura fácil de entender para la toma de decisiones y estadísticamente robusta. Para el caso de este proyecto, la metodología de minería de datos a aplicar es la metodología SEMMA, los modelos estadísticos que se enmarcan en la resolución de la problemática descrita son de carácter supervisado,

donde la variable objetivo a explicar es el estado del Examen Nacional de Educación Superior (aprobación o no aprobación)

Características de los aspirantes

- Características de su educación
- Características del hogar
- Factores con oportunidad de mejora
- Expectativas sobre la educación superior
- Características de la comunidad

2.8. APLICACIÓN DE LA METODOLOGÍA SEMMA

En el capítulo II sección 2.6.3, se hizo una detalla descripción de la metodología SEMMA, donde las etapas o pasos principales del proceso de minería están conformados por los siguientes: Muestrear, Explorar, Modificar, Modelar y Evaluar.

A continuación, se analizarán cada una de las etapas mencionadas con la información levantada para el examen de admisión del régimen Sierra y el resultado del examen obtenido por los aspirantes.

2.8.1. Etapa Muestrear

La información levantada del examen de admisión del régimen Sierra consta de 242,324 registros y 310 variables. Donde la variable objetivo a explicar es el Estado del Examen, a continuación, se presenta su frecuencia en la Tabla 2:

Tabla 2
Objetivo Estado del Examen

ESTADO EXAMEN	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
APROBADO	219957	90.77	219957	90.77
NO APROBADO	22367	9.23	242324	100.00

En la Tabla 2 se muestra la distribución de la variable Estado del Examen, el total de registros son 242,324, los cuales se distribuyen en dos categorías: Aprobado y No Aprobado que representa el 90.7 % y 9.23 % respectivamente. En este caso como el objetivo es estimar un modelo supervisado, dependiente del modelo o técnica utilizada se deben equiparar o igualar las proporciones de las categorías de la variable objetivo antes de realizar el modelo, una técnica común mente utilizada en el caso de los árboles de decisión es el sobremuestro, el mismo que consiste en seleccionar aleatoriamente individuos de cada una de las categorías de la variable objetivo, hasta tener una muestra representativa y equiparable de las dos categorías. En el caso de la regresión logística la corrección de la desproporción de categorías se la realiza mediante la corrección de la constante.

A la base de información se la segmentó en tres conjuntos de datos, correspondientes a las bases de Entrenamiento (70%), Validación (15%) y Prueba (15%). A continuación, se presenta el número de registros y porcentaje para cada uno de los conjuntos de datos (Ver Tabla 3).

Tabla 3
Frecuencia base de Entrenamiento, Prueba y Validación

Conjunto de datos	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
TRAIN	169625	70.00	169625	70.00
VALIDATE	36349	15.00	205974	85.00
TEST	36350	15.00	242324	100.00

2.8.2. Etapa Explorar

La información levantada presenta 310 variables entre variable objetivo y explicativas, en esta etapa se realiza el análisis univariante y bivariante entre cada una de las variables explicativas y la variable objetivo, todo esto con el fin de seleccionar las variables que tienen una mayor importancia al momento de explicar la variable objetivo. El estadístico utilizado para identificar la importancia de cada

variable con respecto al variable objetivo es el chi-cuadrado, el mismo que responde a la prueba de hipótesis siguiente:

H0: Las variables X y Y son independientes.

H1: Las variables X y Y no son independientes.

Valores del estadístico cercano a 0 significa que las variables no son independientes, y valores altos significan que las variables son independientes. A continuación, se presenta el análisis univariante y bivariante realizado:

Análisis Univariante

El análisis univariante nos permite observar de manera rápida y sencilla la distribución de los elementos de cada una de las variables, nos permite identificar los valores más representativos, valores perdidos o missing. A continuación, se presentan las frecuencias de las variables más significativas que forman parte del modelo

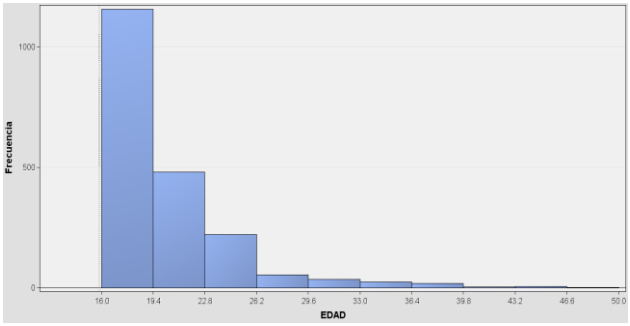


Figura 20 Frecuencia de Edad

En las frecuencias de la variable edad, la mayoría de registros se concentran en el intervalo entre 16 y 19 años, existen también aspirantes que superan la edad de 30 años, siendo esto importante dado que muestra la inclusión y participación de personas que están intentando retomar los estudios superiores (Ver Figura 20).

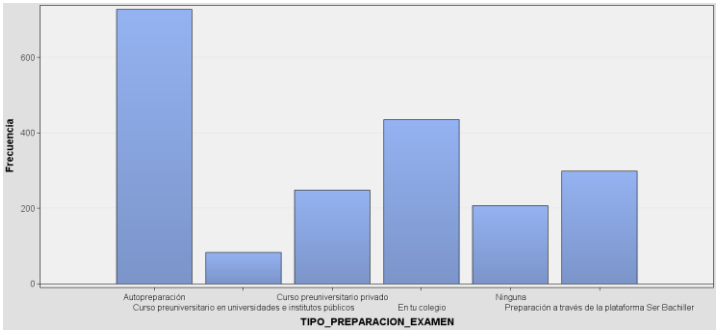


Figura 21 Frecuencia Preparación de Examen

Se observa que existe un alto número de aspirantes que se preparan por sus propios medios para rendir el examen, seguido de aquellos que tienen ayuda por parte de su colegio. Aquí se puede evidenciar un aspecto a mejorar, dado que las autoridades deberían ayudar a que la preparación del examen sea más accesible o investigar porque los aspirantes no logran tener un apoyo para su preparación (Ver Figura 21).

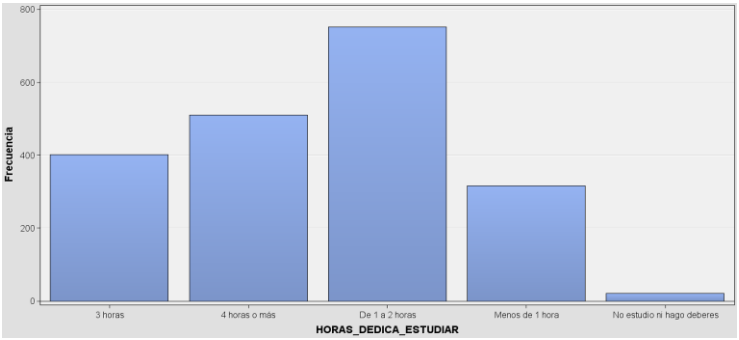


Figura 22 Frecuencia Horas dedicadas a estudiar

Los datos muestran que el número de horas que los aspirantes se dedican a estudiar diariamente, y es interesante ver que un gran número de los mismos dedican menos de 2 horas para estudiar, factor que podría afectar al resultado obtenido en el examen (Ver Figura 22).

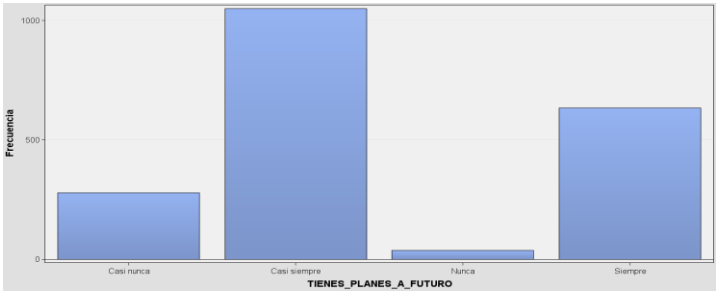


Figura 23 Frecuencia Tienes planes a futuro

Se observa la planificación de algún tipo de actividad a futuro, que tienen los aspirantes, en general solo un número muy bajo de los mismos no tiene planes hacia el futuro (Ver Figura 23).

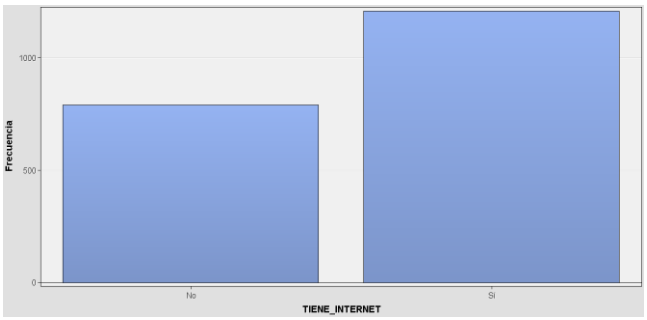


Figura 24 Frecuencias Tiene Internet

En cuanto al ámbito de acceso a internet en casa, es interesante debido a que existen muy poca diferencia entre los que tienen internet vs los que no tienen, este hecho podría ser fundamental al momento de tener acceso a información en línea o algún tipo de capacitación que le permita mejorar en el resultado del examen (Ver Figura 24).

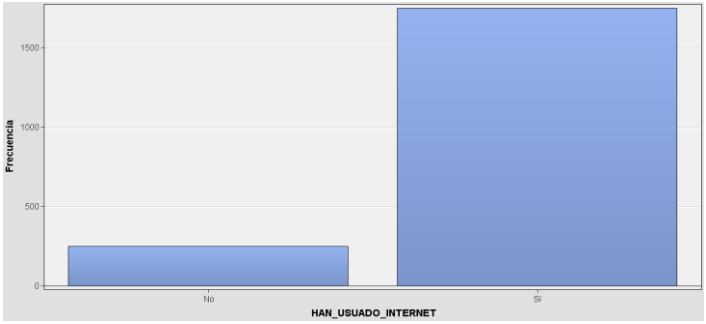


Figura 25 Frecuencia Uso Internet

Sobre el uso del internet por parte de los aspirantes, esta información se puede complementar con la de la figura 24, porque se observa que un número considerable de aspirante han usado internet, aunque no lo tenga disponibles en su hogar. Es aquí donde se podría identificar que el internet forma parte ya de las necesidades básicas de los aspirantes (Ver Figura 25).

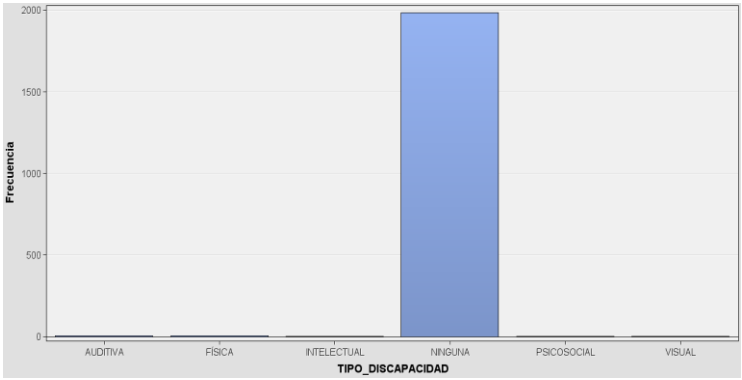


Figura 26 Frecuencia Discapacidad

En cuanto al tipo de discapacidad que tendría un aspirante que rinde el examen, se puede ver que en su mayoría no tienen ninguna discapacidad, es decir la participación de personas con algún tipo discapacidad es mínima o casi nula (Ver Figura 26).

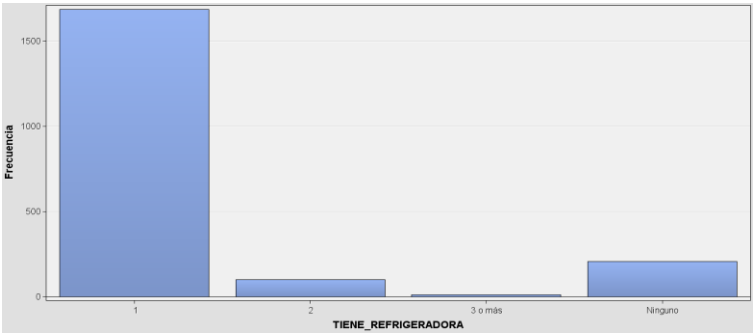


Figura 27 Tiene refrigerado Hogar

Sobre los electrodomésticos y específicamente la tenencia de una refrigeradora en el hogar, en su mayoría de los aspirantes por lo menos tienen 1 en el hogar. Por otro lado, hay una pequeña parte de la población que no la tiene, esto podría tomar

como un indicador de bajos recursos por parte del núcleo familiar del aspirante y donde se podría tratar de indagar algunos otros factores que permitan generar algún tipo de ayuda económica para el aspirante (Ver Figura 27)

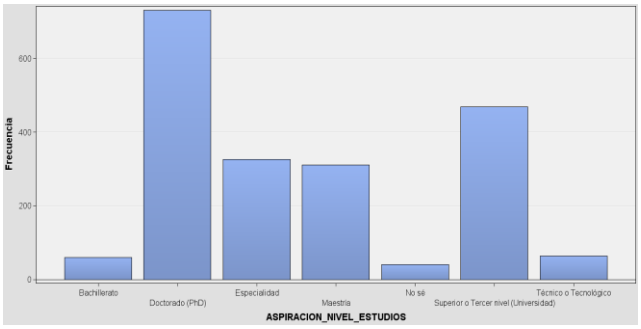


Figura 28 Frecuencia Aspiración de Estudio

Según las expectativas de continuar sus estudios superiores, se observa que en su mayoría aspiran llegar a un título de PhD y universidad, no obstante también hay un porcentaje mínimo que no tiene algún tipo de aspiración estudios (Ver Figura 28).

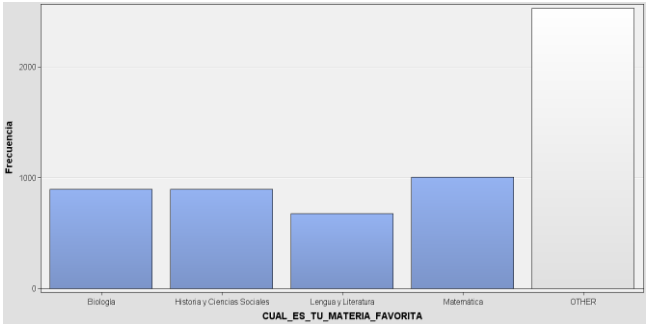


Figura 29 Frecuencia Materia favorita

Sobre las materias favoritas de los aspirantes que rinden el examen, la distribución muestra que existen varias materias que son importantes para los aspirantes entre las que se destaca las Matemáticas, Ciencias Sociales, Lenguaje y Biología, pilares fundamentales o que forman parte de las ciencias básicas iniciales de la universidad (Ver Figura 29).

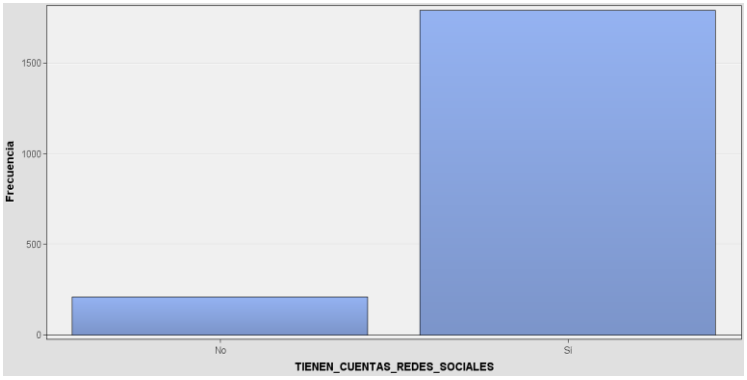


Figura 30 Frecuencia tiene redes sociales

La tenencia de redes sociales por parte del aspirante, es muy claro que en su mayoría tiene una cuenta de red social, siendo este un factor clave que se podría utilizar para mejorar la comunicación entre los aspirantes y las autoridades (Figura 30).

