ANÁLISIS DE RIESGOS ACADÉMICOS EN EL PROGRAMA DE INGENIERIA INDUSTRIAL DE LA INSTITUCIÓN POLITÉCNICO GRANCOLOMBIANO.

Autor:

Lina Maria Martinez Galindo

Tutor:

Juan Calos Gutiérrez

Institución Universitaria Politécnico Grancolombiano
Ingeniería Industrial
Bogotá D.C., Colombia

2019

CONTENIDO

CONTENIDO DE ILUSTRACIONES

CONTENIDO DE TABLAS

CONTENIDO DE GRAFICOS

1. RESUMEN:

El análisis de riesgo académico ha sido de interés en investigación a lo largo de los años. El principal motivo de esta investigación es demostrar aquellos factores que predijeran un rendimiento bajo con un alto riesgo de deserción académica. La muestra estuvo constituida por 7529 estudiantes del programa de ingeniería industrial de la modalidad presencial desde el semestre 2014-1 hasta 2018-2 en la institución Politécnico Grancolombiano. Resulto muy interesante hacer uso de la técnica de árboles de clasificación con el objetivo de encontrar las variables independientes que se asocien al bajo rendimiento de los estudiantes. La validación de este modelo se hace con una regresión logística, tomando únicamente las variables significativas resultantes del método árbol de clasificación y así viendo el grado de relevancia que cada una le aportaba al modelo. Se concluye que la edad, la sede, la jornada, el tipo de convenio y el semestre, son las variables predictoras del modelo de análisis de riesgos académicos.

PALABRAS CLAVES: Riesgo y deserción académicos.

2. TITULO DE LA PROPUESTA:

ANÁLISIS DE RIESGOS ACADEMICOS EN EL PROGRAMA DE INGENIERIA INDUSTRIAL DEL POLITÉCNICO GRANCOLOMBIANO POR MEDIO DE UN MODELO ESTADÍSTICO.

3. PLANTEAMIENTO DEL PROBLEMA:

Actualmente en las instituciones educativas existen diversos problemas que afectan el desarrollo de la educación, una de las principales dificultades a las que se enfrenta el sistema educativo nacional, es la deserción de los alumnos del programa de ingeniería industrial durante el transcurso de sus estudios, a causa de factores que ponen en riesgo alto de abandono a cada persona en su periodo estudiantil. La deserción parece ser una enfermedad educativa sus efectos no se detienen y cada vez va creciendo y trayendo efectos nocivos para el sistema pedagógico en un país como Colombia.

El rendimiento académico genera una estructura esencial dentro de la educación superior, lo cual para los estudiantes es difícil de lograr; existen múltiples estudios que a través de la aplicación de técnicas de minería de datos, métodos de supervivencia, métodos de clasificación, ecuaciones logísticas, entre otros, buscan realizar predicciones del rendimiento académico de los estudiantes, puesto que, el historial depende de factores personales, laborales, económicos, sociales, psicológicos; etc.

Así mismo, es importante para la institución educativa Politécnico Grancolombiano, que el nivel de éxito alcanzado por los estudiantes sea un reflejo de calidad de la enseñanza que ofrece su grupo de docentes; y por ello, se encuentra en total disposición para tomar medidas en cuanto a los estudiantes que se encuentran en alto riesgo de perder sus materias y de tomar la decisión de abandonar sus estudios.

La institución demuestra bastante interés en hacer este tipo de análisis, puesto que el objetivo es identificar las personas con mayor riesgo de deserción, teniendo en cuenta su

historial académico y factores de impacto, para poder tomar decisiones y bajar la tasa de abandono estudiantil. Este análisis se hará por medio de técnicas analíticas y estadísticas.

La investigación se divide en las siguientes fases para lograr el objetivo esperado:

- Determinar los factores de impacto que pueden afectar el historial académico de los estudiantes desde el periodo 2014-1 hasta el periodo 2018-2.

- Demarcar estadígrafos que puedan aplicarse para medir el rendimiento académico en estudiantes de la universidad Politécnico Grancolombiano.

- Extracción y consolidación de bases datos de las diferentes fuentes de información.

- Determinar un modelo algorítmico para medir el rendimiento académico de los estudiantes.

- Definir que aplicativo computacional que pueda ser herramienta para medir el rendimiento académico por cohortes de los estudiantes de ingeniería industrial.

4. OBJETIVOS:

4.1 OBJETIVO GENERAL:

Diseñar un modelo que mida el nivel de riesgo académico de los estudiantes de Ingeniería Industrial en modalidad presencial, con base en los datos del semestre 2018-2, con el fin de promover la permanencia exitosa de los estudiantes en el Politécnico Grancolombiano.

4.2 OBJETIVOS ESPECIFICOS:

- Seleccionar una metodología que permita guiar el proyecto de minería de datos.

- Diseñar estructura en donde se almacenarán y unificarán los datos recolectados.

- Cuantificar e Identificar los diferentes factores personales y académicos que afectan el rendimiento universitario, para analizar, definir y consolidar los más relevantes.

- Especificar algoritmos que permitan realizar la detección de factores afectantes en el buen rendimiento académico.

5. JUSTIFICACIÓN:

Una de las principales prioridades por parte del estado colombiano, ha sido garantizar la calidad de la educación en todos sus niveles, sin embargo, los esfuerzos realizados por cumplir dicho propósito no han sido suficientes ya que los indicadores internacionales muestran que Colombia tiene puntajes bajos en comparación a países que se encuentran en procesos de desarrollo de similares a este. (Sánchez, Roberto Mauricio (2016).

Según (Diaz & Garzón), desde la década de 1990, la deserción ha mostrado un aumento gradual y constante, sin diferencias significativas entre universidades públicas y privadas o personas con pocos o muchos ingresos; es una característica estructural del sistema universitario colombiano y centro de preocupación para las universidades y el Estado.

Esta problemática abarca una gran variedad de factores como dificultades financieras, poca preparación escolar, la carrera no convence al estudiante, falta de apoyo, etc. Otra de las circunstancias que se observan en los estudiantes, es que, al graduarse no están preparados para enfrentar las demandas del mercado laboral. Lo que preocupa es que se calcula que solo el 50% de los estudiantes que inician sus estudios superiores llegan a terminar y se gradúan. (Dinero, 2017).

Las escuelas de ingeniería en todo el mundo tienen una tasa de deserción relativamente alta. En Colombia uno de cada dos estudiantes no logra graduarse y el 72% de los estudiantes abandona la educación superior entre el primero y el cuarto semestre académico. (Portafolio, 2017). Por lo general, alrededor del 35% de los estudiantes de primer año en diversos programas de ingeniería no llegan al segundo, de los estudiantes restantes, con frecuencia abandonan o fallan en su segundo o tercer año de estudios.

En la institución la tasa de deserción en el periodo 2018-2 fue del 23%, aproximadamente 11.191 estudiantes desertaron en ese periodo en la universidad, por otro lado, en la facultad de ingeniería diseño e innovación la tasa de deserción fue del 23% con una deserción de 2.185 estudiantes en la facultad y en el programa de ingeniería industrial la tasa de deserción fue del 24%, abandonando sus estudios un total de 867 estudiantes. (Inteligencia Competitiva - Politécnico Grancolombiano, 2018)

Es interesante hacer este tipo de análisis en el programa académico de Ingeniería Industrial, según (Universa, 2019), el programa está dentro de las veinte carreras con mayor demanda y mejores pagas en el país, figura una tasa de empleabilidad del 85,6%. Es importante para la universidad Politécnico Grancolombiano llevar un seguimiento sobre las notas y rendimiento académico de sus alumnos de este programa, ya que la demanda laboral para estos profesionales es grande, y quiere cerciorarse que sus egresados sean los mejores en el campo profesional.

La investigación se enfoca en el área de Inteligencia competitiva del Politécnico Grancolombiano, la cual, es la encargada de recolectar, almacenar, analizar y mostrar con métodos estadísticos los datos generados por la operación de la universidad. En ese orden de ideas, el motivo de este trabajo es brindar a la Institución Universitaria herramientas de seguimiento y control de los riegos académicos que corren los estudiantes por cada cohorte en su periodo estudiantil.

6. MARCO TEORICO:

Frente a la situación que actualmente afronta el programa de ingeniería industrial de la institución, hay diferentes bases de datos de los cuales se puede hacer la extracción y recolección de información necesaria para desarrollar este proyecto.

La minería de datos se define como el proceso de extracción de información implícita u oculta, no plenamente explotada y en la mayoría de los casos no conocida previamente, a partir de datos disponibles en repositorios (bases de datos) o en tiempo real. Con este proceso se pretende generar conocimiento útil a partir del establecimiento de relaciones entre los datos, tales como patrones o asociaciones. (Vélez Garcia, 2018).

En resumen, los procesos de minería de datos transforman grandes conjuntos de datos que sean de utilidad para la toma de decisiones o para comprender mejor el problema en concreto. No existe un modelo único para el desarrollo de un proyecto de minería de datos. (Vélez Garcia, 2018).

Existen varias técnicas de minería de datos, en la imagen número 1 se presenta el resultado de encuestas realizadas en el 2007 y el 2014 por: (Piatesky, 2014)

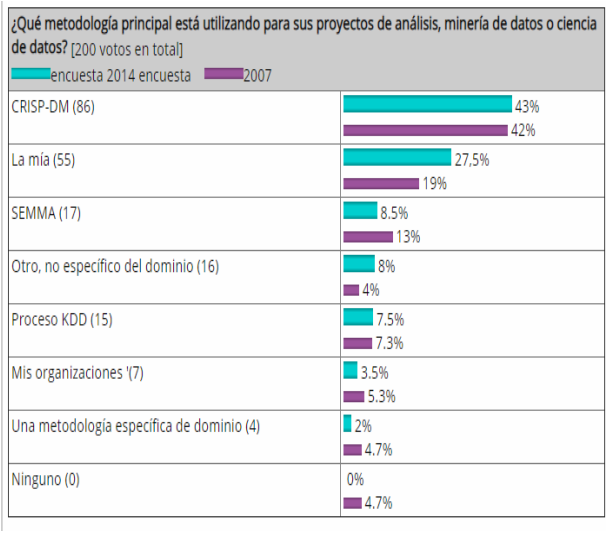

Ilustración 1-Metodología principal CRISP-DM (Piatesky, 2014)

En la imagen número 1 se observa que la metodología que continúa siendo más utilizada es la de **CRISP-DM**, y es la escogida para el desarrollo de este proyecto, a continuación, se describe esta técnica y cada una de sus fases.

