MODELO DE MINERÍA DE DATOS APLICADO A LA DETECCIÓN DE PROPENSIÓN A DESAPROBACIÓN EN EXAMEN APRENDER EN EL ÁREA DE MATEMÁTICAS

Estudiante:
VIOLI PABLO EZEQUIEL

Departamento de Desarrollo Productivo y Tecnológico
Universidad Nacional de Lanús

Tutores:
Dr. Pytel Pablo
Lic. Loidi Laura Gabriela

TRABAJO FINAL INTEGRADOR PRESENTADO PARA OBTENER EL TÍTULO DE LICENCIADO EN SISTEMAS

Año 2021

Agradecimientos

A Mg. Pablo Pytel. Sus consejos fueron siempre útiles cuando no salían de mi pensamiento las ideas para escribir lo que hoy he logrado. Formó parte importante de esta historia con sus aportes profesionales que lo caracterizan. Muchas gracias por estar allí cuando mis horas de trabajo se hacían difíciles. Gracias por sus orientaciones.

A Lic. Laura Loidi, por orientarme y ayudarme en los inicios del trabajo, donde más confusiones había. Muchas gracias por su acompañamiento y paciencia.

A la Universidad Nacional de Lanús, por permitirme formarme como profesional y hacerme crecer como persona.

A mis profesores. Donde quiera que vaya, los llevaré conmigo en mi transitar profesional. Gracias por su paciencia, por compartir sus conocimientos de manera profesional e invaluable, por su dedicación perseverancia y tolerancia.

A mis Padres, que han sido siempre el motor que impulsa mis sueños y esperanzas, quienes estuvieron siempre a mi lado en mi desarrollo profesional. Hoy cuando concluyo mis estudios, les dedico este logro, como una meta más conquistada.

Una mención especial a mis amigos y compañeros de la universidad, y a mis compañeros de trabajo. Hoy nos toca cerrar un capítulo maravilloso en esta historia de mi vida y no puedo dejar de agradecerles por su apoyo y constancia, al estar en las horas más difíciles, por compartir horas de estudio y compartir sus conocimientos. Gracias por estar siempre allí.

Resumen

La minería de datos puede definirse como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, a partir de grandes volúmenes de datos. Hoy en día, las distintas fuentes de información disponibles en un mismo negocio generan la necesidad de buscar nuevos procesos para poder transformar dichos datos en información útil para las organizaciones. El rápido crecimiento en la capacidad para almacenar datos, que están experimentando los procesos de recolección de información sobre la población, proporciona nuevas posibilidades para analizar su comportamiento. En el siguiente trabajo se aplican en la encuesta Aprender realizada por el Ministerio de Educación Anualmente en los Colegios de la nación Argentina dichas técnicas de minería de datos. El objetivo es detectar tempranamente aquellos estudiantes que tendrán una actuación calificada como regular o sobresaliente en los exámenes de Matemáticas, para poder agruparlos y describir sus características más influyentes. Los resultados de este trabajo pueden aplicarse para generar un espacio de ayuda y apoyo en las primeras etapas del estudiante, para acompañarlos y ayudarlos en la enseñanza. Asimismo, este trabajo busca también aportar mecanismos para detectar aquellos estudiantes con perfiles orientados a ciencias exactas, para poder brindarles ayuda pedagógica y mostrarle los posibles caminos universitarios que puede seguir en su futura educación.

Abstract

Data mining can be defined as the process of extracting previously unknown useful and understandable knowledge from large volumes of data. Today, the different sources of information available in the same business, generate the need to search for new processes to be able to transform data into useful information for organizations. The rapid growth in the capacity to store data that the processes of collecting information on the population are undergoing, provides new possibilities to analyze their behavior. In the following work, data mining techniques are applied in the “Aprender” survey carried out by “Ministerio de Educacion Argentina” annually in the Schools of the Argentine nation. Our objective is to detect early those students who will have a performance rated as regular or outstanding in Mathematics exams, in order to group them and describe their most influential characteristics. The results of this work can be carried out to generate a space for help and support in the early stages of the student, to accompany them and help them in teaching. Likewise, this work also seeks to provide mechanisms to detect those students with profiles oriented to exact sciences, to be able to provide them with pedagogical help and show them the possible university paths that they can follow in their future education.

Índice

Índice de figuras

Índice de Tablas

Índice de Nomenclaturas

AUC: Área bajo la curva

CBC: Contenidos Básicos Comunes

CRISP-DM: Proceso estándar de la industria para la minería de datos

DINIECE: Dirección Nacional de Información y Evaluación de la Calidad Educativa

FN: Falso Negativo

FP: Falso positivo

IA: Inteligencia Artificial

KDD: Descubrimiento de Conocimientos en Bases de Datos

KNN: K-Vecinos Más Próximos

KS: Kolmogorov-Smirnov

NAP: Núcleos de Aprendizaje Prioritarios

ODDS: logaritmo natural de las probabilidades

ONE: Operativos Nacional de Evaluación

R2: R cuadrado

ROC: Característica Operativa del Receptor

SEMMA: Muestrear, Explorar, Modificar, Modelar y Evaluar

TDIDT: Árboles de Decisión de tipo Inducción de arriba hacia abajo

TN: Verdadero Negativo

TP: Verdadero Positivo

TRC: pruebas referidas a criterio

TRI: Teoría de Respuesta al Ítem

TRN: pruebas referidas a normas

1.Introducción

Este proyecto es presentado como trabajo final integrador con el fin de obtener el título de Licenciado en Sistemas en la universidad Nacional de Lanús. Su desarrollo comenzó en el año 2020 y se da por finalizado durante el mes de diciembre del 2021.