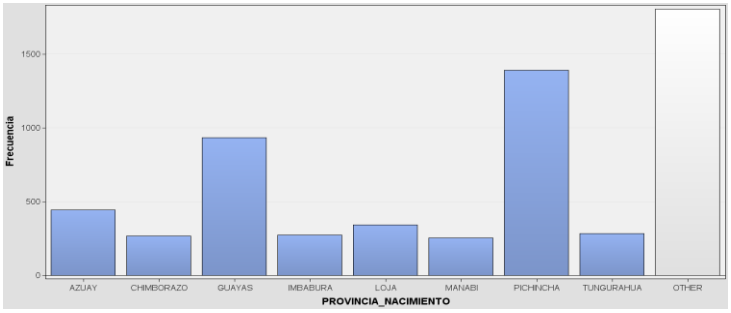


Figura 31 Frecuencia provincia nacimiento

En cuanto a la provincia de nacimiento del aspirante, se observa que la mayoría nacieron en las provincias de Pichincha, Guayas y Azuay (Figura 31).

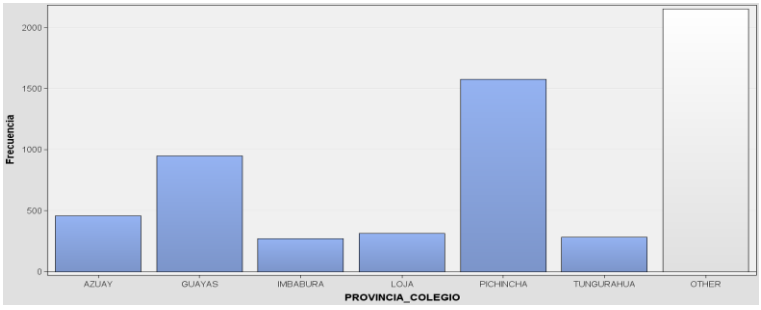


Figura 32 Frecuencia Provincia Colegio

La mayor concentración de aspirantes encuentran en las provincias de Pichincha, Guayas y Azuay, (Figura 32).

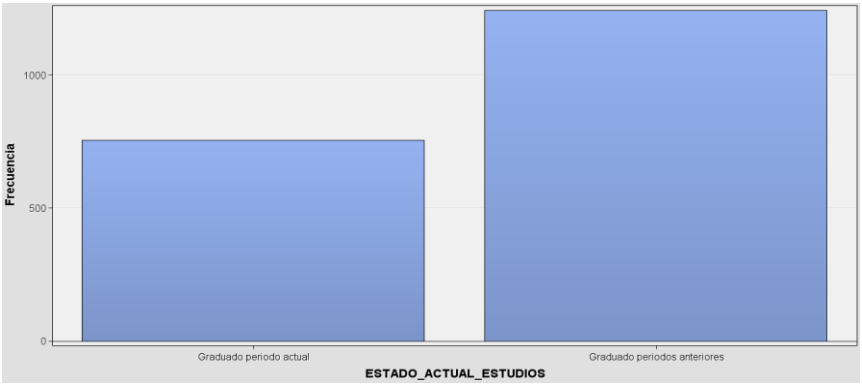


Figura 33 Frecuencia Estudio actuales

Se observa que si el aspirante es graduado en el periodo actual o en periodos anteriores, de esto se observa una mayor cantidad de aspirantes que se han graduado en periodos anteriores, esto podría ser un factor relevante al momento de rendir el examen y la aprobación del mismo (Ver Figura 33).

Obs #	Nombre de la variable	Tipo	Porcentaje de ausentes	Mínimo	Máximo	Media	Número de niveles	Porcentaje moda	Moda
1	ASPIRACION_NIVEL_ESTUDIOS	CLASS	0	.	.	.7		36.55	DOCTORADO (PHD)
2	CUAL_ES_TU_MATERIA_FAVORITA	CLASS	0	.	.	.12		15.55	MATEMÁTICA
3	ESTADO_ACTUAL_ESTUDIOS	CLASS	0	.	.	.2		62.25	GRADUADO PERIODOS ANTERIORES
4	ESTADO_EXAMEN	CLASS	0	.	.	.2		90.2	APROBADO
5	FORMA_DESECHAR_AGUAS_SERVIDAS	CLASS	0	.	.	.6		68.95	CONECTADO A RED PÚBLICA DEL ALCA
6	HAN_USUADO_INTERNET	CLASS	0	.	.	.2		87.5	SÍ
7	HORAS_DEDICA_ESTUDIAR	CLASS	0	.	.	.5		37.6	DE 1 A 2 HORAS
8	NACIONALIDAD_INDIGENA	CLASS	0	.	.	.5		94.85	NO APLICA
9	OCUPACION_ASPIRANTE	CLASS	0	.	.	.9		39.75	ESTUDIANTE O RECIÉN GRADUADO (HA
10	PROVINCIA_COLEGIO	CLASS	0	.	.	.26		27.45	PICHINCHA
11	PROVINCIA_NACIMIENTO	CLASS	0	.	.	.24		24.6	PICHINCHA
12	PROVINCIA_RESIDENCIA	CLASS	0	.	.	.25		28.3	PICHINCHA
13	RESALTAS_LAS_PARTES_IMPORTATES	CLASS	0	.	.	.4		44.4	CASI SIEMPRE
14	TIENEN_CUENTAS_REDES_SOCIALES	CLASS	0	.	.	.2		89.6	SÍ
15	TIENES_PLANES_A_FUTURO	CLASS	0	.	.	.4		52.55	CASI SIEMPRE
16	TIENE_INTERNET	CLASS	0	.	.	.2		60.45	SÍ
17	TIENE_REFRIGERADORA	CLASS	0	.	.	.4		84.31	
18	TIPO_DISCAPACIDAD	CLASS	0	.	.	.6		99.2	NINGUNA
19	TIPO_EMPLEO_PADRE	CLASS	0	.	.	.9		27.1	TIENE UN TRABAJO PAGADO ESTABLE
20	TIPO_PREPARACION_EXAMEN	CLASS	0	.	.	.6		36.4	AUTOPREPARACIÓN
21	EDAD	VAR	0	16	50	20.313.		.	

Figura 34 Estadísticos generales

El resumen de las frecuencias y estadísticos estándares utilizados para la exploración de los datos, muestra que la variable clase aspiración de nivel de estudio se tiene 7 categorías o niveles y su moda o valor más común es Doctorado. Para la variable Edad dado que es numérica presenta un valor mínimo de 16 años y un máximo de 50 años, la media de la edad en la base de información es de 20 años (Ver Figura 34).

Análisis Bivariante

El análisis bivariante nos permite identificar la relación existente entre la variable objetivo y cada una de las variables explicativas, de tal forma que ordena

las variables de mayor a menor la importancia, así elegir aquellas más significativas para el estudio.

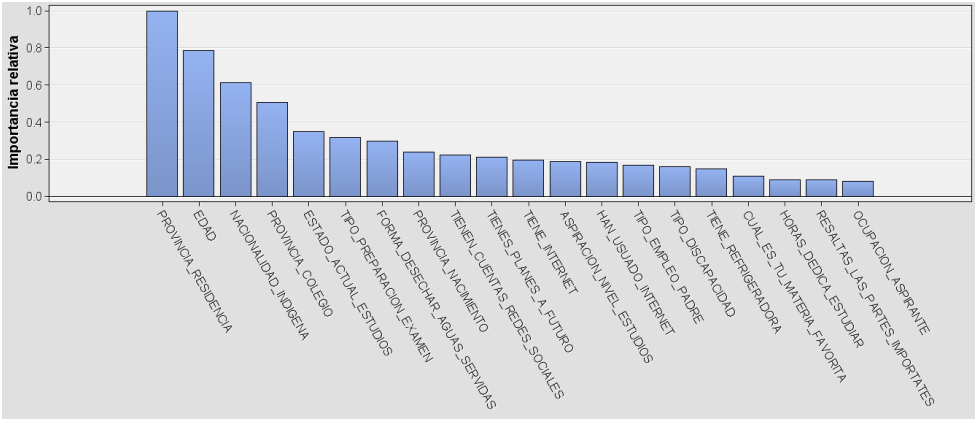


Figura 35 Importancia de variable para el modelo.

Entre las variables más influyentes y consideradas para el modelo, se encuentran la provincia de residencia con un 100% de importancia, seguida de la variable edad con un 80%, la variable ocupación del aspirante con un valor del 7% es la última presente en la elección (Ver Figura 35).

2.8.3. Etapa Modificar

Esta etapa consiste e identificar la transformación más relevante de cada una de las variables, a fin de que su transformación ayude a explicar de una mejor manera la variable objetivo. Generalmente existen transformaciones estándares dependiendo del tipo de variable que se presenta en el análisis, en el caso de la base de estudio la mayoría de variables son categóricas o nominales para la cual la transformación utilizada es la agrupación de categorías menos representativas, para el caso de la edad que es la única variable numérica se utilizó la transformación de cuartiles la cual agrupa a los registros de manera idénticamente distribuida. A continuación se muestran las variables a las cuales se realizó algún tipo de transformación (Ver Figura 36).

Name	Formula	Input Type	Input Level
TG_ASPIRACION_NIVEL_ESTUDIOS	Group:ASPIRACION_NIVEL_ESTUDIOS	C	NOMINAL
TG_CUAL_ES_TU_MATERIA_FAVORITA	Group:CUAL_ES_TU_MATERIA_FAVORITA	C	NOMINAL
PCTL_EDAD	Quantile(4)	N	INTERVAL
TG_FORMA_DESECHAR_AGUAS_SERVIDAS	Group:FORMA_DESECHAR_AGUAS_SE...	C	NOMINAL
TG_NACIONALIDAD_INDIGENA	Group:NACIONALIDAD_INDIGENA	C	NOMINAL
TG_OCUPACION_ASPIRANTE	Group:OCUPACION_ASPIRANTE	C	NOMINAL
TG_PROVINCIA_COLEGIO	Group:PROVINCIA_COLEGIO	C	NOMINAL
TG_PROVINCIA_NACIMIENTO	Group:PROVINCIA_NACIMIENTO	C	NOMINAL
TG_PROVINCIA_RESIDENCIA	Group:PROVINCIA_RESIDENCIA	C	NOMINAL
TG_TIENES_PLANES_A_FUTURO	Group:TIENES_PLANES_A_FUTURO	C	NOMINAL
TG_TIENE_REFRIGERADORA	Group:TIENE_REFRIGERADORA	C	NOMINAL
TG_TIPO_DISCAPACIDAD	Group:TIPO_DISCAPACIDAD	C	NOMINAL
TG_TIPO_EMPLEO_PADRE	Group:TIPO_EMPLEO_PADRE	C	NOMINAL
TG_TIPO_PREPARACION_EXAMEN	Group:TIPO_PREPARACION_EXAMEN	C	NOMINAL

Figura 36 Resumen de transformación de Variables

En las siguientes figuras se presentan las variables provincia de nacimiento y edad transformada, respectivamente.

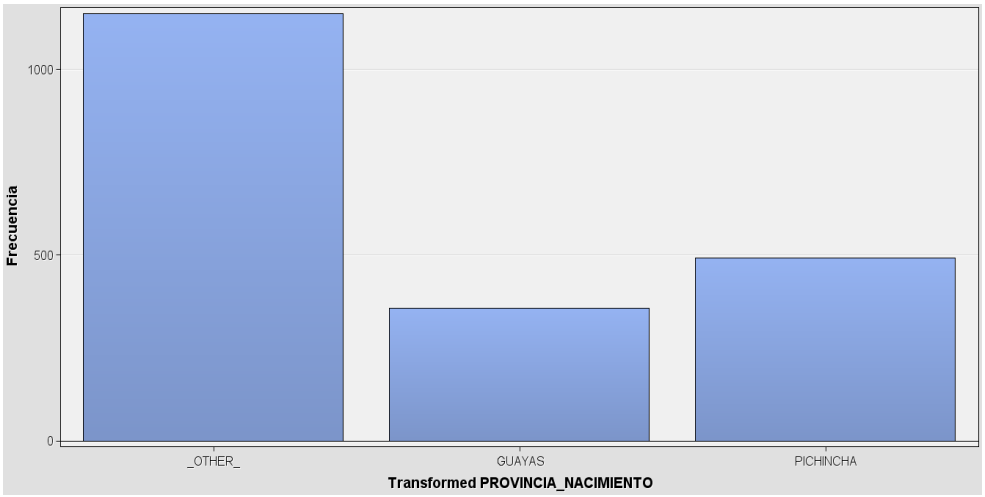


Figura 37 Provincia Nacimiento transformada

El resultado de aplicar una transformación a la provincia de nacimiento muestra las provincias más representativas como Guayas y Pichincha y agrupa al resto de provincias con menor representatividad en la categoría otros (Ver Figura 37).

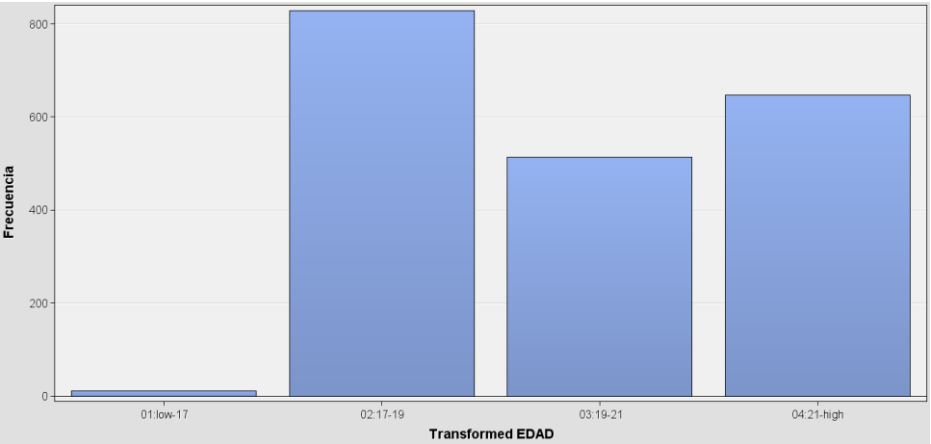


Figura 38 Edad de aspirante transformada

De igual forma, la categorización automática que se ha realizado sobre la variable edad, agrupa aquellos aspirantes que tienen edades menores a 17 años (Ver Figura 38). También se puede observar que existe una parte de la población representativa superior a los 21años que aspira a un nivel de estudios de tercer nivel.