METODOLOGÍA CRIPS-DM:

Es una de las guías de referencia más utilizadas en proyectos de minería de datos. Incluye descripciones de las fases de un proyecto y las tareas a realizar para cada una de estas, por otro lado, como modelo de proceso ofrece un resumen del ciclo vital de minería de datos. A continuación, se describe cada una de las fases en las que se divide el ciclo vital de la metodología CRISP-DM (Fernandes, 2017):

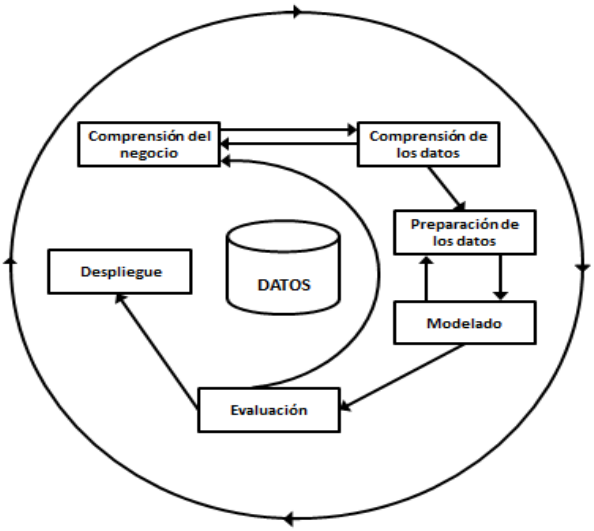

Ilustración 2.-Metodología CRISP-DM (Peralta, 2014)

- **Fase de comprensión del negocio o problema:**

La primera fase de la metodología se denomina comprensión del negocio o problema, es la fase con mayor importancia ya que en ella se especifican los requisitos y se define el plan a seguir durante las demás fases del proyecto. En esta fase se trata de comprender el problema desde un punto de vista empresarial o institucional para seleccionar la información y los datos que permitan resolverlo. Para la realización de esta fase, CRISP-DM propone seguir las siguientes tareas: determinar el objetivo del negocio, valorar la situación, determinar los objetivos de la minería de datos y por último realizar un plan de proyecto.

- **Fase de comprensión de datos:**

La segunda fase se conoce como la comprensión de los datos. En esta fase se recolectan los primeros datos a analizar, con el fin de tener un primer contacto real con el problema planteado en la primera fase y familiarizarse con la información obtenida. En primer lugar, se debe identificar la calidad de los datos y definir una primera hipótesis en base a éstos. Las principales tareas propuestas por la guía en esta fase son: la recolección de datos iniciales, la descripción de los datos, la exploración de los datos y la verificación de la calidad de los datos.

- **Fase de preparación de los datos:**

Con los datos ya recopilados, se procede a su preparación para ajustarlos a la técnica de minería de datos que se desea utilizar y construir un conjunto de datos que se ajuste al problema. Esta fase incluye las tareas de selección de datos a los que se aplicará una técnica de modelado, y la limpieza y generación de datos variables adicionales. Las tareas para realizar en esta fase son: la selección de datos, limpieza de datos, la estructuración de los datos, la integración de los datos y el formateo de los datos.

- **Fase de modelado:**

Es el proyecto de minería de datos. Debido a que hay varias técnicas que requieren un formato específico de datos, en muchos casos es necesario volver a la fase de preparación de datos. Las técnicas para utilizar en esta fase se eligen basándose, entre otros, en los siguientes criterios:

1. Ser apropiado al problema.

2. Cumplir con los requisitos del problema.
3. Disponer de datos adecuados.

Las tareas propuestas por la guía son las siguientes: selección de la técnica de modelado, generación de un plan de prueba, construcción del modelo, y evaluación del modelo.

- **Fase de evaluación:**

En esta fase se evalúa el modelo seleccionado, y se verifica el cumplimiento de los criterios de éxito establecidos como objetivo del estudio. Se revisa el proceso y los resultados obtenidos, por si fuera necesario repetir algún paso previo para corregir algún error cometido anteriormente. Si finalmente se concluye que el modelo seleccionado es válido para el estudio, y cumple con los criterios de éxito, se procede a la explotación del modelo. Las tareas que componen esta fase son las 5 siguientes: evaluación de los resultados, revisión del proceso y determinar próximos pasos.

- **Fase de distribución:**

Una vez que el modelo ha sido seleccionado, construido y validado, en esta fase transformaremos el conocimiento y resultados obtenidos en acciones para el negocio. Identificaremos las acciones a realizar por la empresa o cliente, basándonos en los resultados obtenidos a partir del modelo. En esta fase también se documentarán los resultados obtenidos de manera comprensible para el usuario. Las tareas que se llevan a cabo en esta fase son: plan de implantación, plan de monitoreo y mantenimiento, realización del informe final y revisión del proyecto.

7. ESTADO DEL ARTE:

El siguiente es el contexto de investigación e indagación de las diferentes fuentes proporcionadas por la universidad en este proyecto. Se enfatizo en las fuentes SCOPUS y GOOGLE ACADÉMICO, con el fin de realizar una retroalimentación de documentos nacionales e internacionales para identificar y comparar los resultados obtenidos.

La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar amplia cantidad de información. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha trasformado el análisis de los datos orientándolos hacia determinadas técnicas especializadas englobadas bajos el nombre de minería de datos o Data Minning. (Rodriguez Flores , Flores Pulido, & Dávila de la Rosa, 2016). La minería de datos en la educación no es un tópico nuevo, su estudio y aplicación ha sido muy relevante durante los últimos años. El uso de estas técnicas permite predecir cualquier fenómeno dentro del ámbito educativo. (Timarán Pereira & Jiménez Toledo, 2014). De esta forma, utilizando las técnicas de minería de datos, se logra el objetivo de predecir los factores más relevantes para que un estudiante tome la decisión de desertar.

Por otro lado, se propone, la construcción de un modelo predictivo de rendimiento académico, con datos de 932 estudiantes, en dos etapas: La primera, se refiere a la extracción y preparación de la información a través de la limpieza y estructuración de datos. La segunda, constituye la aplicación de minería de datos, por medio de actividades de carga y procesamiento, así también la formulación del modelo e interpretación de resultados (Mérchan Rubiano & Duarte Garcia, 2016). Para este proyecto se utiliza la

técnica CRISP-DM, la cual es un proceso que se encuentra dividido en seis fases que son de gran importancia para realizar minería de datos, siendo el método más utilizado en la última década. La minería de datos se define como se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. (Galán Cotina, 2015).

Se realizó un modelo analítico para la predicción de rendimiento académico en estudiantes de ingeniería, en el año 2015 en la ciudad Santiago de Chile, este estudio tiene como objetivo mostrar herramientas de Learning Analytics que puedan ser usadas para generar modelos predictivos que sirvan para apoyar a aquellos estudiantes con riesgo de deserción o insuficientes desempeños académicos. Se demuestra que las técnicas de minería de datos son capaces demostrar aprendizaje momento a momento para estos estudiantes que se encuentren en riesgo. (Celis, Moreno, Poblete, Villanueva, & Weber, 2015).

En una universidad privada de Perú, se realizó un artículo presentando los resultados obtenidos de los diseños y aplicaciones de siete modelos predictivos correspondientes al mismo número de cursos, denominados por la universidad como críticos por alto índice de reprobación. (Sifuentes Bitocchi, 2018). El caso de estudio de este proyecto es el programa de ingeniería industrial y se clasificaron 35 materias, las cuales son las más enfocadas al programa, para obtener la nota parcial y la definitiva de cada una de ellas.

El estudio comparativo de algoritmos para predecir la deserción académica utiliza información personal y académica de los estudiantes. Los autores (Hernández Gonzales , Meléndez Armenta , & Morales Rosales, 2016), desarrollaron un sistema predictivo para detectar al alumno con probabilidad de deserción. Utilizando la herramienta, Microsoft SQL Server Analysis Services, se crea un modelo de predicción de atributos discretos y continuos, utilizando un algoritmo de clasificación y regresión, para encontrar los estudiantes con elevado porcentaje de deserción mediante un árbol de decisión. De esta manera se logró realizar la extracción de datos necesarios para la generación del modelo estadístico en este proyecto.

Se realizo el artículo, que recibe por nombre "Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°", en el cual se presentan los resultados obtenidos el modelo de clasificación basado en árboles de decisión, con el fin de detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media, que presentaron las pruebas Saber 11° en los años 2015 y 2016. La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. (Pereira Timarán , Zambrano Caicedo, & Troya Hidalgo, 2018).

Según (Maya Lopera, 2018), existen diferentes tipos de árboles, que son catalogados según la situación y el resultado deseado, él los cataloga en los siguientes tipos:

1. Árbol de clasificación o binario: se usa cuando hay varias alternativas que se han calculado anteriormente para obtener resultados más predecibles, para hacer uso de estos, hay que trazar esquemas binarios y proyectar las diferentes variables o ramas del árbol, gracias a las probabilidades de estas se puede predecir un poco el resultado.

2. Árbol de regresión: este tipo de árbol ayuda a determinar un único resultado ya que se tiene la información necesaria para identificar una "ruta" optima. Cuando se construye este árbol se divide la información en secciones o subgrupos. Este tipo de árbol es usado en bienes raíces.

3. Árbol de mejora: es usado cuando se desea tener un resultado más preciso; el árbol se construye, luego se toma una variable, esta se calcula y se estructura para reducir la incertidumbre y los errores a la hora de tomar decisiones. Este árbol es usado en contabilidad y matemática.

4. Árboles mixtos entre regresión y clasificación: es usado para prever un resultado con variables impredecibles, generalmente se usan indicadores 16 que muestren lo que ya ha sucedido en un pasado (Sesgando un poco el resultado), Se usa generalmente en ciencia.

5. Arboles de binomiales y tri-nomiales: son árboles donde cada rama es en sí un modelo binomial su función es recrear el modelo binomial varias veces en el tiempo suponiendo que el precio o costo de un elemento sube o baja con el tiempo. Estos árboles con muy usados en la aplicación de opciones reales.

La principal justificación para la aplicación de esta técnica es que la misma ha demostrado su utilidad descriptiva, exploratoria y explicativa, así como su aplicación para la clasificación de datos, en el ámbito del rendimiento académico. En este proyecto se aplicará la del árbol de clasificación o binario, ya que la variable dependiente es el riesgo "Alto" o "Bajo" que un estudiante tiene al perder una materia.

En la universidad de Barcelona, se realizó un artículo de la aplicación de árboles de clasificación en la plataforma SPSS, los autores explican los usos generales de este método y el procedimiento paso por paso para su desarrollarlo e interpretación, el cual se tomó como guía para la elaboración del modelo en este proyecto. (Berlanga Silvente, Rubio Hurtado, & Rith Vilá , 2014).

En la investigación con nombre "Modelo de regresión logística como alternativa para medir la probabilidad de deserción temprana en la universidad de los llanos periodo 2015-2 – 2018-1", realizada por (Arismendy Fuentes & Morales Parrado, 2018), usan la regresión logística para dicha predicción, en donde las variables con mayor significancia en el modelo son el puntaje de la prueba saber 11, género, haber cursado estudios antes de ingresar a la universidad, no haber reprobado años en el bachillerato y si los padres conviven juntos, disminuye la probabilidad de deserción temprana, por otro lado si el estudiante es egresado de un colegio privado aumenta la probabilidad de deserción. Así mismo, de acuerdo con las pruebas de bondad de ajuste, se considera que el modelo se adapta totalmente a los datos observados.

En la universidad de Atacama, Chile, se realizó un análisis de deserción académica donde los propósitos son determinar los factores asociados al buen desempeño de los alumnos en el primer año lectivo 2014 de las carreras de ingeniería de la universidad, e identificar aquellas variables determinantes a la deserción académica. En la segunda parte utilizaron un modelo de Regresión Logística en el cual descubrieron que las variables asociadas a la deserción tienen que ver con las características de inscripción a la universidad, créditos

inscritos y rendimiento académico. (Barahona Urbina, Aliaga Prieto , & Veres Ferrer, 2016)

Se realiza un análisis de predicción del rendimiento académico de los alumnos ingresantes a la Facultad de Ciencias Exactas y Naturales y Agrimensura de la Universidad Nacional del Nordeste (FACENA-UNNE) en la ciudad de Corrientes, Argentina, durante el primer año de carrera con las características socioeducativas de los estudiantes. El rendimiento fue medido por la aprobación de los exámenes parciales o finales de la primera materia de Matemática que los alumnos cursan. Se ajustó un modelo de regresión logística binaria, el cual clasificó adecuadamente el 75% de los datos. Entre las variables más relevantes para explicar el rendimiento académico se encuentran el título secundario obtenido, la carrera elegida y el nivel educacional alcanzado por la madre. (Porcel, Dapazo, & López, 2019).

8. METODOLOGÍA:

8.1 Fase de comprensión del problema:

Este proyecto se desarrolla en la Institución Universitaria Politécnico Grancolombiano, ofrece dos modalidades de estudio: presencial y virtual. Esta institución ha tenido un promedio de 2.182 estudiantes desde el año 2014 hasta el 2018 en el programa de ingeniería industrial de la modalidad presencial.

Se pretende detectar las causas que estén provocando que los estudiantes pierdan materias en el programa de ingeniería industrial, en la modalidad presencial, para promover la permanencia de los estudiantes en la universidad.

Los efectos que causan la reprobación de materias y que preocupan a la institución son:

1. Retrasos en el avance normal de la titulación, por lo tanto, el tiempo para graduarse aumenta.
2. Deserción de la institución o programa
3. Para la universidad la afectación se ve en los ingresos, puesto que es más rentable mantener a un estudiante que no deserte durante su carrera que traer un estudiante nuevo para ocupar el espacio de otro que ya deserto.
4. Disminuye la tasa de graduación para los primíparos.

- **Recursos para el proyecto:** A continuación, se presenta un inventario de los recursos a utilizar en la investigación y desarrollo del proyecto:

RECURSO	DESCRIPCIÓN
Fuente 1: Base de datos de matriculados.	De esta fuente se obtendrán datos básicos de los estudiantes del programa de ingeniería industrial desde el 2014 hasta 2018
Fuente 2: Base de datos Insumos	De esta fuente se obtendrá datos referentes al historial académico del estudiante, materias y profesor.
Fuente 3: Base de datos Notas Estudiantes	De esta fuente se obtiene el promedio de las calificaciones de los estudiantes por cada semestre.

Fuente 4: Base de datos unificada y centralizados.	Esta fuente será el repositorio de la unión de todos los datos antes mencionados.
Herramientas para realizar limpieza y transformación de los datos	Se usará la herramienta Microsoft Excel
Herramientas para aplicar algoritmos de minería de datos que permitan segmentación, asociación, clasificación, etc., y así como validar los resultados.	Se utilizará la herramienta Rstudio.
Computador	En este equipo se hará todo el proceso de minería de datos.

Tabla 1.- Inventario de recursos (Elaboración propia)

- **Restricciones**: Se determinaron restricciones de seguridad y legales, las cuales se representan en la siguiente tabla.

RESTRICCIONES	
Seguridad	**Legales**
1. Acceso a las fuentes: Para la recolección de los datos se tendrá acceso a la lectura de bases de datos transaccional y a la base de datos centralizada. **2. Niveles de seguridad:** Se darán acceso a los resultados por el nivel de información a la que el visualizador final deba tener acceso.	**1. Privacidad de los datos:** Se debe tener en cuenta un responsable tratamiento de los datos personales de los estudiantes de la institución. No se deben hacer públicos los resultados ni listados de los estudiantes. **2. Uso de los datos:** Solo se pueden utilizar los datos para los fines definidos y autorizados por la institución.

Tabla 2. Restricciones (Elaboración propia)

8.2 Fase de comprensión de los datos:

La segunda fase que propone la técnica CRISP-DM es la de recolección, definición y comprensión de los datos.

- **Recolección de los datos:**

En la ilustración No 1 se indica el grupo de datos a recolectar y las fuentes de datos de donde fue extraída la información:

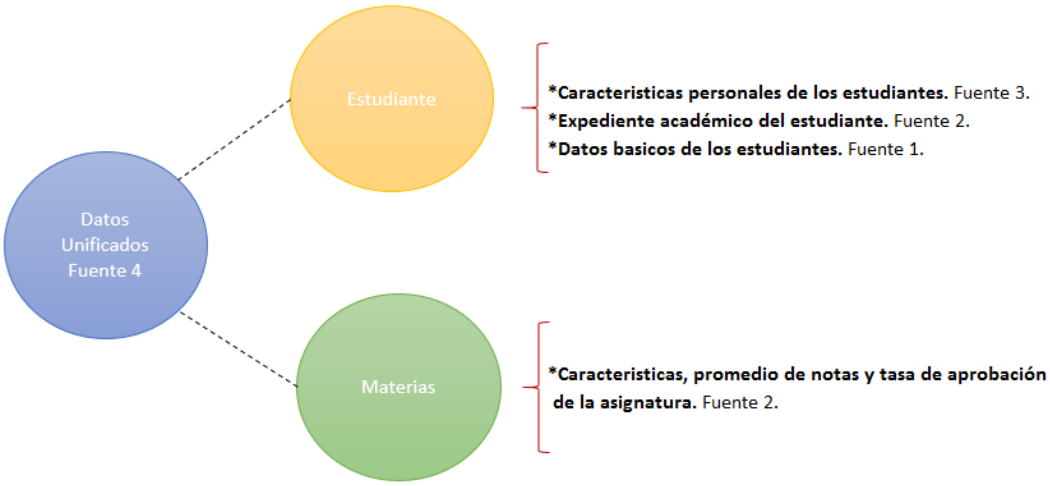

Ilustración 3.-Grupo de datos a recolectar (Elaboración propia)

- **Descripción de los datos:** En este punto se presenta una visión inicial del tipo de datos encontrados y también de las variables que existen en las bases de datos, al unificar los datos en la Fuente 4, se descubren 9 variables a las cuales se presentaran a continuación:

Variable	Tipos	Descripción	Base de datos
Nota	Definitiva	Nota final obtenida en el semestre.	Notas parciales ingeniería industrial
	Parciales	Nota de cada uno de los cursos vistos en el periodo académico	
Jornada	Diurno	Tiempo diario que dedica el establecimiento educativo desde las 6:40 pm hasta las 9:50 pm.	Matriculados
	Nocturno	Tiempo diario que dedica el establecimiento educativo desde las 7:00 am hasta las 6:00.	
Primíparo	Si	Estudiante que ingresa por primera vez a la universidad, facultad, jornada o programa.	Matriculados
	No	Estudiante antiguo en la universidad, facultad, jornada, y programa.	
Edad		El rango de edades que hay en los estudiantes de ingeniería industrial de la jornada presencial es de 17 a 54 años	Caracterización del estudiante
Género	Femenino	Define a la mujer	Caracterización del estudiante
	Masculino	Define al hombre	
Estado civil	Soltero	Individuo que no contrajo matrimonio ni tiene vinculo sentimental estable	Caracterización del estudiante
	Casado	Persona que ha contraído matrimonio	
	Divorciado	Persona cuyo matrimonio se ha disuelto	
	Viudo	Persona cuyo cónyuge ha fallecido	

	Unión libre	Vinculo sentimental sin ningún tipo de contrato matrimonial	
Tipo de convenio	Sena	Estudiantes que realizaron estudios en el SERVICIO NACIONAL DE APRENDIZAJE (SENA), y su título es homologado.	Notas parciales ingeniería industrial
	Regular	Estudiantes que no tiene ningún tipo de título de posgrado en otra institución.	
Semestre	1	Número de periodos académicos que se deben cursar en el ciclo del programa de ingeniería industrial	Notas parciales ingeniería industrial
	2		
	3		
	4		
	5		
	6		
	7		
	8		
	9		
Sede	ANTIOQUIA - MEDELLIN CAMPUS BARRIO LOS COLORES (PG)	Son las dos sedes que pertenecen a modalidad presencial de la institución.	Matriculados
	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)		

Tabla 3.-Definición de variables (Elaboración propia)

- **Exploración de los datos:**
  En esta fase se hace un análisis descriptivo por cada una de las variables encontradas y en algunos casos se hará dicha distribución por periodo.

  1. **Distribución de las notas:** Se realiza una tabla de frecuencia del promedio de notas por periodo, en un rango de 1 a 5, se cuenta con un total de 7.259 registros desde el periodo 2014-1 hasta el 2018-2. La variable de distribución de notas hace parte de un factor académico del estudiante.

Rango Notas	*Frecuencia*	*% acumulado*
0.0 – 1.0	33	0.45%
1.1 – 2.0	190	3.07%
2.1 – 3.0	1386	22.17%
3.1 – 4.0	4200	80.04%
4.1 – 5.0	1449	100.00%
y mayor...	0	100.00%

Tabla 4.-Frecuencia de notas (Elaboración propia).

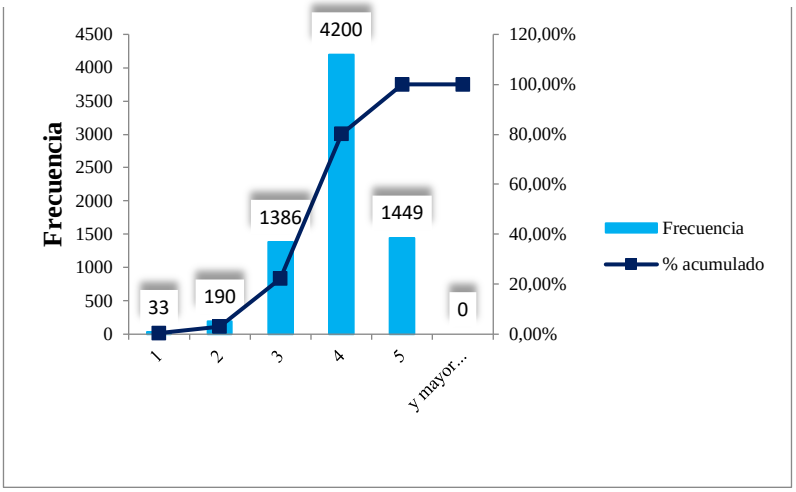

Gráfico 1.-Gráfico de distribución de las notas (Elaboración propia)

Se evidencia que la mayor parte de las personas obtienen como calificaciones notas entre 4 y 5.

2. **Distribución del género:** A continuación, se presenta una tabla y grafico de distribución del género durante los periodos 2014 -1 a 2018 – 2. La variable de distribución de género es un factor personal del estudiante.