2.8.4. Etapa Modelar

En esta etapa se aplicaron los modelos de carácter supervisado, con el fin de identificar las variables independientes que explique la variable objetivo, más explícitamente se aplicaron dos modelos:

- Árbol de decisión
- Regresión Logística

2.8.5. Resultados Árbol de decisión

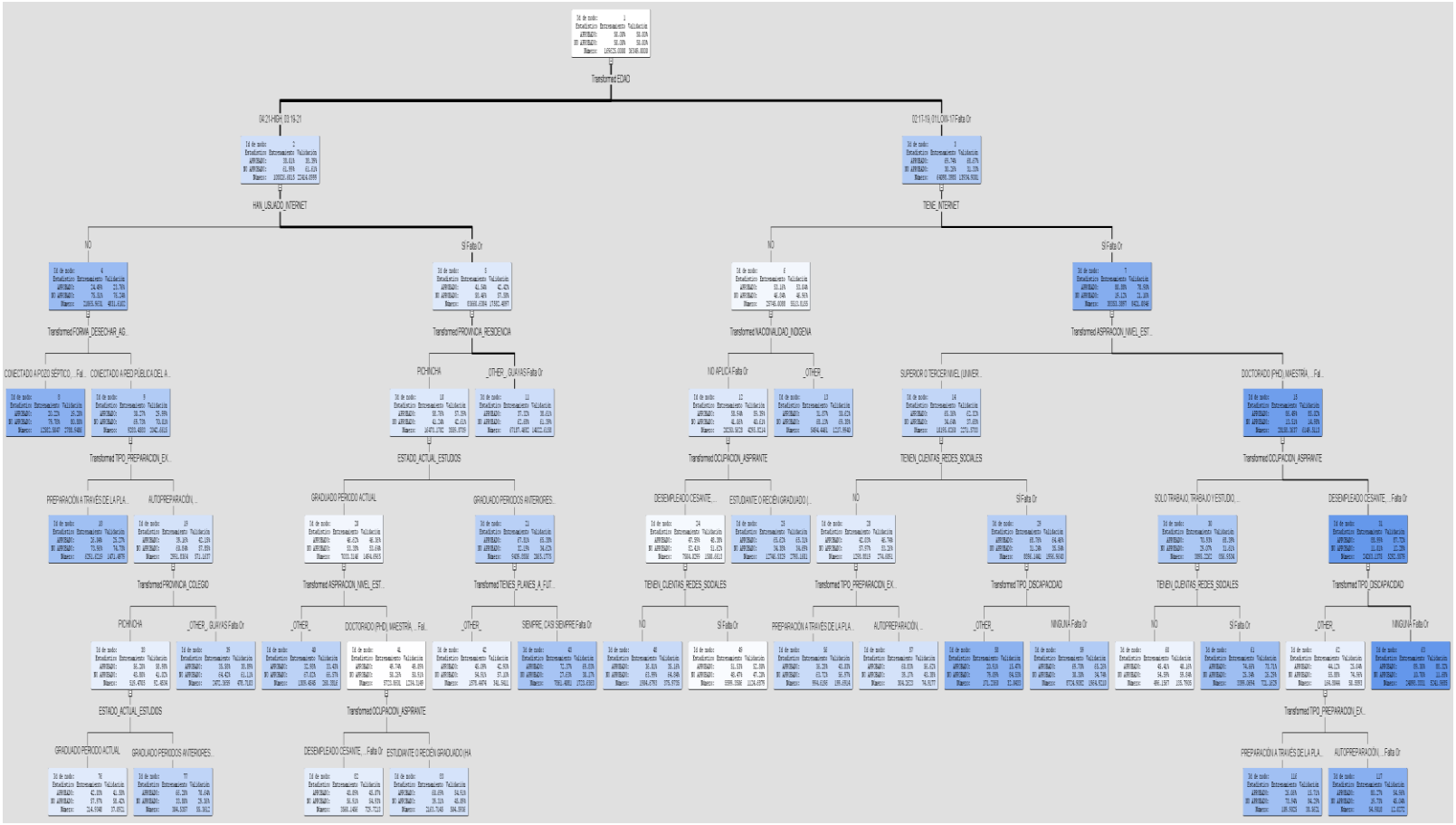


Figura 39 Árbol de decisión modelo

El árbol de decisión (Ver Figura 39), donde se identifica como variable objetivo el estado examen con sus dos categorías Aprueba y No aprueba, cada una con un porcentaje 50% en el nodo raíz. Como primera variable significativa aparece la transformación de la variable edad, la cual divide al nodo raíz en dos nodos 2 y 3. En el nodo 2 se encuentra los aspirantes con edades superiores a 19 años, en este nodo el porcentaje de la categoría no Aprueba aumenta a un 62%, es decir 12 puntos a partir del nodo raíz, esto quiere decir que los aspirantes con edades más avanzadas tienen un mayor porcentaje de no aprobar el examen. En el nodo 3 sucede todo lo contrario, los aspirantes con edades menores o iguales a 19 años presenta un porcentaje de no aprobación menor con un valor de 30%, es decir 20 puntos menos con respecto al nodo raíz.

Los nodos hijos a partir del nodo 2 y nodo 3 van identificando tanto las variables significativas para explicar el comportamiento de la variable objetivo y van delineando el perfil tanto para los aspirantes que aprueban el examen, así como también el perfil de los aspirantes que no aprueban.

Por ejemplo, el perfil para los Aspirante que aprueban el examen, son aquellos que tienen una edad menor o igual a 19 años, tienen internet en su domicilio, tiene aspiración a maestría o doctorado no trabaja y no tiene una discapacidad, este perfil llega a un porcentaje de No Aprobación del 10%. Un perfil para los aspirantes que no aprueban el examen son aquellos que tienen una edad superior a 19 años no han usado internet y no tienen acceso a agua potable llegando a una tasa de No Aprobación del 80%.

Es claro que el árbol de decisión nos permite una lectura más clara de las variables y atributos que explican la Aprobación o No del examen por parte del aspirante, a continuación se presenta la importancia que tienen las variables en el modelo del árbol de decisión, siendo la transformación de la edad la más importante con un valor 100% y las horas que dedican a estudiar la menos importante 0 %, (Ver Figura 40).

Nombre de la variable	Número de reglas de división	Importancia	Importancia de validación	Ratio de validación para la importancia de entrenamiento
PCTL_EDAD	1	1.0000	1.0000	1.0000
HAN_USUADO_INTERNET	1	0.5742	0.6581	1.1460
TG_PROVINCIA_RESIDENCIA	1	0.4540	0.4055	0.8932
TIENE_INTERNET	1	0.3872	0.3853	0.9953
TG_NACIONALIDAD_INDIGENA	1	0.3143	0.3562	1.1334
TG_FORMA_DESECHAR_AGUAS_SERVIDAS	1	0.2647	0.3082	1.1643
TG_OCUPACION_ASPIRANTE	3	0.2199	0.1996	0.9078
TG_ASPIRACION_NIVEL_ESTUDIOS	2	0.2030	0.2264	1.1153
ESTADO_ACTUAL_ESTUDIOS	2	0.1897	0.1836	0.9681
TG_TIPO_PREPARACION_EXAMEN	3	0.1879	0.2271	1.2083
TIENEN_CUENTAS_REDES_SOCIALES	3	0.1807	0.1772	0.9804
TG_TIENES_PLANES_A_FUTURO	1	0.1264	0.1374	1.0869
TG_TIPO_DISCAPACIDAD	2	0.1238	0.1639	1.3241
TG_PROVINCIA_COLEGIO	1	0.0835	0.0690	0.8260
TG_PROVINCIA_NACIMIENTO	0	0.0000	0.0000	.
RESALTAS_LAS_PARTES_IMPORTATES	0	0.0000	0.0000	.
TG_CUAL_ES_TU_MATERIA_FAVORITA	0	0.0000	0.0000	.
TG_TIPO_EMPLEO_PADRE	0	0.0000	0.0000	.
TG_TIENE_REFRIGERADORA	0	0.0000	0.0000	.
HORAS_DEDICA_ESTUDIAR	0	0.0000	0.0000	.

Figura 40 Importancia Variables Árbol de decisión

Tabla de decisión

Rol de los datos=TRAIN Variable objetivo=ESTADO_EXAMEN Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total	Porcentaje ajustado de la variable predicha/decisión
APROBADO	APROBADO	96.6901	57.5635	88630	52.2506	28.7818
APROBADO	NO APROBADO	83.8099	42.4365	65339	38.5197	21.2182
NO APROBADO	APROBADO	3.3099	19.3792	3034	1.7887	9.6896
NO APROBADO	NO APROBADO	16.1901	80.6208	12622	7.4411	40.3104

Rol de los datos=VALIDATE Variable objetivo=ESTADO_EXAMEN Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total	Porcentaje ajustado de la variable predicha/decisión
APROBADO	APROBADO	96.4404	57.3983	18938	52.1005	28.6991
APROBADO	NO APROBADO	84.1072	42.6017	14056	38.6696	21.3008
NO APROBADO	APROBADO	3.5596	20.8346	699	1.9230	10.4175
NO APROBADO	NO APROBADO	15.8928	79.1654	2656	7.3069	39.5835

Figura 41 Comparación Entre la variable predicha vs la Real

La comparación entre el porcentaje de aciertos entre la predicción del variable objetivo y su valor real (Figura 41). Los resultados presentados toman en cuenta a los datos tanto de entrenamiento como validación, los mismos que son semejantes para cada una de las categorías comparadas y mostrando que el modelo estable para diferentes

muestras de datos. Como se observa en la comparación Aprobado de la variable real vs Aprobado de la variable predicha la misma tiene un valor de 28% de coincidencias, lo mismo sucede para la comparación No aprobado para las dos variables donde se obtiene un porcentaje del 40%, En conjunto la coincidencia entre las categorías de las variables real y predicha suman un total del 68% aproximadamente para la dos muestra de datos. A continuación, se presenta de manera gráfica los resultados anteriormente analizados (Ver Figura 42).

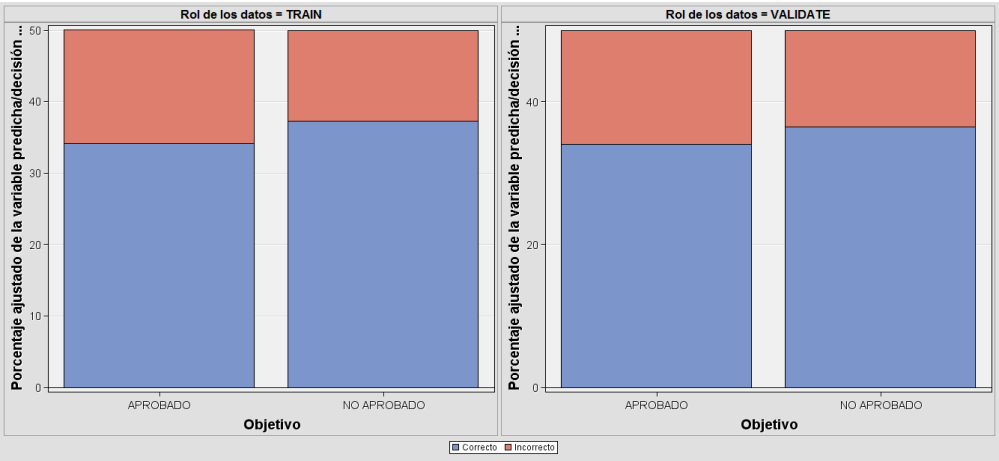


Figura 42 Comparación entre variable objetivo y real

2.8.6. Resultados Regresión Logística

La significancia de las variables explicativas seleccionadas para el modelo de regresión logística, siendo todas significativas con un P-valor menor a 0.05. Como se puede observar la mayoría de variables utilizadas en el árbol de decisión están presente en la regresión logística, lo cual muestra una consistencia en la selección de variable que explican en el modelo (Ver Figura 43).

Effect	DF	Wald Chi-Square	Pr > ChiSq
ESTADO_ACTUAL_ESTUDIOS	1	113.0457	<.0001
HAN_USUADO_INTERNET	1	107.3864	<.0001
HORAS_DEDICA_ESTUDIAR	4	310.4900	<.0001
PCTL_EDAD	3	1861.6505	<.0001
TG_ASPIRACION_NIVEL_ESTUDIOS	4	653.3606	<.0001
TG_CUAL_ES_TU_MATERIA_FAVORITA	4	159.2170	<.0001
TG_NACIONALIDAD_INDIGENA	1	343.1625	<.0001
TG_OCUPACION_ASPIRANTE	4	285.0525	<.0001
TG_PROVINCIA_COLEGIO	2	8.0718	0.0177
TG_PROVINCIA_NACIMIENTO	2	65.2201	<.0001
TG_PROVINCIA_RESIDENCIA	2	38.8124	<.0001
TG_TIENES_PLANES_A_FUTURO	2	421.5897	<.0001
TG_TIPO_DISCAPACIDAD	1	135.6566	<.0001
TG_TIPO_EMPLEO_PADRE	4	69.1842	<.0001
TG_TIPO_PREPARACION_EXAMEN	4	661.2922	<.0001
TIENEN_CUENTAS_REDES_SOCIALES	1	244.1719	<.0001
TIENE_INTERNET	1	251.3144	<.0001

Figura 43 Significancia de las Variables Regresión Logística

Los coeficientes estimados que forman parte de la ecuación de la regresión logística (Ver Figura 44).

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-1.5703	0.0522	904.12	<.0001	0.208
ESTADO_ACTUAL_ESTUDIOS	Graduado periodo actual	1	0.1340	0.0126	113.05	<.0001	1.143
HAN_USUADO_INTERNET	No	1	0.1252	0.0121	107.39	<.0001	1.133
HORAS_DEDICA_ESTUDIAR	3 horas	1	-0.1924	0.0245	61.85	<.0001	0.825
HORAS_DEDICA_ESTUDIAR	4 horas o más	1	-0.0991	0.0236	17.66	<.0001	0.906
HORAS_DEDICA_ESTUDIAR	De 1 a 2 horas	1	-0.0251	0.0215	1.36	0.2437	0.975
HORAS_DEDICA_ESTUDIAR	Menos de 1 hora	1	0.3066	0.0247	154.16	<.0001	1.359
PCTL_EDAD	01:low-17	1	-0.5794	0.0888	42.55	<.0001	0.560
PCTL_EDAD	02:17-19	1	-0.4820	0.0332	210.31	<.0001	0.618
PCTL_EDAD	03:19-21	1	0.3958	0.0325	147.96	<.0001	1.486
TG_ASPIRACION_NIVEL_ESTUDIOS	DOCTORADO (PHD)	1	-0.3377	0.0172	387.70	<.0001	0.713
TG_ASPIRACION_NIVEL_ESTUDIOS	ESPECIALIDAD	1	-0.1026	0.0206	24.80	<.0001	0.902
TG_ASPIRACION_NIVEL_ESTUDIOS	MAESTRÍA	1	-0.0555	0.0203	7.46	0.0063	0.946
TG_ASPIRACION_NIVEL_ESTUDIOS	SUPERIOR O TERCER NIVEL (UNIVERS	1	0.1980	0.0154	164.73	<.0001	1.219
TG_CUAL_ES_TU_MATERIA_FAVORITA	BIOLOGÍA	1	-0.0524	0.0217	5.81	0.0159	0.949
TG_CUAL_ES_TU_MATERIA_FAVORITA	HISTORIA Y CIENCIAS SOCIALES	1	0.0711	0.0197	13.07	0.0003	1.074
TG_CUAL_ES_TU_MATERIA_FAVORITA	LENGUA Y LITERATURA	1	0.1937	0.0202	91.79	<.0001	1.214
TG_CUAL_ES_TU_MATERIA_FAVORITA	MATEMÁTICA	1	-0.1990	0.0207	92.91	<.0001	0.820
TG_NACIONALIDAD_INDIGENA	NO APLICA	1	-0.2610	0.0141	343.16	<.0001	0.770
TG_OCUPACION_ASPIRANTE	DESEMPLEADO CESANTE	1	-0.0431	0.0221	3.79	0.0516	0.958
TG_OCUPACION_ASPIRANTE	ESTUDIANTE O RECIÉN GRADUADO (HA	1	-0.2595	0.0191	184.14	<.0001	0.771
TG_OCUPACION_ASPIRANTE	SOLO TRABAJO	1	0.1405	0.0178	62.52	<.0001	1.151
TG_OCUPACION_ASPIRANTE	TRABAJO Y ESTUDIO	1	-0.0632	0.0210	9.05	0.0026	0.939
TG_PROVINCIA_COLEGIO	GUAYAS	1	0.0776	0.0499	2.42	0.1199	1.081
TG_PROVINCIA_COLEGIO	PICHINCHA	1	-0.1393	0.0513	7.36	0.0067	0.870
TG_PROVINCIA_NACIMIENTO	GUAYAS	1	0.0885	0.0400	4.90	0.0269	1.093
TG_PROVINCIA_NACIMIENTO	PICHINCHA	1	-0.2408	0.0364	43.66	<.0001	0.786
TG_PROVINCIA_RESIDENCIA	GUAYAS	1	0.3168	0.0508	38.81	<.0001	1.373
TG_PROVINCIA_RESIDENCIA	PICHINCHA	1	-0.2384	0.0502	22.52	<.0001	0.788
TG_TIENES_PLANES_A_FUTURO	CASI SIEMPRE	1	-0.1334	0.0127	110.14	<.0001	0.875
TG_TIENES_PLANES_A_FUTURO	SIEMPRE	1	-0.1481	0.0123	145.47	<.0001	0.862
TG_TIPO_DISCAPACIDAD	NINGUNA	1	-0.4204	0.0361	135.66	<.0001	0.657
TG_TIPO_EMPLEO_PADRE	NO SÉ	1	0.0200	0.0205	0.96	0.3282	1.020
TG_TIPO_EMPLEO_PADRE	TIENE TRABAJO POR TEMPORADAS	1	0.0295	0.0178	2.74	0.0976	1.030
TG_TIPO_EMPLEO_PADRE	TIENE UN TRABAJO PAGADO ESTABLE	1	-0.1229	0.0164	56.11	<.0001	0.884
TG_TIPO_EMPLEO_PADRE	TRABAJA OCASIONALMENTE	1	-0.00970	0.0192	0.25	0.6142	0.990
TG_TIPO_PREPARACION_EXAMEN	AUTOPREPARACIÓN	1	-0.0437	0.0176	6.16	0.0131	0.957
TG_TIPO_PREPARACION_EXAMEN	CURSO PREUNIVERSITARIO PRIVADO	1	-0.7164	0.0365	384.95	<.0001	0.489
TG_TIPO_PREPARACION_EXAMEN	EN TU COLEGIO	1	0.3440	0.0179	368.95	<.0001	1.411
TG_TIPO_PREPARACION_EXAMEN	PREPARACIÓN A TRAVÉS DE LA PLATA	1	0.1981	0.0207	91.31	<.0001	1.219
TIENEN_CUENTAS_REDES_SOCIALES	No	1	0.1903	0.0122	244.17	<.0001	1.210
TIENE_INTERNET	No	1	0.1548	0.00976	251.31	<.0001	1.167