PERIODO	FEMENINO	MASCULINO
20141	380	593
20142	231	353
20151	280	450
20152	198	343
20161	320	464
20162	330	468
20171	358	493
20172	375	454
20181	246	351
20182	228	344
Total, general	**2946**	**4313**

Tabla 5.-Tabla de distribución del género (Elaboración propia)

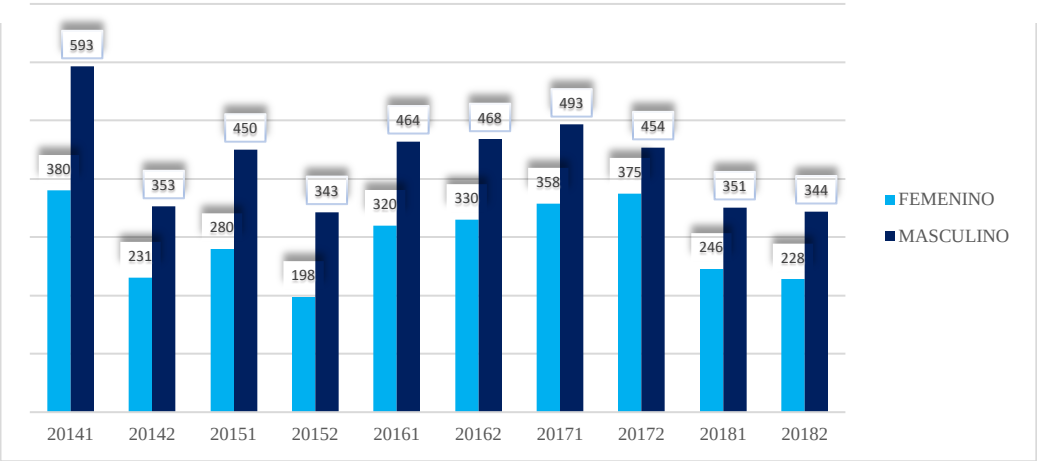

Gráfico 2.-Gráfico de distribución del género (Elaboración propia)

3. **Distribución del estado civil: S**e presenta una tabla y grafico de distribución del Estado civil de los estudiantes durante los periodos 2014 -1 a 2018 – 2. El factor de estado civil del individuo es de información personal.

PERIODO	CASADO	DIVORCIADO	SOLTERO	UNION LIBRE
20141	64	2	880	27
20142	39		530	15
20151	38	1	663	28
20152	29	1	502	9
20161	32		719	33
20162	40		725	33
20171	25	1	798	27
20172	42	5	751	31
20181	26	1	549	21
20182	23	2	517	30
Total, general	**358**	**13**	**6634**	**254**

Tabla 6.-Tabla de distribución de estado civil (Elaboración propia)

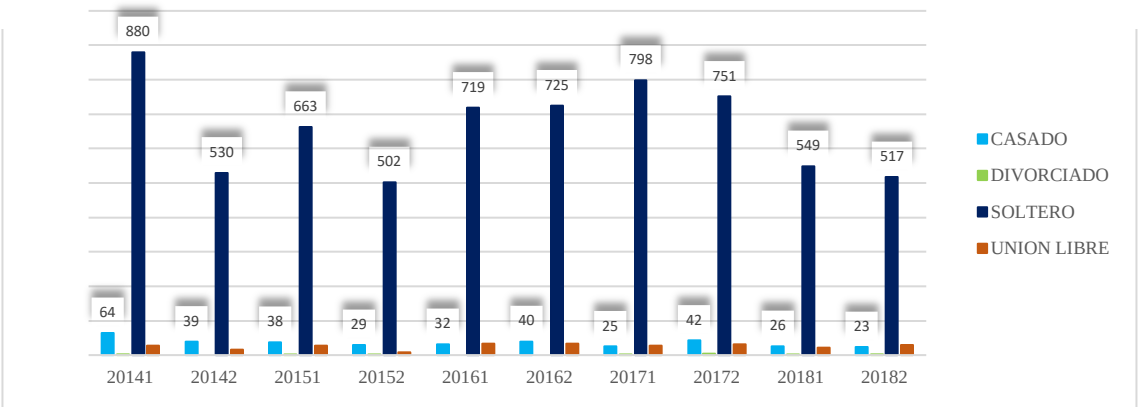

Gráfico 3.-. Gráfico de distribución de estado civil (Elaboración propia)

4. **Distribución Primíparo:** A continuación, se presenta una tabla y grafico de distribución de la variable primíparo durante los periodos 2014 -1 a 2018 – 2. El factor primíparo hace parte de la información académica del estudiante.

PERIODO	ANTIGUO	PRIMIPARO
20141	834	139
20142	536	48
20151	643	87
20152	477	64
20161	643	141
20162	693	105
20171	724	127
20172	696	133
20181	489	108
20182	463	109
Total, general	**6198**	**1061**

Tabla 7.-Tabla de distribución de primíparos (Elaboración propia)

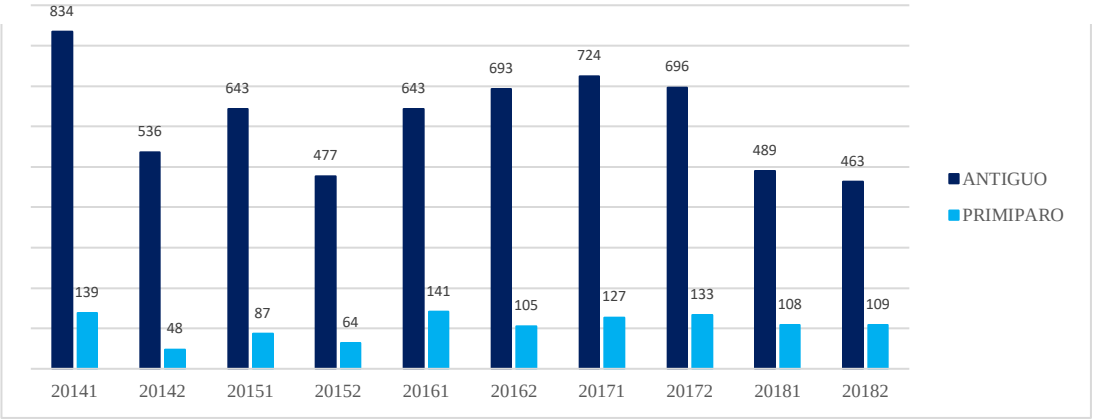

Gráfico 4.-Gráfico de distribución de primíparo - antiguo (Elaboración propia)

5. **Distribución de jornada:** A continuación, se presenta una tabla y grafico de distribución de la variable jornada durante los periodos 2014 -1 a 2018 – 2. La variable jornada es un factor de tipo académico del estudiante.

Etiquetas de fila	DIURNO	NOCTURNO
20141	169	804
20142	65	519
20151	125	605
20152	68	473
20161	61	723
20162	71	727
20171	97	754
20172	134	695
20181	103	494

20182	75	497
Total, general	**968**	**6291**

Tabla 8.-Tabla de distribución de jornada (Elaboración propia)

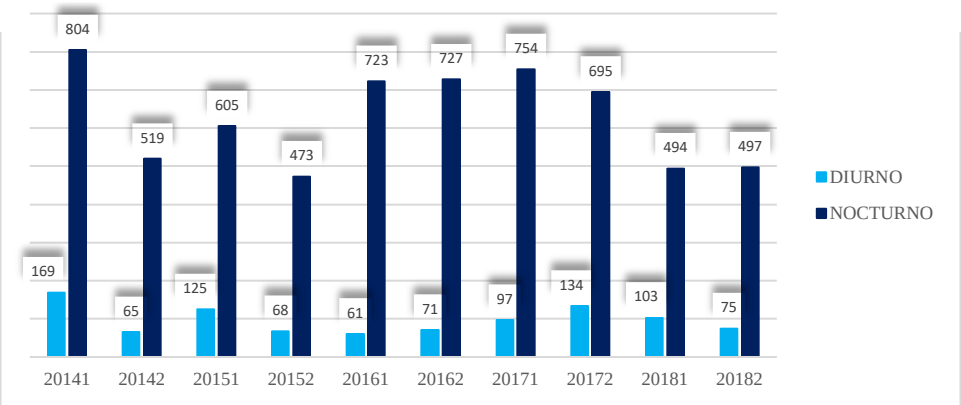

Gráfico 5.-Gráfico de distribución de jornada (Elaboración propia)

6. **Distribución convenio tipo:** A continuación, se presenta una tabla y grafico de distribución de la variable convenio durante los periodos 2014 -1 a 2018 – 2. El factor de convenio_tipo hace parte de los factores académicos del estudiante, puesto que se clasifica en dos categorías muy importantes, como son el convenio Sena y el convenio regular.

PERIODO	Regular	Sena
20141	695	278
20142	511	73
20151	495	235
20152	527	14
20161	474	310
20162	467	331
20171	490	361
20172	469	360
20181	310	287
20182	261	311
Total, general	**4699**	**2560**

Tabla 9.-Tabla de distribución de convenio (Elaboración propia)

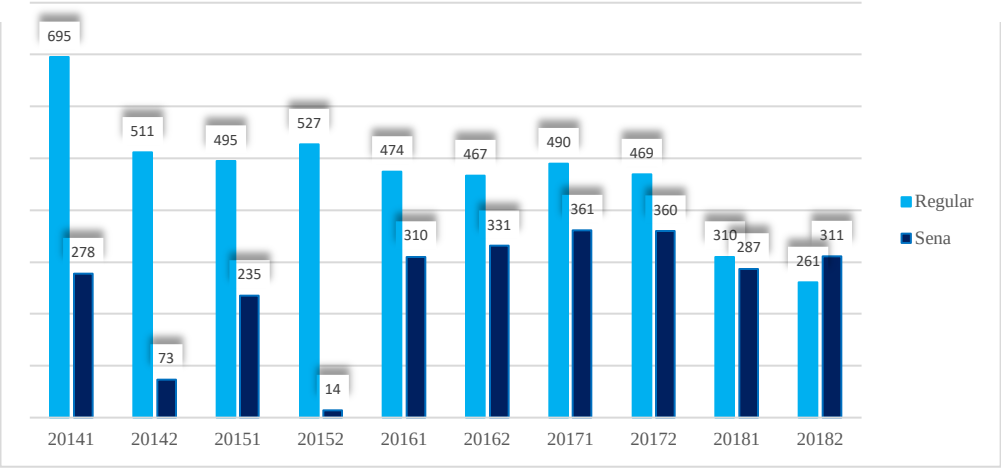

Gráfico 6.-Gráfico de distribución convenio tipo (Elaboración propia)

7. **Distribución sede:** A continuación, se presenta una tabla y grafico de distribución de la variable sede durante los periodos 2014 -1 a 2018 – 2.

PERIODO	ANTIOQUIA	BOGOTA DC
20141	33	940
20142	2	582
20151	67	663
20152	31	510
20161	155	629
20162	147	651
20171	172	679
20172	191	638
20181	146	451
20182	163	409
Total, general	**1107**	**6152**

Tabla 10.-Tabla de distribución de convenio (Elaboración propia)

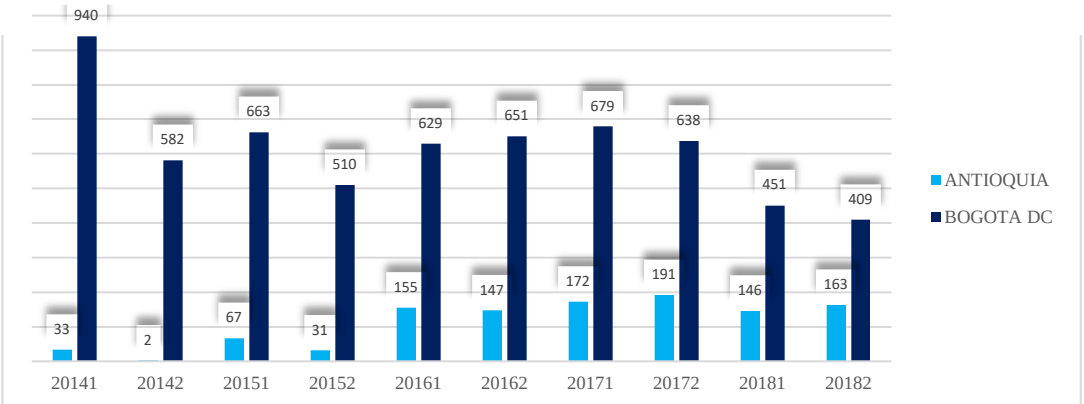

Gráfico 7.- Grafico de distribución sede (Elaboración propia)

Verificación de la calidad de los datos:

En este paso se realiza la tarea de revisar minuciosamente la información y así verificar los campos que estén vacíos o que tienen información incorrecta para realizar su debida corrección manual, con la información que proporciona la plataforma Analitycs.

8.3. Fase de preparación de los datos:

Luego de tener los datos y obtener un conocimiento previo de su contenido, es necesario definir una preparación de los datos antes de aplicar las técnicas de minería de datos, lo que beneficiara los resultados del modelado que depende directamente de la calidad de los datos que se utilicen. Esta fase incluye integración, limpieza y transformación de datos.

- **Limpieza de datos:**

  La base de datos solo suministraba la fecha de nacimiento de los estudiantes, así que en otra columna se aplicó la formula "AÑO (HOY ()-Fecha de nacimiento)-1900)", para obtener la edad del estudiante. La columna con la información de la edad del estudiante recibió el nombre de "Edad". En el campo de edad se identifican personas con edades fuera del rango normal (17- 54), en este caso al ser pocos los estudiantes, se hace una corrección manual de la fecha de nacimiento utilizado los datos del expediente de cada estudiante suministrado por la plataforma Analytics.

- **Integración de datos:**

  En este paso se selecciona la información necesaria y se construye una base compacta que contendrán los registros con las variables a utilizar y los atributos finales para hacer el análisis del proyecto.

- **Transformación de datos:**

  En esta sección se realiza toma la variable "Materia" y se selecciona las materias que más influyen en el programa, quedando así únicamente las notas de 35 materias para realizar el análisis, se organizó la base de datos quedando únicamente 7.529 registros únicos, y esa vendría a ser definida como la fuente 4.

  Se categorizo un nivel de riesgo en la variable explicativa, dividiendo el número de materias perdidas entre el número de materias inscritas por cada estudiante, y así clasificándose en riesgo "Bajo" y "Alto" en donde las personas que tuvieran una proporción mayor al 60% tendrían un riesgo alto de tomar la decisión de desertar.

8.4. Fase de modelado:

Árbol de clasificación:

Para la creación del árbol de clasificación en el programa SPSS se propone el caso práctico de identificar los factores más significativos a la hora de tener un riesgo alto o deserción de un estudiante. La variable que se quiere explicar, es decir, la variable dependiente del modelo es el "Riesgo" de que un estudiante de ingeniería industrial tome la decisión de desertar durante el periodo académico y como variables independientes se seleccionaron las siguientes: *Jornada, Primíparo, Edad, Género, Estado civil, Tipo de*

convenio, Semestre y Sede. Con esta información se realiza el árbol de clasificación seleccionando el método de crecimiento CHAID.

Ilustración 4.- Cuadro de variables (Elaboración propia)

Pulsando el botón de Resultados, se abre un cuadro de dialogo con pestañas, en el que se pueden seleccionar distintos tipos de opciones.

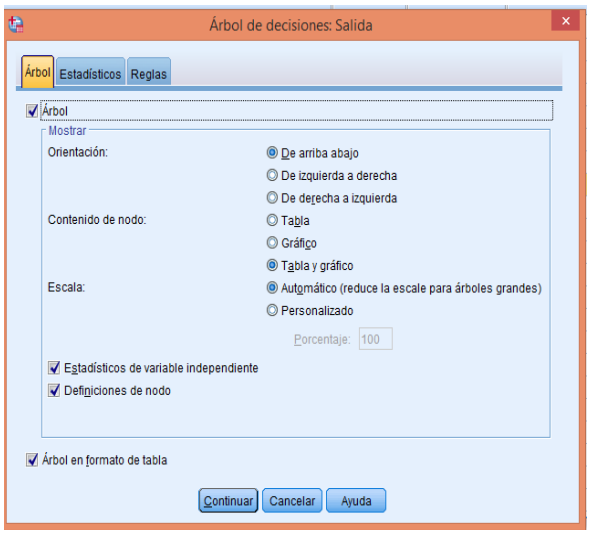

Ilustración 5.- Cuadro de Resultados (Elaboración propia)

La pestaña "Árbol" permite controlar el aspecto inicial del árbol o suprimir completamente su presentación.

En la pestaña "Estadísticos" las opciones disponibles dependen del nivel de medida de la variable dependiente, del método de crecimiento y de valores de configuración como: resumen, riesgo y tabla de clasificación.

La pestaña "Reglas" ofrece la capacidad de generar condiciones de selección o de clasificación, en forma de sintaxis de comandos o de solo texto.

En el botón de Validación del modelo, se puede evaluar la bondad de la estructura del árbol, cuando se generaliza para una mayor población. Existen dos métodos de validación

disponibles, la validación cruzada y la validación por división muestral. Para el desarrollo de este modelo se escogió la validación cruzada para dividir la muestra en un número de submuestras, y a continuación, se generan el modelo del árbol.

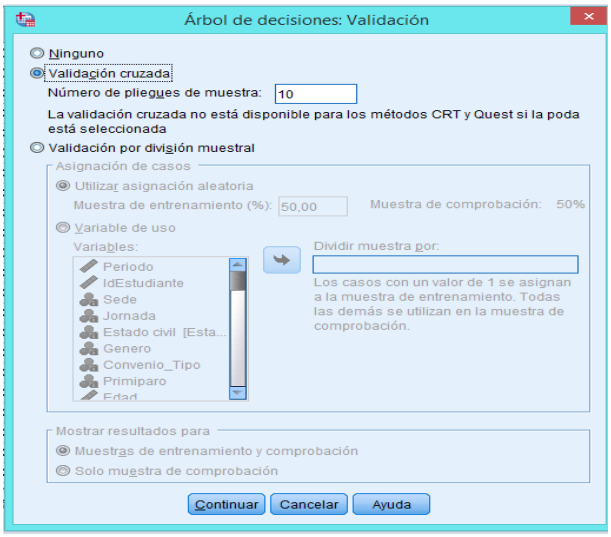

Ilustración 6.- Cuadro de validación (Elaboración propia)

El botón Criterios, permite establecer el dictamen de crecimiento del árbol, en este caso se opta por el más sencillo y dejaremos las filiales tal como aparecen por defecto del programa.

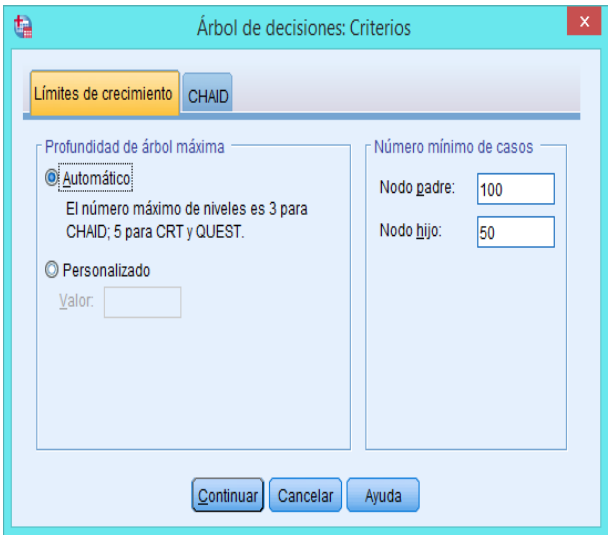

Ilustración 7.- Cuadro de criterios (Elaboración propia)

La pestaña “Limites de crecimiento” permite limitar el número de niveles del árbol y controlar el número de casos mínimo para nodos padres e hijo.

La pestaña “CHAID” puede controlarse el nivel de significancia para la división de los nodos y la fusión de categorías. Para ambos criterios el nivel de significancia por defecto es igual a 0.05.

El procedimiento excluirá de manera automática las variables que no generen ningún tipo de significancia, es decir, que no haya alguna contribución de parte de ellas al modelo.

Regresión logística:

Para la creación del modelo de regresión Logística en la herramienta SPSS también se tomó en cuenta a la variable dependiente como el riesgo que tiene un estudiante de desertar, clasificándose como riegos “Alto” y “Bajo” que será una categoría dicotómica codificada con valores 1 y 0, pero utilizando como variables independientes únicamente las variables que fueron significativas en el modelo de árbol de clasificación anteriormente explicado, estas variables son: *Edad, Jornada, Convenio, Semestre y Sede.*

Con la base de datos abierta, se activa la secuencia: Analizar > Regresión > Logística binaria y se mostrará el siguiente recuadro.

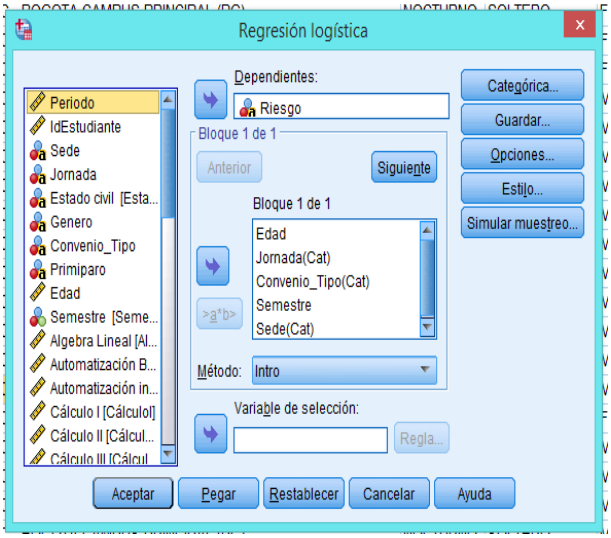

Ilustración 8.-Cuadro de variables (Elaboración propia)

Pulsando el botón de Categórica, se abre un recuadro en el cual se ve el tipo de variables que hay en el modelo de Regresión Logística, y el contraste de las categorías de referencia.