Figura 44 Valor de los Coeficientes Regresión Logística

Cada uno de los valores de los coeficientes aporta a aumentar o disminuir la probabilidad de que el aspirante no apruebe el examen (Evento). En el caso en los que el coeficiente es negativo un aumento en esa variable la probabilidad del evento disminuye y viceversa en el caso de coeficientes positivos, un aumento en la variable hace que la probabilidad de que el evento suceda aumente. En el caso del estudio en mención, todas las variables son del tipo categórico por lo que, para cada variable se presentará un valor de coeficiente para todas sus categorías excepto una (su valor es recogido por la constante de la ecuación) y su aporte a la probabilidad de que suceda el evento es igual a lo anteriormente descrito. Por ejemplo, la variable edad ha sido dividida en 4 categorías, de la cual la última mayor a 21 años no se muestra en la estimación y su valor será considerado en el coeficiente de la regresión (cuando el resto de categorías tomen el valor de cero). Ahora, observando las categorías que, si tienen coeficientes, dos de ellas son de signo negativo (menor a 19 años), esto quiere decir que disminuyen o castigan a la probabilidad de no aprobar el examen; mientras que la categoría que tiene signo positivo (mayor o igual a 19 años) aporta o aumenta la probabilidad de no aprobar el examen.

Tabla de clasificación

Rol de los datos=TRAIN Variable objetivo=ESTADO_EXAMEN Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total	Porcentaje ajustado de la variable predicha/decisión
APROBADO	APROBADO	96.2838	67.8143	104413	61.5552	33.9072
APROBADO	NO APROBADO	80.9977	32.1857	49556	29.2150	16.0928
NO APROBADO	APROBADO	3.7162	25.7409	4030	2.3758	12.8705
NO APROBADO	NO APROBADO	19.0023	74.2591	11626	6.8539	37.1295

Rol de los datos=VALIDATE Variable objetivo=ESTADO_EXAMEN Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total	Porcentaje ajustado de la variable predicha/decisión
APROBADO	APROBADO	96.1120	67.8063	22372	61.5478	33.9031
APROBADO	NO APROBADO	81.2576	32.1937	10622	29.2223	16.0968
NO APROBADO	APROBADO	3.8880	26.9747	905	2.4898	13.4876
NO APROBADO	NO APROBADO	18.7424	73.0253	2450	6.7402	36.5134

Figura 45 Comparación Entre la variable predicha vs la Real

El grafico anterior presenta la comparación entre el porcentaje de aciertos entre la predicción de la variable objetivo y su valor real (Ver Figura 45). Los resultados presentados toman en cuenta a los datos tanto de entrenamiento como validación, los mismos que son semejantes para cada una de las categorías comparadas y mostrando que el modelo estable para diferentes muestras de datos. Como se observa en la comparación Aprobado de la variable real vs Aprobado de la variable predicha la misma tiene un valor de 33% de coincidencias, lo mismo sucede para la comparación No aprobado para las dos variables donde se obtiene un porcentaje del 37%, En conjunto la coincidencia entre las categorías de las variables real y predicha suman un total del 70% aproximadamente para la dos muestra de datos. A continuación, se presenta de manera gráfica los resultados anteriormente analizados (Ver Figura 46).

Figura 46 Comparación entre variable objetivo y real

2.8.7. Etapa Evaluar

En esta etapa se realiza tanto la comparación entre modelos estimados en la etapa anterior, así como también se presenta los indicadores de validación del modelo elegido. El grafico de Curva Roc (Ver Figura 47), para las tres muestras Entrenamiento, Validación y Prueba y para los dos modelos estimados. Como se observa el Modelo de Regresión Logística presenta una mayor área bajo la curva, aunque no muy alejada de la del árbol de decisión. Con respecto a las muestras de datos no se observa una mayor diferencia entre las misma lo que hace pensar que los resultados de los dos modelos son robustos.

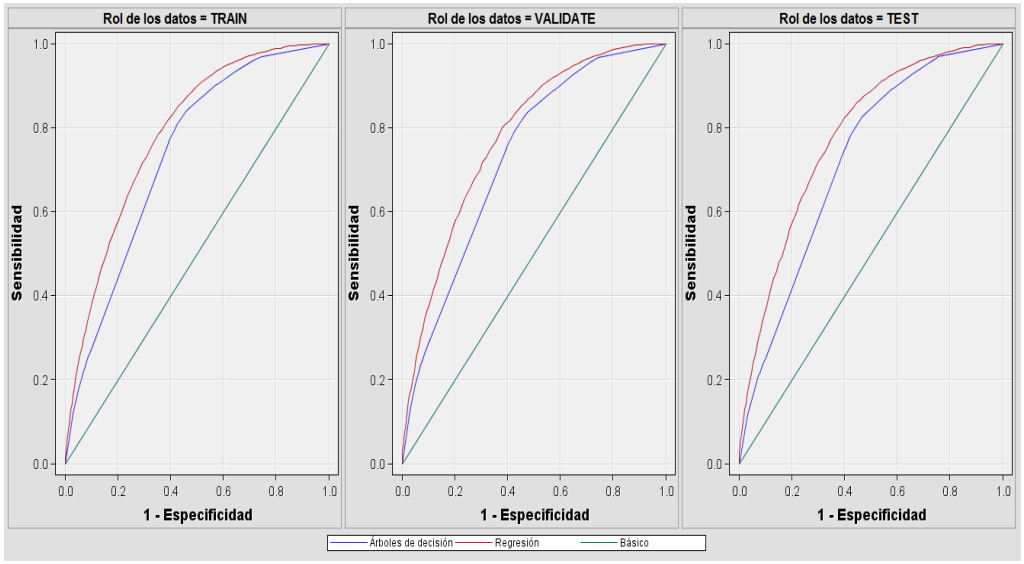


Figura 47 Curva ROC Comparación de Modelos

Los resultados obtenidos entre la variable real y la predicha de las muestras de entrenamiento y validación, donde se observa una pequeña ganancia del modelo de regresión logística frente al árbol de decisión, pero no es significativa. Con respecto a las dos muestras los resultados obtenidos son semejantes, lo cual muestra la consistencia del modelo (Ver Figura 48).

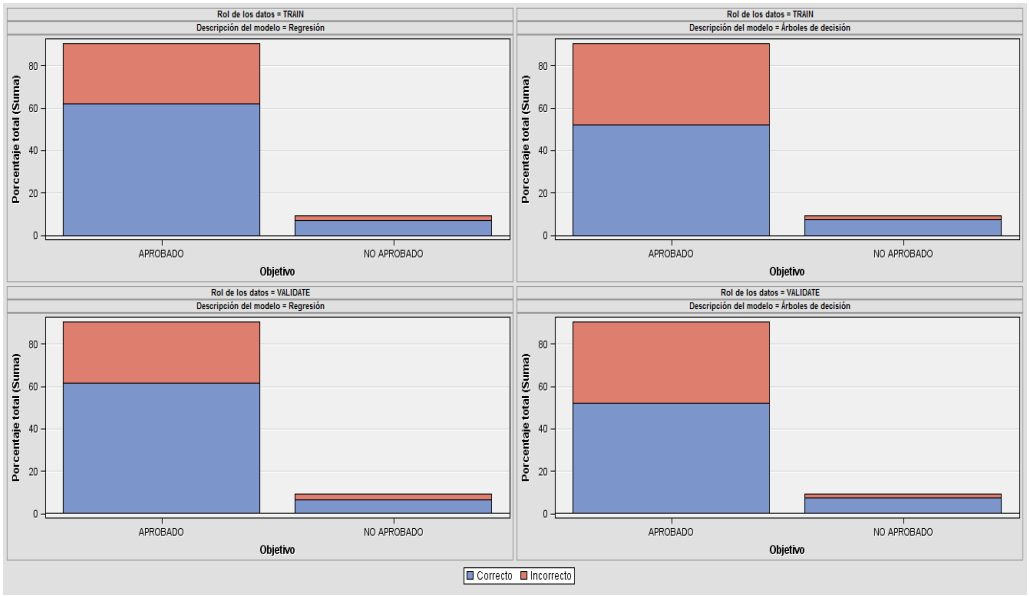
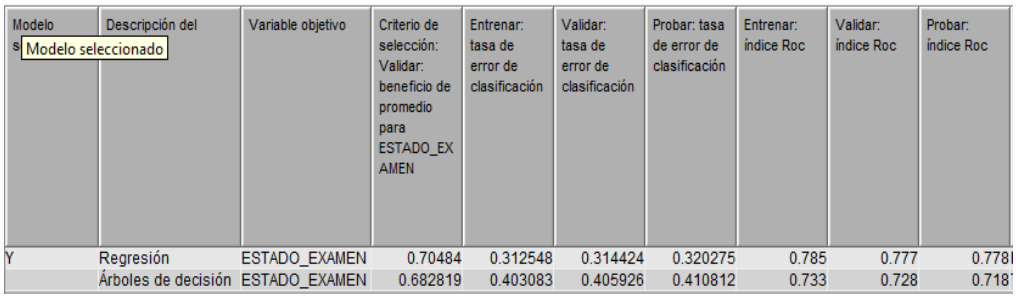


Figura 48 Comparación entre variable objetivo y real

Finalmente, se observa los indicadores de la Curva Roc y la tasa de Error de clasificación tanto para el modelo de regresión como árbol de decisión. El valor de la curva ROC para las diferentes muestras de datos tiene un valor aproximado de 0.77, el mismo que es adecuado para la estimación del modelo; En el caso de clasificación incorrecta la misma no supera el 32% en las diferentes muestras, este valor para construcción de modelos en minería de datos es aceptable y muestra una eficiencia de estimación aproximada del 70% (Ver Figura 49).



Modelo s	Descripción del Modelo seleccionado	Variable objetivo	Criterio de selección: Validar: beneficio de promedio para ESTADO_EXAMEN	Entrenar: tasa de error de clasificación	Validar: tasa de error de clasificación	Probar: tasa de error de clasificación	Entrenar: índice Roc	Validar: índice Roc	Probar: índice Roc
Y	Regresión	ESTADO_EXAMEN	0.70484	0.312548	0.314424	0.320275	0.785	0.777	0.778
	Árboles de decisión	ESTADO_EXAMEN	0.682819	0.403083	0.405926	0.410812	0.733	0.728	0.718

Figura 49 Indicadores de comparación entre Modelos y muestras de datos.

CAPÍTULO III

3. SOLUCIÓN DE BUSINESS INTELLIGENCE

3.1. INTRODUCCIÓN BUSINESS INTELLIGENCE

Una de las claves para llegar al éxito en una organización o empresa, es tener la capacidad de tomar decisiones en base a información que esté disponible cuando y donde se la requiera, para alcanzar estos objetivos es necesario migrar de los sistemas tradicionales cuyos fines están generalmente orientados a recopilar datos hacia sistemas de información que procesen datos y generen información.

Es en este ámbito donde la Inteligencia de Negocio ofrece un conjunto herramientas informáticas que facilitan el análisis, extracción, depuración y carga de datos alimentando indicadores que apoyan la toma de decisiones de directivos y usuarios de la organización, para asegurar el éxito de una implementación de este tipo, es importante organizar el trabajo en varias fases como se muestra en la Figura 51.

El Sistema Nacional de Nivelación y Admisión (SNNA) hasta la fecha actual, no cuenta con una solución de Business Intelligence para realizar un análisis del comportamiento histórico de los procesos de admisión por lo tanto es complicado generar diferentes escenarios que permitan al nivel estratégico de la institución modificar o crear políticas públicas en beneficio de la ciudadanía.

3.2. ARQUITECTURA DE UNA PLATAFORMA DE BUSINESS INTELLIGENCE

3.2.1. Propuesta de la solución de Business Intelligence

La plataforma informática del SNNA recepta aproximadamente 280.000 inscripciones por semestre generando grandes volúmenes de datos que crecen constantemente según el régimen de estudios que corresponda Sierra o Costa, esto genera alta transaccionalidad en las etapas de inscripción, postulación, asignación y aceptación de cupos, haciendo que el performance del motor de la base de datos llega al límite de su capacidad y en algunos casos se ha optado por controlar el acceso

mediante el ultimo digito de la cedula, por lo expuesto, durante estos días las consultas o procesos masivos que se ejecuten en la base de datos afecta la calidad de los servicios que se brinda a la ciudadanía

La solución que se ha diseñado e implementado en esta investigación estará implementada en una nueva infraestructura sobre un motor de base de datos PostgreSQL para evitar competir por los mismos recursos en picos de consumo, además los indicadores implementados serán los que generalmente han sido solicitados con mayor frecuencia por los *stakeholders*.

El objetivo principal es automatizar las tareas repetitivas y operativas de limpieza de datos para dejar consistente la data con lo cual se reducirán los tiempos de respuesta fundamentalmente.