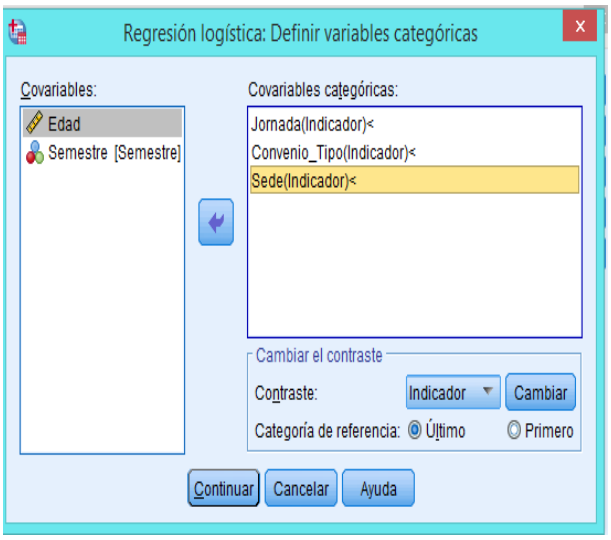

Ilustración 9.-Cuadro categóricas (Elaboración propia)

Para este modelo se dejó la configuración de la ventana en la categoría de referencia con el indicador “ultimo”, por defecto de la herramienta.

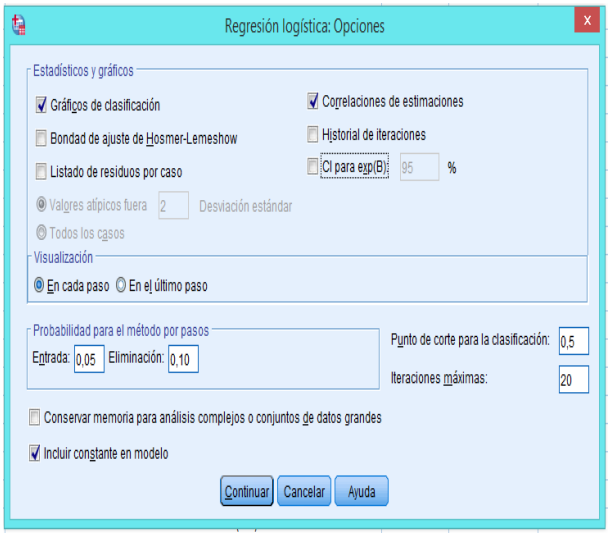

Ilustración 10.-Cuadro de Opciones (Elaboración propia)

En el botón Opciones, se abre un cuadro de dialogo en el cual se observan las distintas opciones seleccionadas para el modelo.

Gráficos de clasificación: Hace referencia a un histograma de los valores actuales y pronosticados por el modelo para la variable dependiente.

Correlaciones de estimaciones: Muestra la matriz de correlaciones de las estimaciones de los parámetros para los términos del modelo.

En la opción Probabilidad para los pasos, permite controlar los criterios por los cuales las variables se introducen y se eliminan de la ecuación. Puede especificar criterios para la Entrada o para la Salida de variables, de manera que una variable se introduce en el modelo si la probabilidad de su estadístico de puntuación es menor que el valor de entrada, y se elimina si la probabilidad es mayor que el valor de salida. Para este modelo se dejaron los valores por defecto que trae el software.

En la opción Punto de corte para la clasificación, permite determinar el punto de corte para la clasificación de los casos. Los casos con valores pronosticados que han sobrepasado el punto de corte para la clasificación se clasifican como positivos (tendrían el evento o resultado que se modeliza), mientras que aquellos con valores pronosticados menores que el punto de corte se clasifican como negativos (no tendrían el evento o resultado). Para este modelo se dejaron los valores por defecto que trae el aplicativo.

Por último, en la opción Incluir constante en el modelo, le permite indicar si el modelo debe incluir un término constante. Si se desactiva, el término constante será igual a 0.

8.5 Fase de evaluación:

Estimando los resultados de los modelos anteriormente mencionados, se hace una valoración de la información resultante y se identifica que hay que hacer algunos cambios.

En el modelo de Regresión Logística, se tuvo que transformar el valor de las categorías de la variable “Riesgo”, porque al dar el resultado estaba clasificando la información por la categoría “Bajo”, lo necesario es que se clasifique la información por la categoría “Alto” la cual es la que explica cuáles son los estudiantes que realmente se encuentran en riesgo de deserción.

La transformación se hizo sustituyendo la categoría “Alto” por el número 1 y la categoría “Bajo” por el número 0.

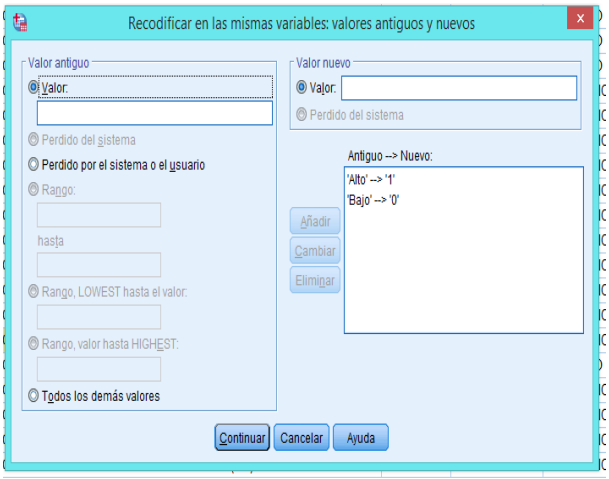

Ilustración 11.- Cuadro Transformación (Elaboración propia)

8.6 Fase de distribución:

Árbol de clasificación:

La tabla de resumen del modelo proporciona información general sobre las especificaciones utilizadas para crear el modelo y sobre el modelo resultante. La sección especificaciones ofrece información sobre valores de configuración utilizados para generar el modelo de árbol, incluidas las variables utilizadas en el análisis. La sección Resultados muestra información sobre el número de nodos totales y terminales, la profundidad del árbol (número de niveles por debajo del nodo raíz) y las variables independientes incluidas en el modelo final. Se decide realizar el análisis únicamente con dos categorías de riesgo, puesto que, al realizarse con tres la significancia era menor al 50%.

Resumen del modelo

Especificaciones	Método de crecimiento	CHAID
	Variable dependiente	Riesgo
	Variables independientes	Sede, Jornada, Estado civil , Genero, Convenio_Tipo, Primiparo, Edad, Semestre
	Validación	Ninguna
	Máxima profundidad del árbol	3
	Casos mínimos en nodo padre	100
	Casos mínimos en nodo hijo	50
Resultados	Variables independientes incluidas	Edad, Jornada, Convenio_Tipo, Semestre , Sede
	Número de nodos	19
	Número de nodos terminales	12
	Profundidad	3

Ilustración 12.-Resumen del modelo (Elaboración propia).

El Diagrama de árbol obtenido es una representación gráfica del modelo del árbol. En el ejemplo, todas las variables son tratadas como nominales y cada nodo contiene una tabla de frecuencias que muestra el número de casos (frecuencia y porcentaje) para cada categoría de la variable dependiente.

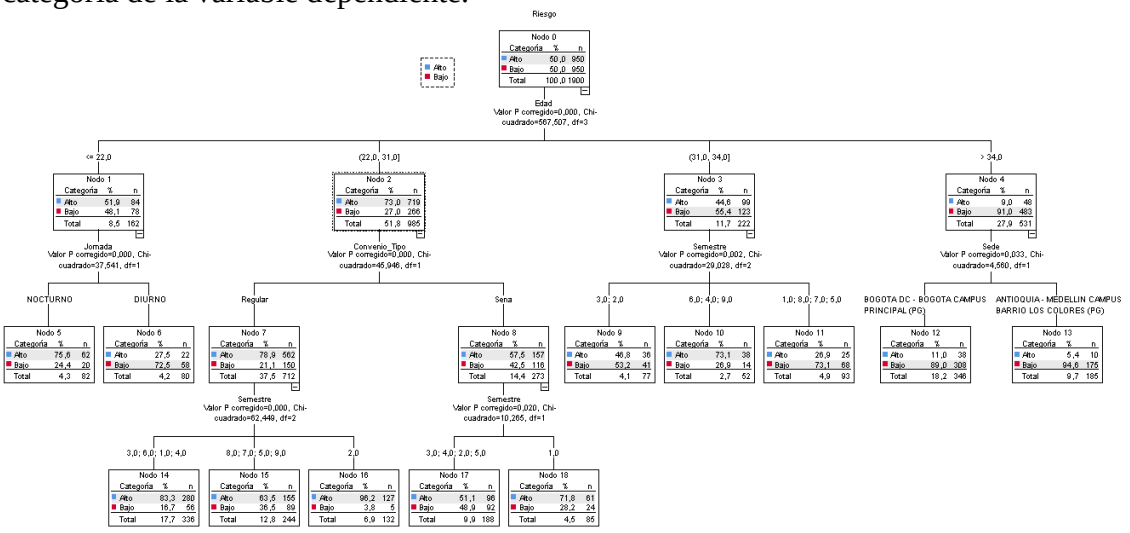

Ilustración 13.-Modelo Árbol de clasificación. (Elaboración propia).

1. En primer lugar, el nodo 0 que describe la variable dependiente: porcentaje de los estudiantes que están en riesgo alto o bajo de desertar.

2. En seguida se observa que la variable dependiente se ramifica en dos nodos: Nodo 1 y 2 pertenecientes a la variable *Edad*, indicando que ésta es la variable principal predictora.

3. Se observa el nodo 2, que es el que tiene una prueba chi cuadrado de 45,94 siendo superior a los demás nodos, diciendo que de un total de 51,8 de personas de 22 y 31 años el 73% se encuentra en riesgo alto.

4. El nodo 2 se vuelve a ramificar en los nodos 7 y 8 pertenecientes a la variable *Tipo de convenio* y se identifica que el nodo 7 con la categoría Regular tiene un valor de chi cuadrado de 62,44 clasificando de un total 37,5 personas el 78,9% están en riesgo

alto de desertar. Por contario el nodo 8 con la categoría Sena tiene un valor de chi cuadrado de 10,25, clasificando un total de 14,4 personas con un 57.5% de riesgo alto.

5. El nodo 7 se ramifica en los nodos 14, 15 y 16 que pertenecen a la variable *Semestre* y se identifica que el nodo 16 tiene una probabilidad de 96,2% de tener riesgo alto, es decir, que las personas que están matriculadas en el semestre 2 tienen mayor probabilidad de tomar la decisión de desertar.

6. El nodo 8 se ramifica en los nodos 17 y 18 pertenecientes a la variable *Semestre* y se evidencia que en el nodo 18, correspondiente al semestre 1 tiene u nivel de riesgo alto con un 71.8% de probabilidad.

7. Por tanto, se puede concluir que los nodos que definen el modelo son: Nodo 0, Nodo 2, Nodo 7 y Nodo 16, es decir, que las variables que influyen son: *Edad*, *Tipo de convenio y Semestre.*

Clasificación

	Pronosticado		
Observado	Alto	Bajo	Porcentaje correcto
Alto	819	131	86,2%
Bajo	300	650	68,4%
Porcentaje global	58,9%	41,1%	77,3%

Método de crecimiento: CHAID
Variable dependiente: Riesgo

Ilustración 14.-Tabla de clasificación. (Elaboración propia).

La tabla de clasificación demuestra que el modelo se ajusta de forma correcta, aproximadamente al 77,3% de los individuos puestos en estudio. De tal forma específica para cada categoría de la variable dependiente ofrece un "porcentaje correcto" más elevado al caso de la categoría "Alto", con un 86,2%.

Regresión Logística:

En la hoja de resultados se evidencia que lo primero que se observa es la tabla de resumen de procesamiento de casos introducidos, los seleccionados para el análisis y los excluidos por tener un valor faltante.

Resumen de procesamiento de casos

Casos sin ponderar[a]		N	Porcentaje
Casos seleccionados	Incluido en el análisis	1750	100,0
	Casos perdidos	0	,0
	Total	1750	100,0
Casos no seleccionados		0	,0
Total		1750	100,0

a. Si la ponderación está en vigor, consulte la tabla de clasificación para el número total de casos.

Ilustración 15.-Resumen de procesamiento de casos (Elaboración propia).

A continuación, aparece una tabla que especifica la codificación de la variable dependiente, la cual debe ser dicotómica, en este caso la categoría "Alto" quedo codificada con el número 1 y la categoría "Bajo" con el número 0.

Codificación de variable dependiente

Valor original	Valor interno
0	0
1	1

Ilustración 16.-Codificación de la variable dependiente (Elaboración propia).

En la siguiente tabla se muestra la codificación de las variables categóricas de manera dicotómica, el modelo tiene 3 variables de este tipo, "*Sede, Convenio_tipo y Jornada*". También señala la frecuencia absoluta de cada uno de los valores

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetro (1)
Sede	ANTIOQUI	85	1,000
	BOGOTA D	1665	,000
Convenio_Tipo	Regular	1293	1,000
	Sena	457	,000
Jornada	DIURNO	261	1,000
	NOCTURNO	1489	,000

Ilustración 17.- Codificación de las variables categóricas (Elaboración propia).

Bloque 0: Bloque inicial

En este bloque inicial se calcula la verosimilitud de un modelo que sólo tiene el término constante (a o b0). Puesto que la verosimilitud L es un número muy pequeño (comprendido entre 0 y 1), se suele ofrecer el logaritmo neperiano de la verosimilitud (LL), que es un número negativo, o el menos dos veces el logaritmo neperiano de la verosimilitud (-2LL), que es un número positivo. El estadístico (-2LL) mide hasta qué punto un modelo se ajusta bien a los datos, esta medición recibe también el nombre de "desviación", cuanto más pequeño sea el valor, mejor será el ajuste.

Historial de iteraciones[a,b,c]

Iteración		Logaritmo de la verosimilitud -2	Coeficientes Constante
Paso 0	1	2413,142	,171
	2	2413,142	,172

a. La constante se incluye en el modelo.

b. Logaritmo de la verosimilitud -2 inicial: 2413,142

c. La estimación ha terminado en el número de iteración 2 porque las estimaciones de parámetro han cambiado en menos de ,001.

Ilustración 18.- Historial de iteraciones (Elaboración propia).

El proceso ha necesitado de 2 ciclos para estimar correctamente el termino constante porque la variación de -2LL ha cambiado en menos del criterio fijado por el programa (0,001), también muestra el valor del parámetro calculado (b0 = 0,172).

Tabla de clasificación[a,b]

			Pronosticado		
			Riesgo		Porcentaje correcto
	Observado		0	1	
Paso 0	Riesgo	0	0	800	,0
		1	0	950	100,0
	Porcentaje global				54,3

a. La constante se incluye en el modelo.

b. El valor de corte es ,500

Ilustración 19.- Tabla de clasificación (Elaboración propia).

La tabla de clasificación permite evaluar el ajuste del modelo de regresión con un solo parámetro en la ecuación, comparando los valores predichos con los valores observados. Por defecto se ha empleado un punto de corte de la probabilidad de Y para clasificar a los individuos de 0,5: esto significa que aquellos sujetos para los que la ecuación con éste único término calcula una probabilidad < 0,5 se clasifican como ESTADO=0 "Bajo", mientras que si la probabilidad resultante es ≥ 0,5 se clasifican como ESTADO=1 "Alto", teniendo en cuenta que se refiere a la clasificación categórica del riesgo de que un estudiante deserte. La tabla demuestra que ha clasificado correctamente a un 54,3% de los casos y ningún caso con riesgos "Bajo" ha sido clasificado correctamente.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	,172	,048	12,826	1	,000	1,187

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Edad	54,583	1	,000
		Jornada(1)	,398	1	,528
		Convenio_Tipo(1)	8,755	1	,003
		Semestre	14,656	1	,000
		Sede(1)	12,529	1	,000
		Estadísticos globales	114,812	5	,000

Ilustración 20.- Variables en la ecuación (Elaboración propia).

Finalmente se presenta el parámetro estimado (B), el error estándar, la significación estadística con la prueba Wald que es un estadístico que sigue una ley Chi cuadrado con 1 grado de libertad. Y la estimación de la OR (Exp(B)). En la ecuación de regresión sólo aparece, en este primer bloque, la constante, habiendo quedado afuera todas las variables de estudio, como tienen una significancia estadística asociada al índice de Wald el proceso automático por pasos continuará, incorporándolas a la ecuación.

Bloque 1: Método = Por pasos hacia adelante (Razón de verosimilitud)

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2292,591[a]	,067	,089

a. La estimación ha terminado en el número de iteración 4 porque las estimaciones de parámetro han cambiado en menos de ,001.

Ilustración 21.- Resumen del modelo (Elaboración propia).

La tabla de resumen del modelo se complementa de tres campos para evaluar de forma global la validez del modelo.

-2 log de la verosimilitud (-2LL): mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de "desviación". Cuanto más pequeño sea el valor, mejor será el ajuste.

R cuadrado de Cox y Snell – R cuadrado de Nagelkerke: es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes). La R cuadrado de Cox y Snell se basa en la comparación del log de la verosimilitud (LL) para el modelo respecto al log de la verosimilitud (LL) para un modelo de línea base. Sus valores oscilan entre 0 y 1. En el análisis realizo, arrojo el valor de 0,067 para el R cuadrado de Cox y Snell y para el R cuadrado de Nagelkerke el valor fue de 0,089.

Tabla de clasificación[a]

			Pronosticado		
			Riesgo		Porcentaje correcto
	Observado		0	1	
Paso 1	Riesgo	0	387	413	48,4
		1	251	699	73,6
	Porcentaje global				62,1

a. El valor de corte es ,500

Ilustración 22.- Tabla de clasificación a (Elaboración propia).

Se observa un porcentaje global de significancia del (62,1%). Y se denota que está haciendo una buena clasificación de la variable predictora.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Edad	-,094	,013	53,062	1	,000	,910	,888	,934
	Jornada(1)	-,385	,145	7,069	1	,008	,681	,512	,904
	Convenio_Tipo(1)	,761	,133	32,991	1	,000	2,141	1,651	2,776
	Semestre	-,109	,024	20,300	1	,000	,897	,855	,940
	Sede(1)	1,358	,278	23,863	1	,000	3,889	2,255	6,706
	Constante	2,667	,371	51,628	1	,000	14,396		

a. Variables especificadas en el paso 1: Edad, Jornada, Convenio_Tipo, Semestre , Sede.

Ilustración 23.- Variables en la ecuación (Elaboración propia).

Si el contexto es predictivo, la probabilidad de un suceso para un perfil de entrada dado a computarse independientemente empleando los coeficientes estimados. Se aplica un modelo ajustado para tener diferentes tipos de casos.

- Se calcula la probabilidad de riesgo alto de que una persona deserte, esta persona tiene una edad de 22 años, estudia en la jornada nocturna, tiene convenio tipo Sena, está en el semestre 1 y pertenece a la sede de Bogotá.

$$\text{P (Alto riesgo)} = \frac{1}{1 + e^{2{,}667 - 0{,}094*22 - 0{,}385*1 + 0{,}761*1 - 0{,}109*1 + 1{,}358*1}} = 0{,}9023$$

Las personas descritas con las anteriores cualidades tienen una probabilidad de desertar de un 90%.

- Se calcula la probabilidad de que una persona de 30 años pertenezca a la jornada diurna, tiene convenio de tipo Sena, se encuentra en el semestre 2 y hace parte de la sede de Antioquia.

$$\text{P (Alto riesgo)} = \frac{1}{1 + e^{2{,}667 - 0{,}094*30 - 0{,}385*0 + 0{,}761*0 - 0{,}109*2 + 1{,}358*0}} = 0{,}5962$$

El porcentaje de alto riesgo de deserción para este tipo de personas es del 60%.

- Una persona que tenga 40 años, que haga parte de la jornada nocturna, tenga un convenio tipo Sena regular, este cursando el semestre 3 y pertenezca a la sede de Bogotá.

$$\text{P (Alto riesgo)} = \frac{1}{1 + e^{2{,}667 - 0{,}094*40 - 0{,}385*1 + 0{,}761*0 - 0{,}109*3 + 1{,}358*1}} = 0{,}3900$$

- El porcentaje de riesgo alto para esta persona es del 39%, pero si se cambia el ejemplo únicamente analizando que esa misma persona pertenezca al convenio tipo Sena, la probabilidad de que deserte incrementa a un 58%.

$$\text{P (Alto riesgo)} = \frac{1}{1 + e^{2{,}667 - 0{,}094*40 - 0{,}385*1 + 0{,}761*1 - 0{,}109*3 + 1{,}358*1}} = 0{,}5778$$

Se evidencia que las variables seleccionadas para el desarrollo de este proyecto son realmente significativas y están explicando la variable independiente.

9. DIAGRAMA DE ACTIVIDADES:

	Fecha		Duración	Agosto			Septiembre				Octubre				Noviembre				Diciembre			
# ACTIVIDAD	Inicio	Fin	(Dias)	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Actividad 01	7/08/2019	7/08/2019	0																			
Ante - proyecto																						
Actividad 02	11/08/2019	20/08/2019	9																			
Desarrollo de proyecto																						
Actividad 03	21/08/2019	30/08/2019	9																			
Actividad 04	16/09/2019	30/09/2019	14																			
Actividad 05	1/10/2019	10/10/2019	9																			
Actividad 06	5/10/2019	21/10/2019	16																			
Actividad 07	22/10/2019	30/10/2019	8																			
Actividad 08	1/11/2019	7/11/2019	6																			
Actividad 09	12/11/2019	18/11/2019	6																			
Actividad 10	19/11/2019	21/11/2019	2																			
Actividad 11	22/11/2019	25/11/2019	3																			
Actividad 12		9/12/2019	4																			

Ilustración 25.- Cronograma de actividades

9.1. DESCRIPCIÓN DE ACTIVIDADES:

# Actividad	Descripción
Actividad 01	Planteamiento del problema
Actividad 02	Se realiza la formulación del problema, Justificación, alcance y objetivo general y objetivos específicos para una primer entrega
Actividad 03	Se realizan investigaciones para el marco teórico y estado del arte
Actividad 04	Se hace investigaciones de metodologías aplicadas a proyectos de minería de datos y se escoge la metodología CRISP-DM.
Actividad 05	Recolección de datos suministrados por el área.
Actividad 06	Se hace control de calidad a la información recibida
Actividad 07	Determinación de variables que se aplicaran al modelo.
Actividad 08	Aplicación del modelo árboles de clasificación.
Actividad 09	Aplicación del modelo regresión logística.
Actividad 10	Interpretación y documentación de resultados.
Actividad 11	Documentación de conclusiones y recomendaciones.
Actividad 12	Presentación del proyecto.