3.3. ETAPAS PARA IMPLEMENTAR LA SOLUCIÓN.

3.3.1. Etapa de Planificación

El RoadMap diseñado para implementar una solución de B.I. en el SNNA, se describe a continuación (Ver Figura 50):

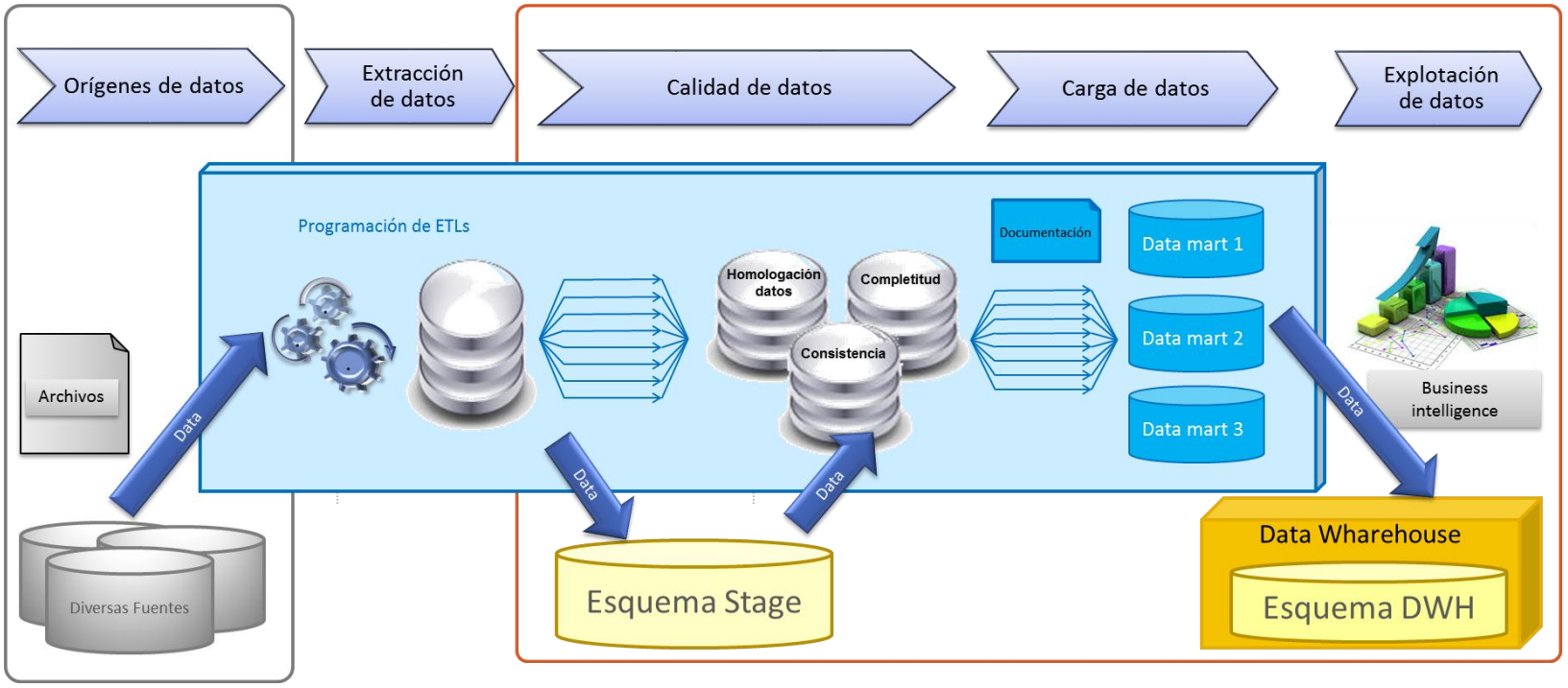


Figura 50 Flujo de Trabajo para implementar una solución de B.I.

La Figura 50 muestra un proceso normal para implementar soluciones de este tipo el cual inicia con la identificación de los orígenes de datos, la base de datos destino, el desarrollo de los ETLs y los procedimientos de validación y control de la data que deben ser aprobados por el usuario funcional y quedar debidamente documentados.

Iniciaremos con la instalación de un servidor virtual con las siguientes características (Ver Tabla 4):

Tabla 4
Recursos Hardware

Concepto	Capacidad
Memoria RAM	8Gb
Procesamiento	8 Procesadores Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz
Espacio en disco duro	50 Gb

Sobre el sistema operativo del servidor, se instalará el motor de base de datos PostgreSQL (Ver Tabla 5)

Tabla 5
Recursos software

Concepto	Software / Versión
Sistema Operativo	CentOS Linux release 7.1.1503
Base de Datos	Postgresql 9.6
Integración de Datos (ETLs)	Pentaho Data Integration Community Edition v4.4.0 (SPOON)
Creación de esquemas (cubos)	Pentaho Schema Workbench v3.5.
Generador de reportes	TABLEU 10.3
Entorno de Ejecución	Java Runtime Environment v1.7.0_79, 64 bits

3.3.2. Etapa de Análisis del negocio.

La visión de este proyecto a largo plazo debe ser, obtener el historial educativo de los estudiantes del sistema educativo del país, para ello se debería llegar a integrar con las fuentes de datos del Ministerio de Educación, entidad rectora de la educación inicial y media, no obstante, se deben llegar a concretar acuerdos o convenios interinstitucionales para logarlo. El alcance del diseño de la presente investigación está enmarcado en los datos con los que ya cuenta el SNNA, no obstante el diseño toma en cuenta la integración con las bases de datos de las Instituciones de Educación Superior (IES) que se podría retomar en una siguiente investigación debido a la envergadura del proyecto y todo lo que involucraría no solo en el ámbito técnico sino administrativo para estandarizar y definir protocolos de transferencia de datos mediante diferentes mecanismos que hoy por hoy ofrece la tecnología actual (Ver Figura 51).

Las fases del proyecto se describen a continuación:

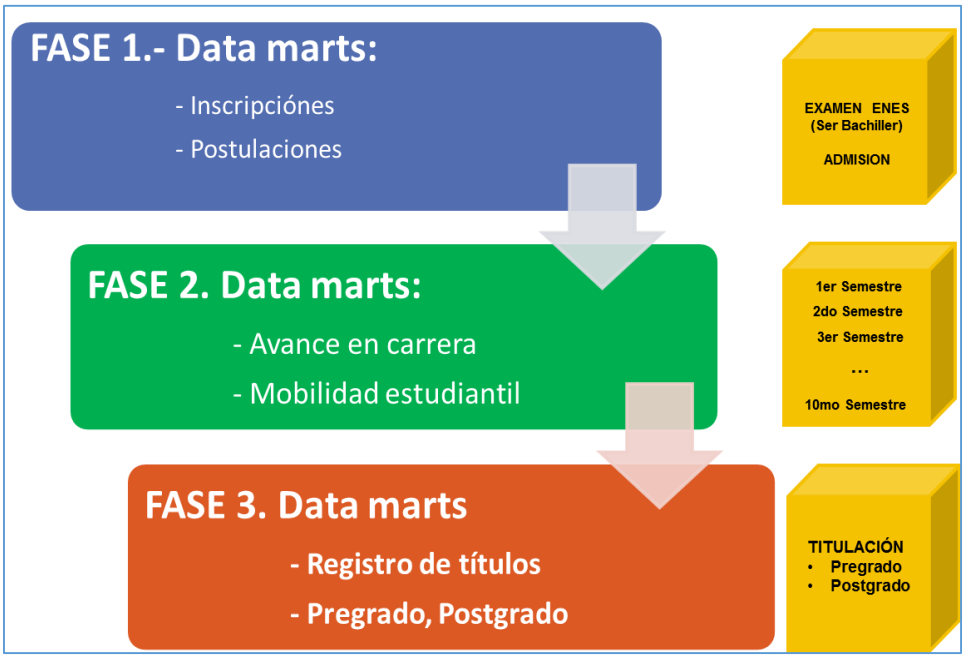


Figura 51 Fases de implementación del proyecto

El alcance de esta implementación estará enmarcado en la Fase Nº 1, es decir se desarrollara un esquema para soportar el proceso de admisión a las IES y permite

visualizar las variables que influyen en la no aprobación del Examen de Acceso a la educación Superior, mismas que se analizaron en el capítulo II de minería de datos, dejando el camino trazado para un siguiente trabajo de investigación que continúe con el resto de fases.

3.3.3. Etapa de Diseño.

El esquema de implementación de la solución muestra los orígenes de datos que alimentan el cubo de información y los principales beneficiarios (Ver Figura 52).

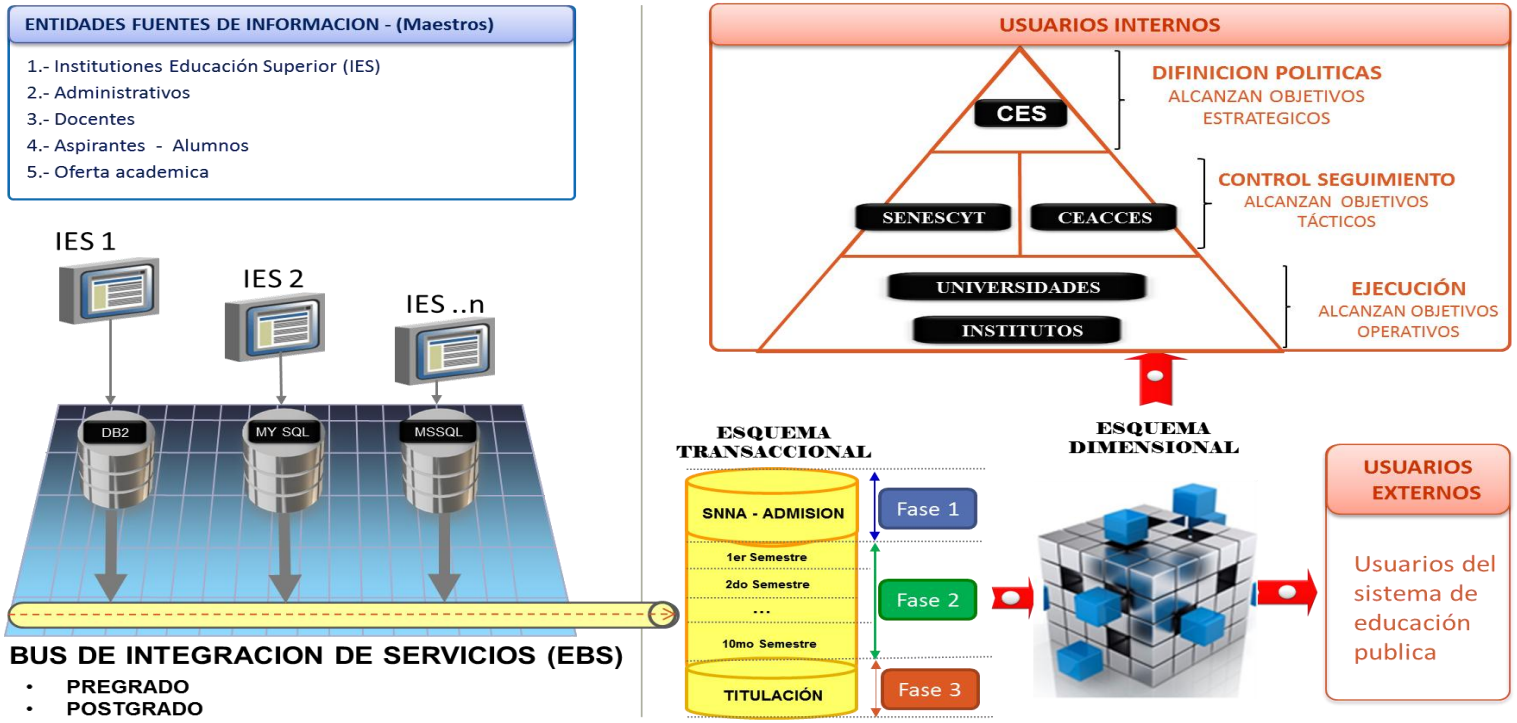


Figura 52 Esquema de implementación de la solución

La Fase 1 consiste en migrar del esquema transaccional al esquema dimensional de la base de datos, mientras que para alcanzar la Fase 2, es necesario la implementación de un Bus para transferencia de datos desde las Instituciones de educación superior hacia la base centralizada de SENESCYT y finalmente la Fase 3 consiste en consumir la información de la base de titulación para cerrar el ciclo de educación superior (Ver Figura 52), tómese en cuenta que en esta solución no toma en cuenta los datos de educación media debido a que no conozco la arquitectura de datos o infraestructura sobre la cual operan.

Esta implementación permitirá examinar los resultados desde diferentes perspectivas, desde lo estratégico (información más cohesionada), hasta lo operativo (Información más granular).

Los indicadores que se desean implementar son los siguientes (Ver Tabla 6)

Tabla 6
Indicadores que se implementaran en el BI

Nº	DETALLE
1	Inscritos por edad y régimen de estudios que aprueban o reprueban el examen
2	Distribución de notas alcanzadas por estado de aprobación
3	Número de aspirantes que trabajan y estudian por estado de aprobación
4	Número de personas por sostenimiento y estado de aprobación
5	Costo de cursos pre universitarios por provincia
6	Relación de aspiración de estudios superiores vs resultados del examen.

3.3.4. Etapa de Construcción

3.3.4.1. Modelo STAGE

El diseño de este modelo muestra las entidades que soportan el paso del esquema transaccional al esquema dimensional (Ver Figura 53)

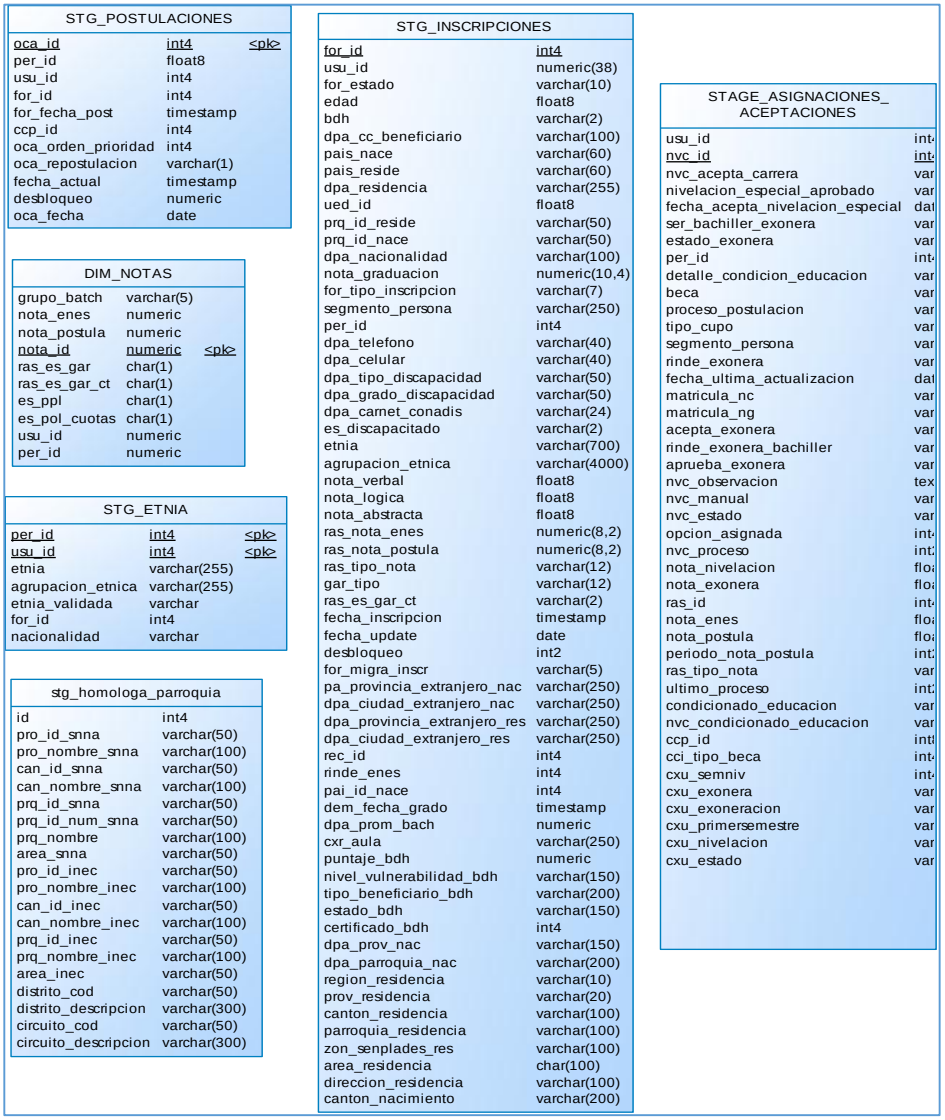


Figura 53 Diagrama del esquema STAGE

3.3.4.2. Modelo PRODUCTIVO

Este modelo muestra el diseño físico del Data Mart de Admisión (Ver Figura 54)

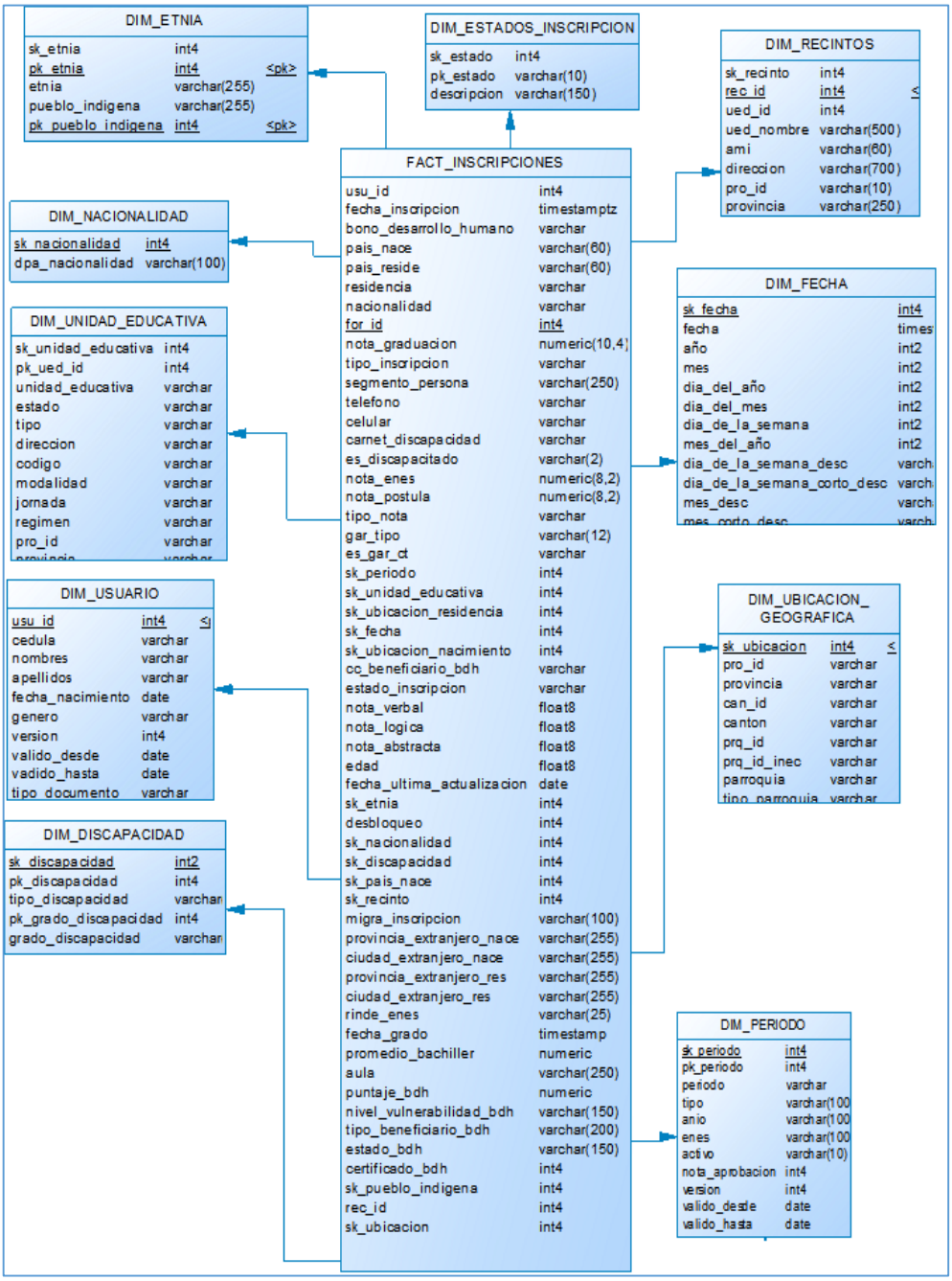


Figura 54 Diagrama del esquema de producción

3.3.5. Etapa de Implementación.

En esta fase se programó los ETLs para popular la data al esquema STAGE tal como se muestra a continuación (Ver Figura 55):

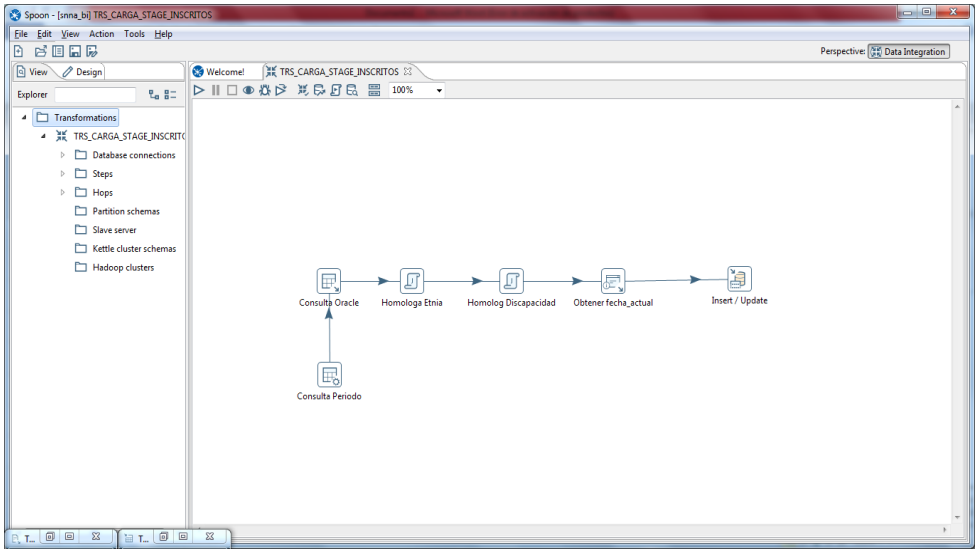


Figura 55 ETL para cargar el esquema STAGE

Luego, se programó el ETL para cargar el esquema Dimensional, se aplicaron algunos procesos de limpieza y completar datos para asegurar consistencia en la data (Ver Figura 56)

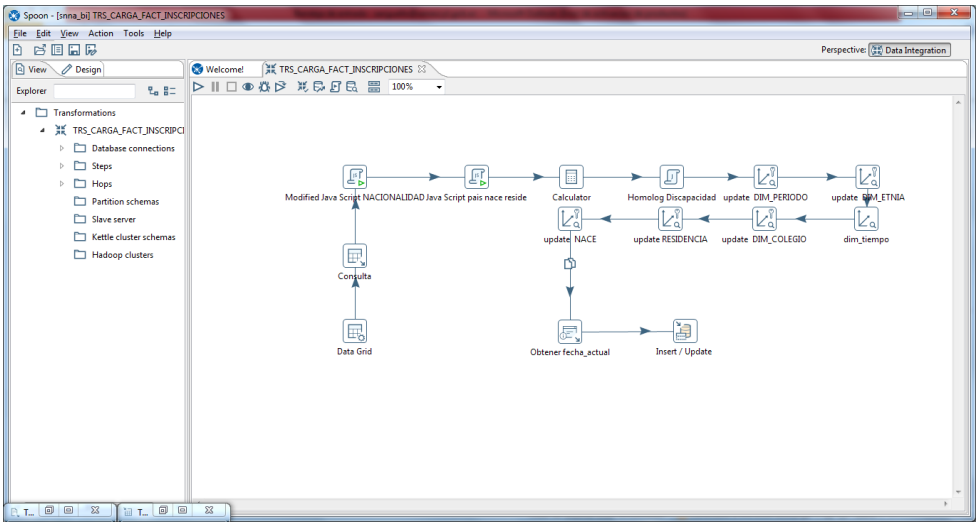


Figura 56 ETL para cargar el esquema dimensional

Para completar o actualizar los datos personales también se programó una rutina que actualiza los registros desde el servicio web del registro civil (Ver Figura 57).

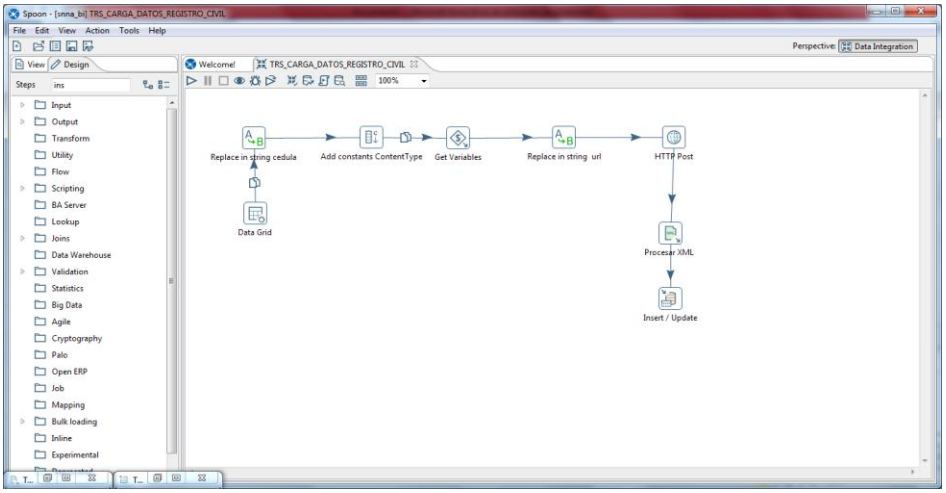


Figura 57 ETL para actualizar los datos desde el Registro Civil

Antes de ejecutar este ETL, se programó uno independiente para realizar la carga inicial de los catálogos del esquema como: la dimensión del tiempo, tipos, periodos, estados entre otros necesarios para soportar la carga de datos a la tabla de hechos (Ver Figura 58).

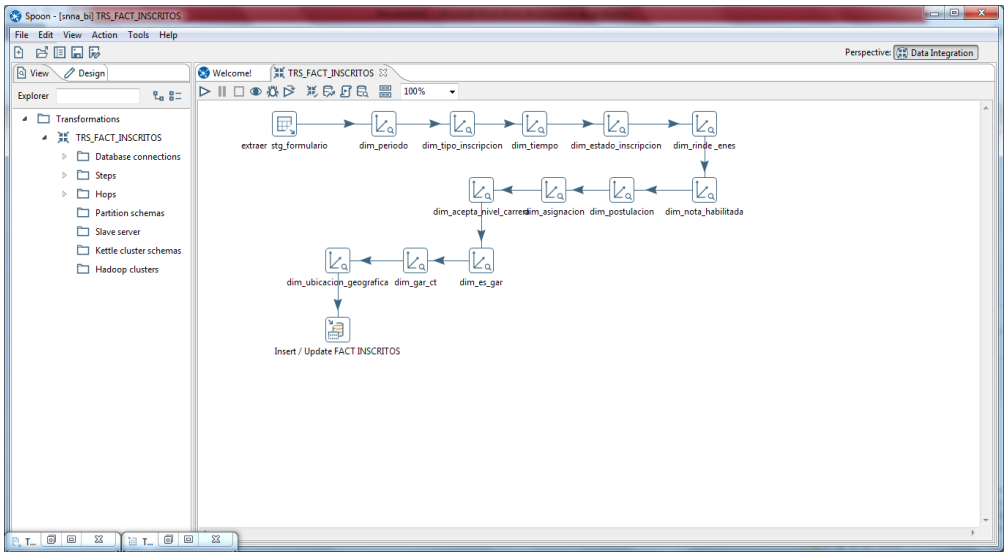


Figura 58 ETL para cargar los catálogos del data mart

Finalmente, se programó un Job para ejecutar la tarea automáticamente (Ver Figura 59).

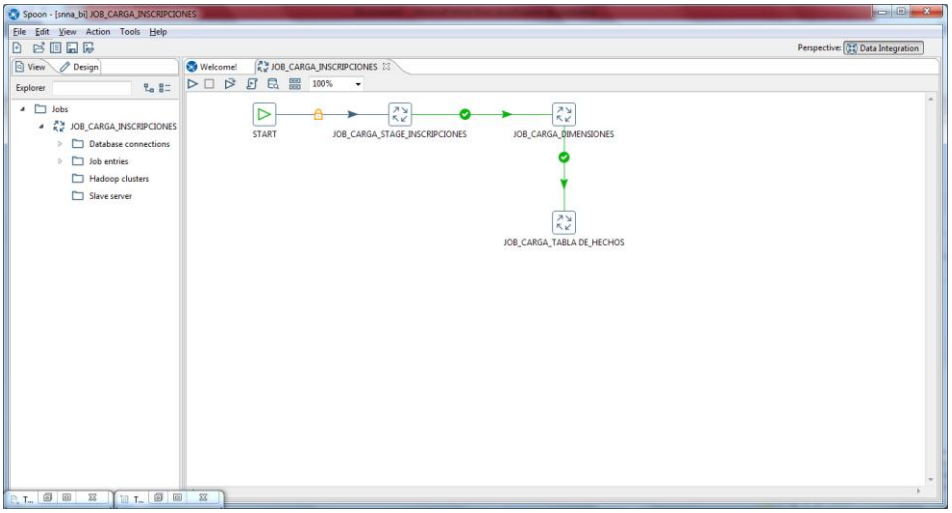


Figura 59 Tarea programada para ejecución automática

3.3.6. INDICADORES

Una vez implementado el Data Mart, hemos colocado un software para generar los indicadores que a continuación se detallan:

Indicador 1. Inscritos por edad y régimen de estudios que aprueban o reprueban el examen.

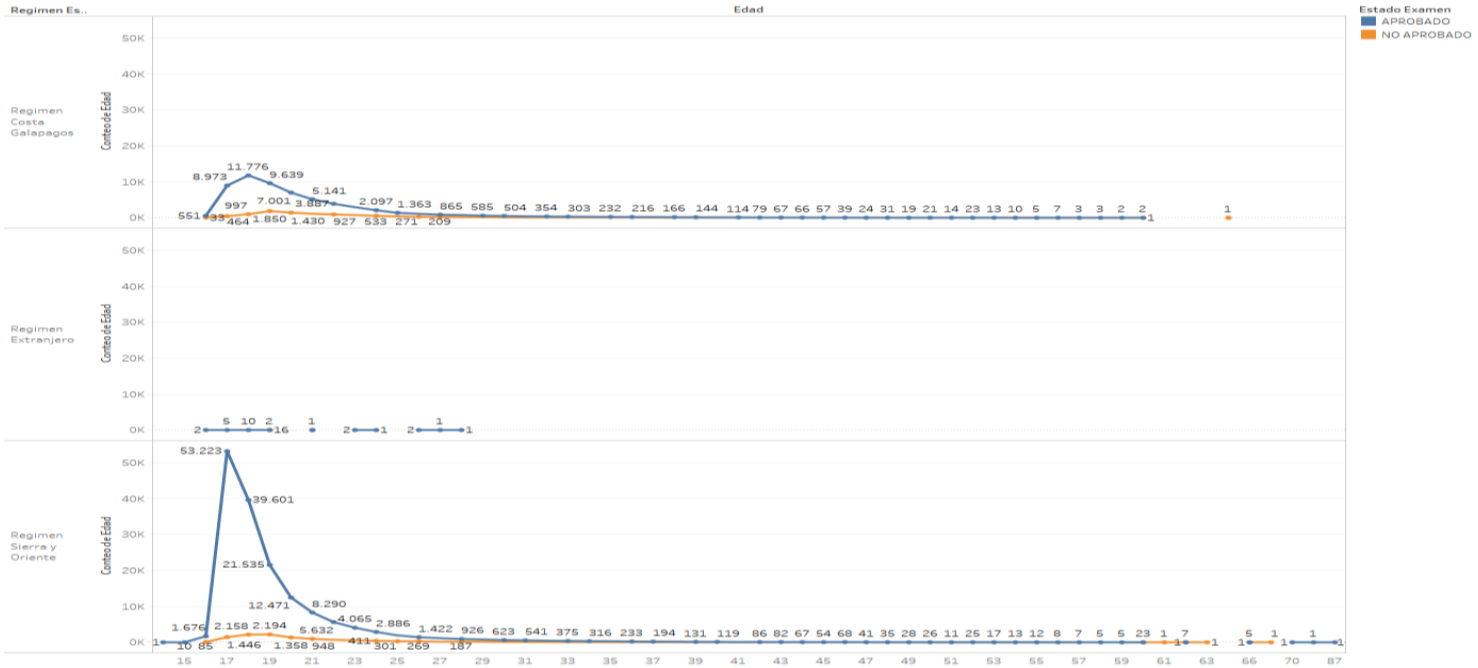


Figura 60 Inscritos por edad y régimen de estudios

La grafica muestra una tendencia clara sobre los aspirantes que aprueban el examen son aquellos que están en edad de 16 a 19 años en general para régimen sierra y costa, nos obstante en el régimen Sierra existe un pico más pronunciado que obedece a que existen un mayor número de inscritos en este régimen (Ver Figura 60).