Tabla 11.- Descripción de actividades (Elaboración propia)

10. CONCLUSIONES:

- La metodología CRISP-DM fue de gran ayuda para el desarrollo de este proyecto, puesto que, se trata de un proyecto de minería de datos, para el que se conocían los objetivos iniciales, pero no las actividades a realizar, los pasos a realizar, ni las técnicas de modelado, al hacer la aplicación de esta técnica en el desarrollo del proyecto se identifica que se está cumpliendo con los objetivos principalmente propuestos.

- Con la identificación y extracción de datos se logró consolidar la información de las diferentes fuentes de información que fueron suministradas, con la finalidad de obtener la información unificada y depurada para la realización de los análisis.

- Los árboles de clasificación resultaron ser una herramienta muy útil, ya que permitió ver de manera gráfica las diferentes opciones que se tenían para la toma de decisiones.

- Los árboles de decisión se pueden tornar extensos debido a la cantidad de variables inciertas que se tienen, lo que vuelve un poco complejo el análisis, en este caso el árbol de clasificación resulto efectivo, puesto que el modelo solo tiene nueve variables inciertas y se ajustaron perfectamente a la predicción.

- Aplicando el modelo de regresión logística, se validó que las variables que dieron como resultante ser significativas *(Sede, Edad, Convenio tipo, Semestre y Jornada)* del modelo de árboles de clasificación realmente si tenían importancia en el modelo de propuesto.

11. RECOMENDACIONES:

- Realizar encuesta de caracterización parametrizada para poder contar con más información personal del estudiante para lograr hacer mejores análisis en cuanto a más variables que pueden aplicar al modelo.

- Tener las fuentes de información consolidadas, para evitar retrasos y reprocesos.

12. REFERENCIAS:

Abrego Almazan , D., Sánchez Tovar , Y., & Medina Quintero, J. (s.f.). *ScienceDirect.* Obtenido de Influencia de los sistemas de información en los resultados organizacionalesInfluence of information systems in organizational performance: https://reader.elsevier.com/reader/sd/pii/S0186104216300432?token=EA0CF781F8E1E6B2A0D2AB24413756F02CAEC5FC4E97C1817F5988536F5A1D9391EEFAF1D5E1A6D8DBC074AE013ECE6A

Ahumada Tello, E., & Perusquia Velasco, J. A. (s.f.). *SciencieDirect*. Obtenido de Inteligencia de negocios: estrategia para el desarrollo de competitividad en empresas de base tecnológicaBusiness intelligence: Strategy for competitiveness development in technology-based firms: https://www.sciencedirect.com/science/article/pii/S0186104215000807

Arismendy Fuentes, C. A., & Morales Parrado, N. L. (2018). *Modelo de regresión logística como alternativa para medir la probabilidad de deserción temprana en la universidad de los llanos periodo 2015-2 – 2018-1.* Obtenido de https://repositorio.unillanos.edu.co/jspui/bitstream/001/1101/1/RUNILLANOS%20ECO%200417%20MODELO%20DE%20REGRESION%20LOGISTICA%20COMO%20ALTERNATIVA%20PARA%20MEDIR%20LA%20PROBABILIDAD%20DE%20DESERCION%20TEMPRANA%20EN%20LA%20UNIVERSIDAD%20DE%20LOS%20LLANOS%2

Barahona Urbina, P., Aliaga Prieto , V., & Veres Ferrer, E. (2 de Diciembre de 2016). *Deserción académica de la universidad de Atacama, Chile.* Obtenido de https://www.redalyc.org/pdf/4498/449849320003.pdf

Berlanga Silvente, V., Rubio Hurtado, M., & Rith Vilá , B. (08 de Enero de 2014). *http://revistes.ub.edu/index.php/REIRE/article/viewFile/reire2013.6.1615/7229.* Obtenido de http://revistes.ub.edu/index.php/REIRE/article/viewFile/reire2013.6.1615/7229

Calvo, M. (2019). *Captio.* Obtenido de Business intelligence: ¿Por qué lo necesito en mi empresa?: https://www.captio.net/blog/algunos-ejemplos-practicos-de-uso-de-business-intelligence

Carlos Márquez Vera, C. R. (3 de Noviembre de 2012). *Predicción del Fracaso Escolar Mediante Técnicas de Minería de Datos.* Obtenido de http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

Castro, J. (12 de Agosto de 2015). *CORPONET*. Obtenido de ¿Qué es la inteligencia de negocios y cómo beneficia a tu empresa?: https://blog.corponet.com.mx/que-es-la-inteligencia-de-negocios

Celis, S., Moreno, L., Poblete, P., Villanueva, J., & Weber, R. (Septiembre de 2015). *Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.* Obtenido de dii.uchile.cl/~ris/RIS2015/rendimientoac.pdf

Diaz, D. B., & Garzón, L. P. (s.f.). *Red de Revistas Científicas de América Latina, el Caribe, España y Portugal Sistema de Información Científica.* Obtenido de Red de Revistas Científicas de América Latina, el Caribe, España y Portugal Sistema de Información Científica: https://www.redalyc.org/pdf/4096/409633954005.pdf

Dinero, R. (29 de Junio de 2017). *Publicaciones Semana S.A.* Obtenido de https://www.dinero.com/pais/articulo/desercion-y-abandono-de-la-educacion-universitaria-en-colombia/247068

Fernandes, R. N. (Julio de 2017). *TRABAJO FIN DE GRADO.* Obtenido de Análisis de datos sanitarios aplicando metodología CRISP-DM: https://repositorio.uam.es/bitstream/handle/10486/680056/nishizaki_fernandes_rafael_tfg.pdf?sequence=1&isAllowed=y

Flores Pulido , L., Dávila de la Rosa, E., & Rodriguez Florez, F. (11 de Octubre de 2016). *Inteligencia de negocios y minería de datos aplicado a la.* Obtenido de Inteligencia de negocios y minería de datos aplicado a la: https://pdfs.semanticscholar.org/ed31/f60175e0b6d3efd82789b1778cfbb363c369.pdf

Galan Cortina, V., & Castro Gálan , E. (Octubre de 2015). *Aplicación de la metodología CRISP-DM a unproyecto de mineria de datos en el entorno universitario.* Obtenido de https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Galán Cotina, V. (Octubre de 2015). *Aplicación de Metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.* Obtenido de https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Hernández Gonzales , A., Meléndez Armenta , R., & Morales Rosales, L. (22 de Diciembre de 2016). *IEEE xplore.* Obtenido de https://ieeexplore.ieee.org/document/7795831

Inteligencia Competitiva - Politécnico Grancolombiano. (2018). *Tasa de permanencia .* Bogotá.

Marquez Vera, C., Romero Morales, C., & Ventura Soto, S. (3 de Noviembre de 2012). *Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos.* Obtenido de http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

Maya Lopera, E. (2018). *Los árboles de decisión como herramienta para el análisis de riesgos de los proyectos.* Obtenido de https://repository.eafit.edu.co/bitstream/handle/10784/12980/Elena_MayaLopera_2018.pdf?sequence=2&isAllowed=y

Mérchan Rubiano, S., & Duarte Garcia, J. (29 de Agosto de 2016). *Análisis de técnicas de minería de datos para construir un modelo predictivo para el rendimiento académico.* Obtenido de https://ieeexplore.ieee.org/document/7555255

Ortiz Lozano, J., Rua Vieites , A., & Bilbao Calabuig , P. (23 de Octubre de 2017). *Aplicación de árboles declasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas.* Obtenido de https://repositorio.comillas.edu/rest/bitstreams/146056/retrieve

Peralta, F. (Octubre de 2014). *ResearchGate.* Obtenido de https://www.researchgate.net/figure/Fases-del-modelo-de-proceso-de-la-metodologia-CRISP-DM-35_fig2_284215308

Pereira Timarán , R., Zambrano Caicedo, J., & Troya Hidalgo, A. (21 de 08 de 2018). *Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°.* Obtenido de http://www.scielo.org.co/pdf/ridi/v9n2/2389-9417-ridi-9-02-363.pdf

Pérez Méndez, J., & Machado Cabezas , A. (Junio de 2015). *ScienceDirect.* Obtenido de ScienceDirect: https://www.sciencedirect.com/science/article/pii/S113848911400017X

Piatesky, G. (2014). *KDnuggets.* Obtenido de https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html#comments

Porcel, E., Dapazo, G., & López, M. (2019). Predicción del rendimiento académico de los alumnos de primer año de la (FACENA-UNNE) en función de su caracterización socioeducativa. *redie*, https://redie.uabc.mx/redie/article/view/264. Obtenido de https://redie.uabc.mx/redie/article/view/264

Portafolio. (06 de Junio de 2017). El 50% de los universitarios colombianos abandonan sus estudios. *Portafolio*, págs. https://www.portafolio.co/tendencias/la-mitad-de-los-universitarios-colombianos-abandona-sus-estudios-506571.

Rico Higuita , D. (Mayo de 2016). *Caracterización de la deserción estudiantil en la Universidad Nacional de Colombia sede Medellín.* Obtenido de https://slidex.tips/download/caracterizacion-de-la-desercion-estudiantil-en-la-universidad-nacional-de-colomb

Rodriguez Flores , F., Flores Pulido, L., & Dávila de la Rosa, E. (11 de Octubre de 2016). *Inteligencia de negocios y minería de datos aplicado a la.* Obtenido de https://pdfs.semanticscholar.org/ed31/f60175e0b6d3efd82789b1778cfbb363c369.pdf

Sifuentes Bitocchi, O. (20 de 08 de 2018). *Modelos predictivos de la deserción estudiantil en una universidad privada peruana.* Obtenido de https://www.researchgate.net/publication/329847510_Modelos_predictivos_de_la_desercion_estudiantil_en_una_universidad_privada_peruana

Timarán Pereira , R., & Jiménez Toledo, J. (14 de Noviembre de 2014). *Detección depatrones de deserción estudiantil en programas de pregrado de instituciones de educación superiror con CRISP-DM.* Obtenido de file:///C:/Users/francy/Downloads/758%20(2).pdf

Triana Marquín , M. (2017). *Predicción del rendimiento académico mediante técnicas de mineria de datos* . Obtenido de http://repositorio.utp.edu.co/dspace/bitstream/handle/11059/8356/37126M357.pdf?sequence=1&isAllowed=y

Universa, C. (14 de Mayo de 2019). *UNIVERSA*. Obtenido de https://noticias.universia.net.co/educacion/noticia/2015/12/17/1134832/20-carreras-universitarias-mayor-demanda-mejor-pagadas-colombia.html

Vélez Garcia, G. (Diciembre de 2018). *Proyecto de Grado presentado como requisito para optar al Título de.* Obtenido de APLICACIÓN DE LA METODOLOGÍA CRISP-DM® A LA RECOLECCIÓN Y ANÁLISIS DE DATOS GEORREFERENCIADOS DESDE TWITTER®: https://repository.unimilitar.edu.co/bitstream/handle/10654/20099/GarciaVelezGustavoAdolfo2018.pdf?sequence=4&isAllowed=y