Indicador 2. Inscritos por edad y régimen de estudios que aprueban o reprueban el examen

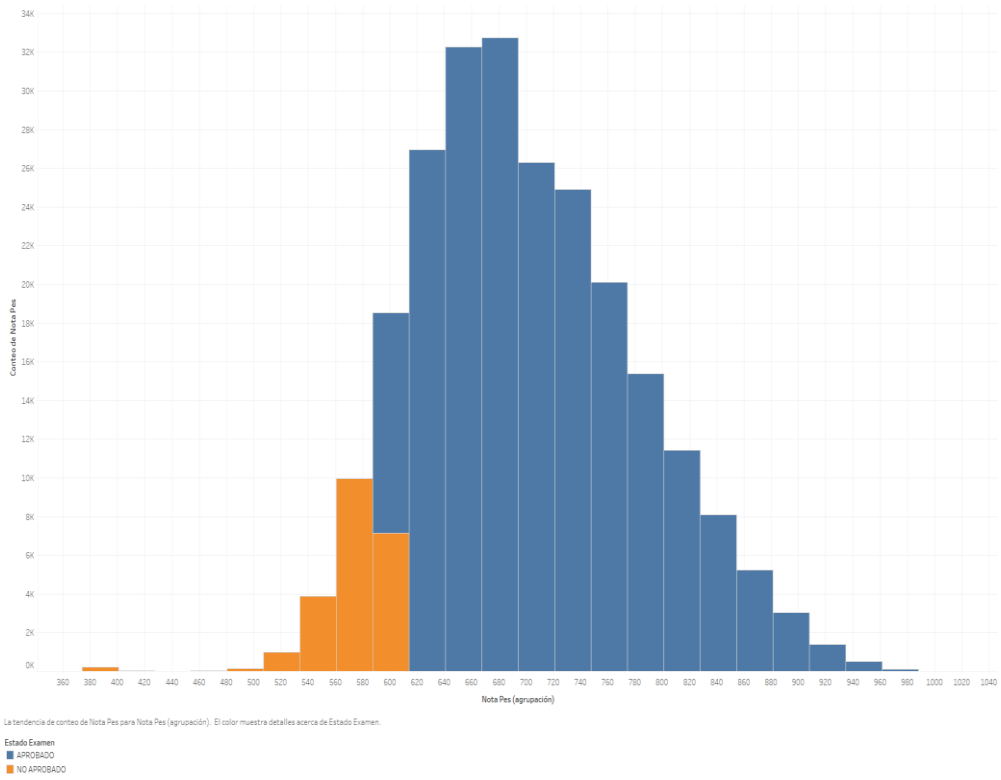


Figura 61 Inscritos por edad y régimen de estudios

La distribución de las notas según la nota y el estado de aprobación, también se visualiza que el promedio de la frecuencia más alta se encuentra entre en el rango de notas de 640 y 700 puntos (Ver Figura 61).

Indicador 3. Número de aspirantes que trabajan y estudian por estado de aprobación

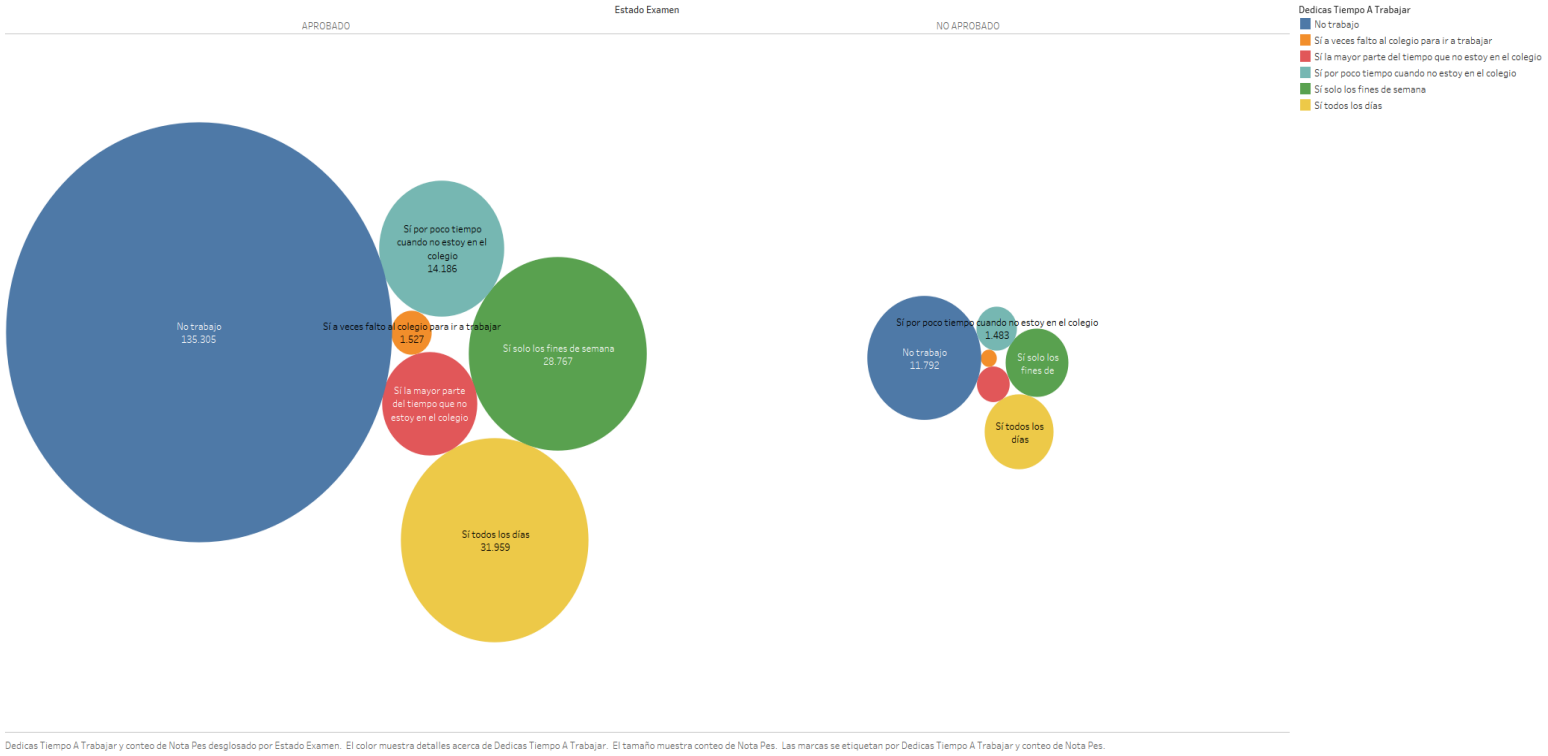


Figura 62 Número de aspirantes que trabajan y estudian

En el grafico se observa que la mayoría de alumnos no trabajan, si existe un grupo de estudiantes que se dedican a trabajar y estudiar y existe correlación entre los que aprueban y reprueban en cada caso (Ver Figura 62).

Indicador 4. Número de personas por sostenimiento y estado de aprobación

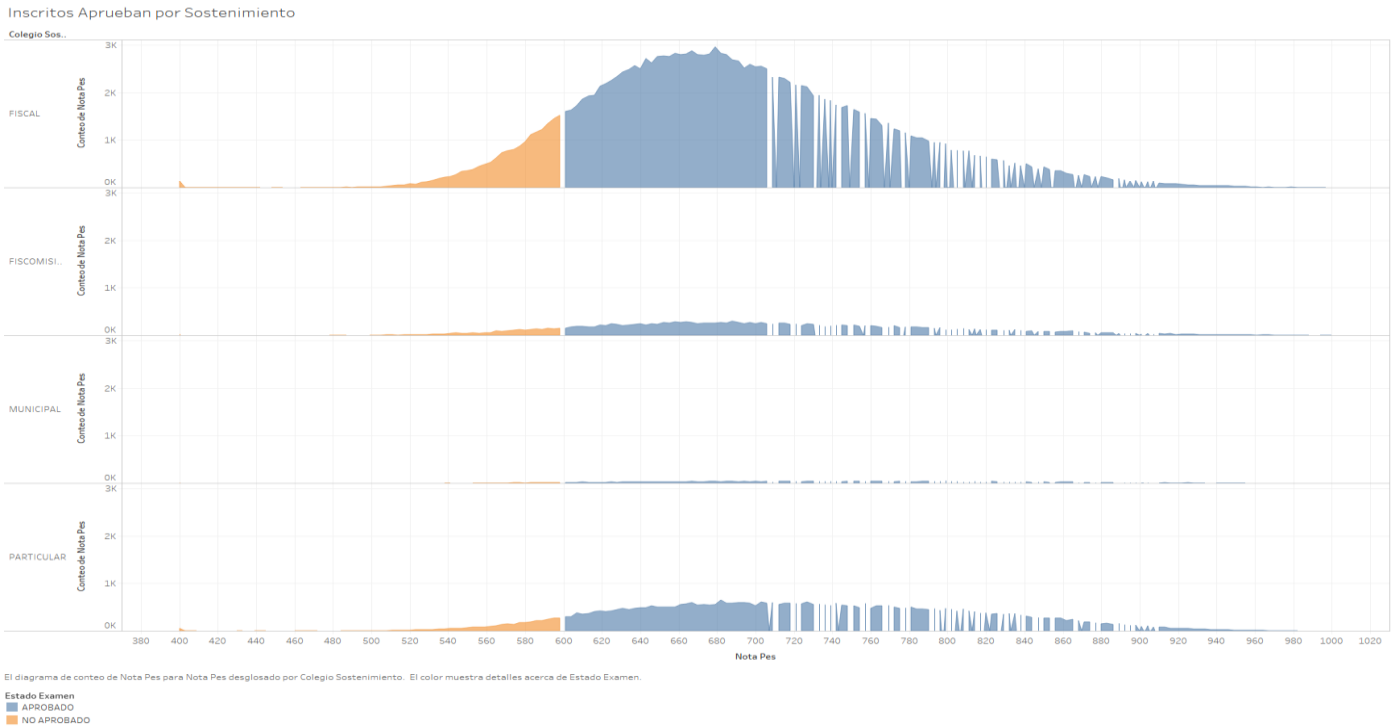


Figura 63 Número de personas por sostenimiento

El grafico muestra que el grupo que el sostenimiento de educación fiscal es el que contiene la mayor cantidad de aspirantes seguido del particular, fisco misional y municipal (Ver Figura 63).

Indicador 5. Costos de cursos pre universitario por provincia

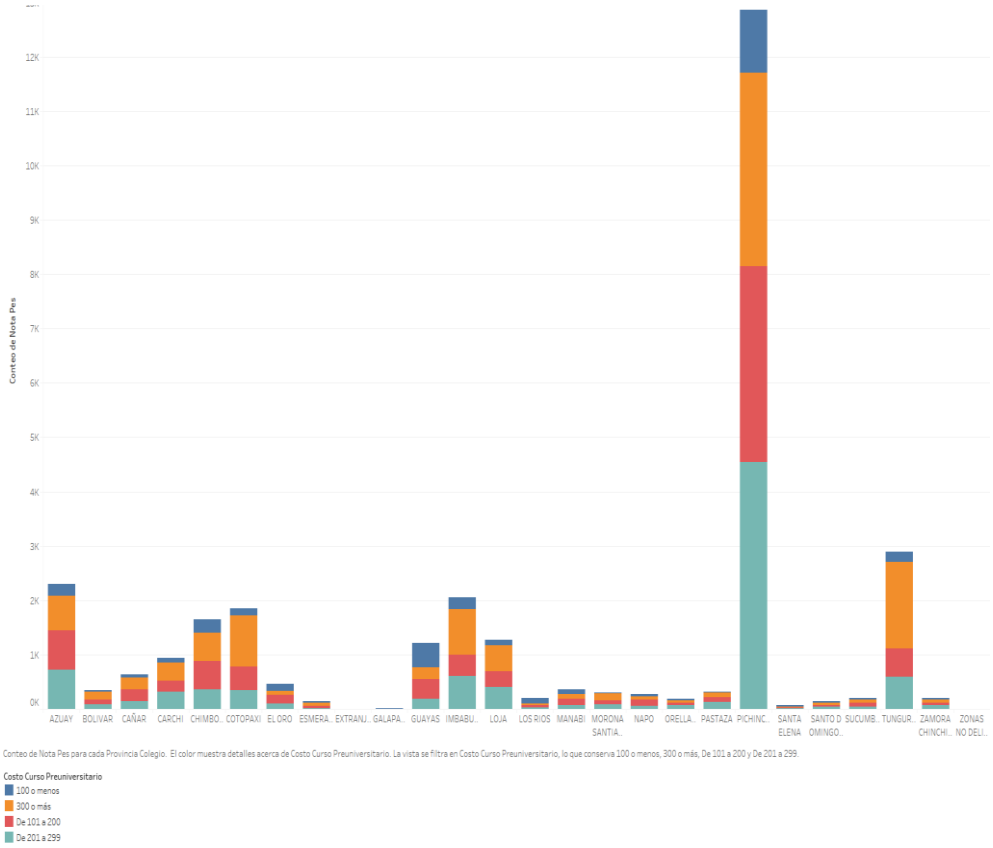


Figura 64 Costos de cursos pre universitario por provincia

Se observa que la provincia en donde más cursos preparatorios para el examen ENES se realizan son en las provincias de Pichincha, Tungurahua, Azuay e Imbabura y generalmente tienen un costo igual o mayor a 300 dólares (Ver Figura 64).

Hay que tomar en cuenta que la base de análisis de este estudio corresponde al régimen sierra, eso explica que las provincias de la región sierra son más representativas en este gráfico.

Indicador 6. Relación de aspiración de estudios superiores vs resultados del examen.

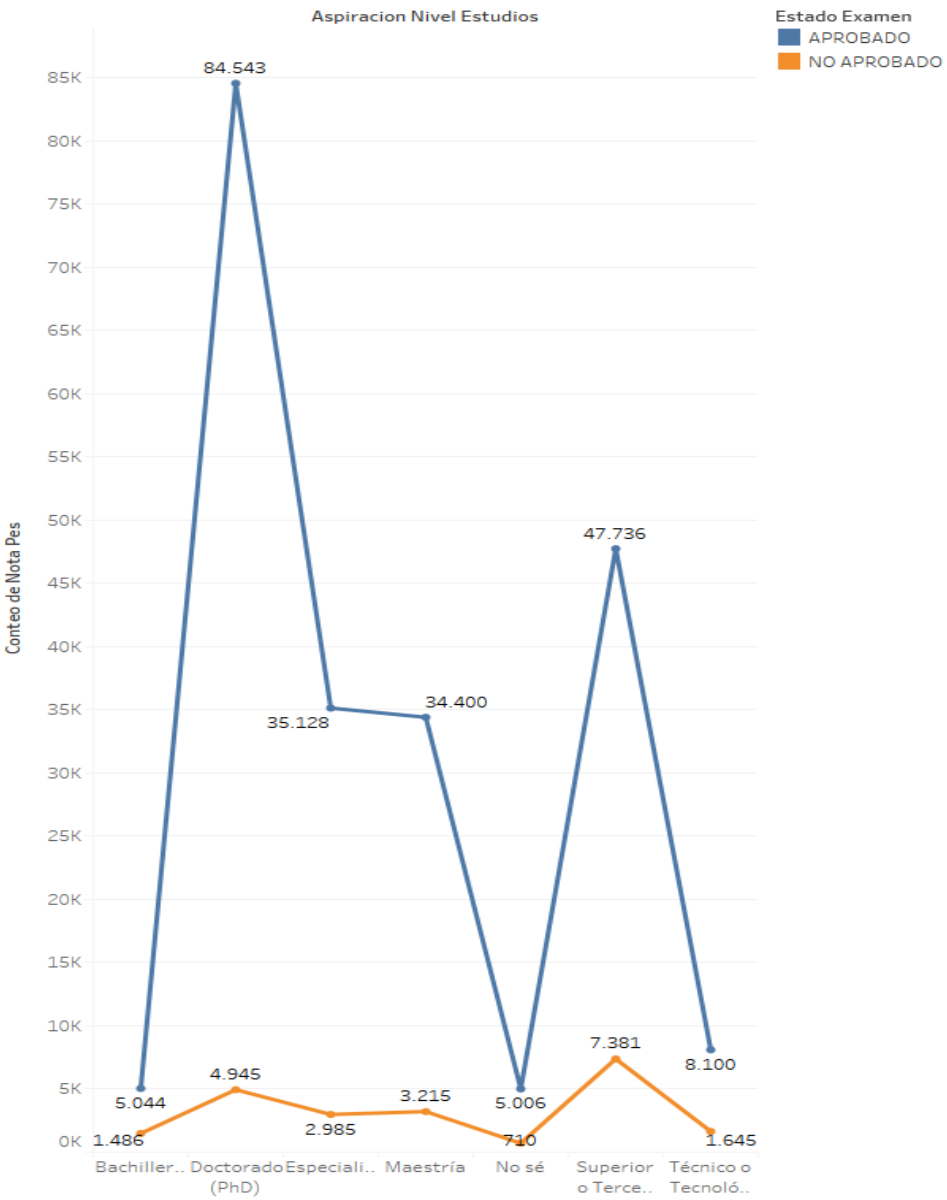


Figura 65 Relación de aspiración de estudios superiores vs resultados del examen

La relación que existe entre los aspirantes que aprueban o reprueban el examen, según las aspiraciones que tienen de continuar sus estudios superiores, se evidencia que aquellos alumnos que desean alcanzar un título de tercer nivel o doctorado obtienen mejores resultados que quienes no desean continuar sus estudios o aún no han decidido (Ver Figura 65).

CAPÍTULO IV

4. CONCLUSIONES Y RECOMENDACIONES

4.1. CONCLUSIONES

- Analizando la data de los aspirantes que rinden el examen de acceso a la Educación Superior, se identificaron 310 variables, para determinar cuáles de ellas explican la aprobación o reprobación del examen se seleccionaron las técnicas de *data mining*: árboles de decisión y regresiones logísticas, para validar los resultados obtenidos se contrastaron los dos modelos comparando el área bajo la curva ROC del modelo de clasificación de las dos técnicas y se confirma que son bastante similares, el valor promedio es aproximadamente 0,77 lo cual es aceptable para un modelo de clasificación y avala que los resultados de los modelos como consistentes y robustos, en cuanto a la tasa de clasificación esta sobre el 70%, por lo expuesto la predicción obtenida es adecuada y útil.

- El nivel de estudio superior más común (moda), al cual aspiran alcanzar los estudiantes que realizaron el examen, es el grado doctor PhD, con un 36 % del total de la población. Lo cual es interesante ya que en los resultados de la aplicación de los algoritmos se observa que esta variable, es una de las más influyentes y tienen relación directa con el desempeño del examen, es decir mientras mayor expectativa tiene los aspirantes mejores resultados obtienen en la evaluación.

- Sobre la discapacidad, los aspirantes en un 99.2% de la población no tienen ningún tipo discapacidad, es decir solo un 0.8% de estudiantes con algún tipo de discapacidad se está incluyendo o incorporando a la educación superior.

- Existe un 62% de aspirantes rezagados, que se han graduado en años anteriores y aun rinden el examen de acceso a la educación superior, manteniendo sus

aspiraciones por alcanzar un cupo y evidencia brecha que existe entre la baja oferta de cupos (65.000 en promedio por cada semestre) que reportan las universidades públicas y la alta demanda de cupos por parte del universo de aspirantes que rinden el examen (250.000 en promedio por cada semestre).

- La edad del aspirante es la característica que más explica la aprobación o reprobación del examen de admisión, la media se encuentra en 20 años, lo cual difiere de los 18 años que debería ser la edad planificada para terminar el bachillerato, esto indica que muchos aspirantes rinden más de una vez el examen ya que la normativa del SNNA no impone ninguna limitación mientras el aspirante no haya aceptado un cupo. El rango de edad de 17 a 20 años agrupa el 75% del universo.

- Otro factor importante es el acceso a internet, se encontró que el 80% de los aspirantes tienen acceso a internet y cuentan con equipos y dispositivos tecnológicos como Computadora, Smartphone, Tablet, Ipad entre otros, sin embargo se estima que la capacidad económica del porcentaje complementario no les permite adquirir o acceder a este tipo tecnología.

- En cuanto a la preparación del examen, en un 36% es auto preparación por parte del aspirante, lo cual es un factor que se debe tomar en cuenta al momento que se quiera mejorar las notas del examen, ya que el 74% se ha preparado en algún tipo de curso. La dedicación a tiempo completo para dedicarse a estudiar, el acceso a servicios básicos sobre todo en sectores rurales, son otros aspectos que influyen focalmente en ciertas ubicaciones del territorio ecuatoriano.

- La implementación de un Business Inteligence en el Sistema Nacional de Nivelación y Admisión es una solución efectiva a los múltiples problemas que ocasiona la generación de reportes *ad hoc*, la tecnología a utilizar no debe

involucrar costos de licenciamiento debido a la normativa del CODIGO INGENIOS que impulsa la Secretaria de Educación Superior Ciencia y Tecnología, por lo tanto la arquitectura que utiliza esta solución está basada en tecnologías libres para la base de datos como PostgreSQL y Data Integration de Pentaho para programar los ETLs, no obstante la herramienta de explotación de datos en esta versión es licenciada y es importante buscar alguna herramienta que sea amigable con el usuario para asegurar la aceptación y sostenibilidad del proyecto a largo plazo.

- Los indicadores implementados en esta solución son aquellos que empíricamente son más demandados, no obstante el diseño y programación del mismo, permite ir incorporando todos los que la línea de negocio demande siempre y cuanto se disponga de los datos en el esquema de datos transaccional.

4.2. RECOMENDACIONES

- En función de afianzar la democratización del acceso a la educación superior y cumplir el principio de igualdad de oportunidades, es importante identificar aquellos aspirantes que por diferentes factores tienen acceso limitado a internet, sobre todo en sectores rurales con el fin de ampliar la cobertura de los servicios básicos y acceso a internet pues esta es una de las variables que influencian para no aprobar el examen de acceso a la educación superior. El impacto de brindar estos servicios deben ser medidos para determinar si las brechas que existen entre el sector urbano y rural se van reduciendo.

- Los organizamos que regulan y operan la Educación Superior del País: CES, CEAACES, SENESCYT y las Universidades y Escuelas Politécnicas deben trabajar en un plan conjunto para ampliar la oferta académica de manera progresiva y sostenible con el fin de atender a la población que hoy por hoy no tiene oportunidad de acceder a la Educación Superior Pública. Esto se vuelve urgente ya que el avance de la sociedad depende en gran parte del nivel de educación de sus ciudadanos.

- El Sistema Nacional de Nivelación y Admisión debe diseñar políticas y ejecutar programas o proyectos que atiendan segmentos específicos de la población ya que existen grandes diferencias en relación al desempeño educativo por diversos factores que se expusieron en esta investigación, estas políticas están plenamente amparadas en la ley y se denominan cuotas o políticas de acción afirmativa y permiten al Estado Ecuatoriano implementar mecanismos que garantizan el acceso a la educación superior de estos grupos que generalmente han sido históricamente excluidos o tienen algún tipo de discapacidad.

- La implementación del Data Wharehouse para Educación Superior que se realizó en el capítulo III, está organizada en tres fases:
    - Fase 1.- Data Mart del proceso de admisión
    - Fase 2.- Data Mart del proceso de carrera.
    - Fase 3.- Data Mart del proceso de titulación

  En el presente trabajo, se llegó a implementar físicamente la Fase 1, no obstante la planificación y diseño si contiene las tres fases del proyecto y tiene una visión integral de todo el sistema de información, por lo tanto es recomendable continuar incorporando los siguientes Data Marts que corresponden a las siguientes fases y así completar el Data Wharehouse, esto permitirá al nivel estratégico de las instituciones que regentan la educación Superior como: Consejo de Educación Superior (CES), Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES) y la Secretaria de Educación Superior Ciencia y Tecnología (SENESCYT), puedan tomar decisiones informadas en base a escenarios y proyecciones que mejoren el futuro del país en ámbito de educación.

- Los resultados que expone esta investigación corresponden al proceso de admisión y específicamente a la etapa de Evaluación, sin embargo es importante continuar con el estudio de las siguientes etapas como: Postulación y Asignación de cupos, en el cual se debe analizar el comportamiento de las carreras más demandadas y el impacto que esto producirá en el cambio de la matriz productiva del país en los siguientes años.

REFERENCIAS BIBLIOGRÁFICAS

ASAMBLEA NACIONAL. (2010). *Constitución de la República.* Quito: Ediciones legales.

ASAMBLEA NACIONAL. (2010). *Ley Orgánica de Educación Superior*. Quito: Ediciones legales.

BELTRÁN, B. (2012). *Minería de datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

BOUZA, C. S. (2012). La minería de datos: árboles de decisión y su aplicación en estudios médicos. *Modelación matemática de fenómenos del medio ambiente y la salud Tomo 2*, 64-78.

CAMARGO, H., & SILVA, M. (2012). Dos caminos en la búsqueda de patrones por medio de la Minería de Datos: SEMMA y CRISP. *Revista de Tecnología Universidad del Bosque. Vol. 9 N° 1*, 11-18.

CAMPOS, M. G. (2012). *Clustering. Algoritmos de agrupamiento. Minería de Datos*. Obtenido de https://www.youtube.com/watch?v=pQrjUt8jMBM

CARVAJAL, H. (2012). *El Data Mining y su incidencia en la toma de decisiones del catastro de establecimientos y la emisión de los permisos de funcionamiento por parte de la Dirección Provincial de Salud de Cotopaxi.* Ambato: Universidad Técnica de Ambato.

CASTAÑEDA, J., & RODRÍGUEZ, M. (2003). *La minería de datos como herramienta de marketing: Delimitación y medidas de evaluación del resultado.* España: Universidad de Granada.

COMPUTER SCIENCE. (Julio de 2013). *Cross Industry Standard Process for Data Mining*. Obtenido de http://compscienceedu.blogspot.com/2013/07/crisp-dm-cross-industry-standard.html

CURTO, J. (2010). *Introducción al Business Intelligence.* Barcelona: UOC.

ESCUELA POLITÉCNICA DEL CHIMBORAZO. (2013). *Capacitación sobre el manejo del sistema UNIVERSITAS XXI* . Obtenido de http://radmision.espoch.edu.ec/index.php?option=com_content&view=article&id=105%3Acapacitacion-sobre-el-manejo-del-sistema-universitas-xxi&catid=1%3Alatest-news&Itemid=72

ESPINOZA, I., & GUTIÉRREZ, L. (2010). *La minería de datos como soporte a la toma de decisiones estratégicas de las organizaciones.* México: Instituto Politécnico Nacional.

FIGUEROA, B. (2007). *Criterios para evaluar información.* Recinto de Ponce.

GONZÁLEZ, J. (2012). *Minería de datos.* Puebla: Universidad Politécnica de Puebla.

GRUPO DE METEREOLOGÍA SANTANDER. (2015). *Minería de Datos. Redes Bayesianas y Neuronales*. Obtenido de http://www.meteo.unican.es/es/research/mineria_datos

MICROSOFT. (2016). *Algoritmo Bayes naive*. Obtenido de https://msdn.microsoft.com/es-es/library/ms174806%28v=sql.120%29.aspx

MICROSOFT. (2016). *Introducción a la minería de datos*. Obtenido de https://msdn.microsoft.com/es-es/library/dn282377%28v=sql.120%29.aspx

MILLER, D. (2007). *Measuring Business Intelligence Success: A capability Maturity Model.* Nueva York: DM Morrisey.

Mpacula. (2011). *k-means clustering*. Obtenido de http://blog.mpacula.com/2011/04/27/k-means-clustering-example-python/

PÉREZ, C. (2007). *Minería de datos: técnicas y herramientas.* Madrid: Paraninfo.

PÉREZ, C., & SANTÍN, D. (2008). *Minería de datos. Técnicas y herramientas.* Madrid: Thomson.

PINTADO, T., & SÁNCHEZ, J. (2014). *Nuevas tendencias en comunicación estratégica.* Madrid: Hesic.

SALAS, O., & CAMPA, F. (2013). *Manual del Controller.* Profit.

SANTACRUZ, R. (2015). *Santacruzramos*. Obtenido de Técnicas y herramientas de la minería de datos: https://santacruzramos.wikispaces.com/3.4.5+T%C3%A9cnicas+y+herramientas+de+la+miner%C3%ADa+de+datos.

SENESCYT. (2013). *Reglamento del Sistema Nacional de Nivelación y Admisión SNNA.* Quito: Ediciones legales.

SENESCYT. (2013). *Reglamento para la administración, uso y acceso a recursos informáticos y tecnológicos de la SENESCYT*. Obtenido de file:///C:/Users/Asistente/Downloads/Acuerdo_018_2013_Reglamento_de_TICS.pdf

SENESCYT. (2016). *Cronograma SNNA*. Obtenido de http://elyex.com/cronograma-snna-senescyt-2014-ecuador-fechas-calendario/

SENESCYT. (12 de Mayo de 2016). *Examen nacional para la Educación Superior - ENES*. Obtenido de ENES: http://www.evaluacion.gob.ec/examen-nacional-para-la-educacion-superior-enes/

SENESCYT. (12 de 05 de 2016). *Reconocimiento al mérito académico*. Obtenido de http://programasbecas.educacionsuperior.gob.ec/reconocimiento-al-merito-academico-2015/

SILVA, A., DOMÍNGUEZ, A., CORTÉS, G., CASTORENA, A., & VÁSQUEZ, M. (2007). Análisis de satisfacción de universitarios mediante la minería de datos. *Revista Iberoamericana para la investigación y el Desarrollo Educativo. Vol. 5, Núm. 10*, 1-15.

SNNA. (2016). *Requisitos para inscripción*. Obtenido de http://www.snna.gob.ec/

TIMARÁN, R. (2010). *Universidad de Nariño San Juan de Pasto*. Obtenido de Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos: file:///C:/Users/Asistente/Downloads/C692YV.pdf

UNIVERSIDAD DE TLAXCALA. (2014). *Minería de datos*. Obtenido de Modelos predictivos y descriptivos en minería de datos: http://es.slideshare.net/lalopg/mtodos-predictivos-y-descriptivos-minera-de-datos

UNIVERSIDAD TECNOLÓGICA METROPOLITANA. (1 de Junio de 2013). *Historia de Minería de Datos*. Obtenido de http://mineriadatos1.blogspot.com/2013_06_01_archive.html

UNIVERSITAT JAUME I. (2013). *XXXIV Congreso Nacional de Estadística e Investigación Operativa.* España: Castellón.

VIEIRA, L., & ORTIZ, L. (2009). *Introducción a la minería de datos.* Río de Janeiro: e-papers.

WIKISPACES. (2012). *Taller de Data Mining*. Obtenido de https://tallerbd.wikispaces.com/DataMining