El aumento del volumen y variedad de información que se encuentra almacenada en grandes bases de datos digitales y otras fuentes ha crecido enormemente en las últimas décadas. Gran parte de esta información es histórica por lo que representa transacciones o situaciones que se han producido en un periodo determinado (Hernandez, Ramirez Quintana, & Ferri Ramirez, 2004). En el mundo actual en que vivimos, donde cada vez es más importante tener todo informatizado y cuantificado en las bases de datos de cada empresa u organización, surge la necesidad de encontrar alguna manera de sacar conclusiones a partir de estos datos, ya que, por sí solos, los datos serían nada más que registros sin significado y es necesarios "explotarlos" para sacarles provecho y obtener algún tipo de información valiosa de ellos (Hernandez, Ramirez Quintana, & Ferri Ramirez, 2004).

Dicha Explotación de los Datos se lleva a cabo a través de la Minería de Datos (Britos & García-Martínez, 2009). Para comenzar un proceso de Minería de Datos, es importante partir de una base de datos estructurada, la cual puede estar almacenada en un data warehouse (almacén de datos). En este contexto, se ha detectado la necesidad de analizar el desempeño de los estudiantes en el examen de matemática dentro de las encuestas Aprender generadas por el Ministerio de Educación. De esta manera, se trata de mejorar el desempeño general de los estudiantes en dicho examen, a través de la detección de manera temprana de aquellos que tendrán dificultad en la resolución del mismo.

Aunque existen diversas técnicas y metodologías que se pueden utilizar, se debe elegir la más adecuada para cada caso concreto. En el presente proyecto se ha seleccionado la metodología CRISP-DM (Chapman, y otros, 2000), la cual es explicada en forma detallada más adelante indicando los motivos que han motivado esta elección.

El trabajo está dividido en cuatro capítulos, diferenciados por el objetivo de los mismos.

El primer capítulo (Introducción) trata de introducir al lector en el problema propuesto y dar una visión general del método de abordaje para la solución del problema.

El segundo capítulo (Marco Teórico & Metodológico) contiene información teórica que rodea al proyecto, para comprender la dimensión, el contexto del mismo e información referente a la actualidad. Por otro lado, también se abarcan temas referentes a la minería de datos, herramientas utilizadas durante el proyecto y metodologías abordadas.

En el tercer capítulo (Modelo Predictivo), se aplican la metodología CRISP-DM previamente explicadas a las bases de encuestas Aprender, realizando el paso a paso hasta la última etapa de implementación del proyecto.

En el cuarto capítulo (Líneas finales), se incluyen la conclusión del proyecto y los avances esperados a futuro.

Por último, se presentan toda la referencia bibliográfica utilizada durante el proyecto y un anexo con información adicional sobre los datos a ser utilizados para la construcción del Modelo Predictivo.

2. Marco Teórico, Metodologías y Antecedentes

En este capítulo se detallan los principales conceptos teóricos que se aplican en el presente trabajo permitiendo de esta manera comprender la dimensión, el contexto del mismo e información referente a la actualidad. En primer lugar, en la sección 2.1 se presentan las Encuestas Aprender. Luego, en las secciones 2.2 y 2.3, se presentan los conceptos de bases de datos y explotación de información o minería de datos respectivamente. En la sección 2.4 se presentan distintas metodologías para trabajar con Minería de datos y, en la 2.5, se referencian las herramientas a utilizar, junto a una breve descripción de las mismas. Por otro lado, en la sección 2.6 se introducen los conceptos de modelado y etapas del mismo. Por último, en las secciones 2.7, 2.8 y 2.9, se presentan detalles sobre el modelado, tales como la limpieza de la base, los algoritmos utilizados en el proceso y las métricas correspondientes al proceso.

2.1 Encuestas Aprender

Las encuestas Aprender son el dispositivo nacional de evaluación de los aprendizajes de los estudiantes y de sistematización de información acerca de algunas condiciones en las que ellos se desarrollan (Ministerio de Educación, s.f.). De esta manera, se busca obtener y generar información oportuna y de calidad para conocer mejor los logros alcanzados y los desafíos pendientes en torno a los aprendizajes de los estudiantes para contribuir a procesos de mejora educativa continua. Por lo tanto, como toda la información que produce el Ministerio de Educación, se constituyen en una herramienta básica para la planificación de la educación a nivel nacional, así como para las investigaciones y proyecciones que se realizan en los ámbitos académico y privado (Ministerio de Educación, s.f.)

Esta evaluación ha sido desarrollada por el Ministerio de Educación, Cultura, Ciencia y Tecnología, a través de la Secretaría de Evaluación Educativa, en acuerdo con el Consejo Federal de Educación y con la participación del Cuerpo Colegiado Federal de docentes y especialistas de todo el país. En este sentido, Gentile (2015) afirma:

> *"El cuestionario complementario indaga en información que permite analizar los logros de aprendizaje en clave de contexto. De esta forma, Aprender brinda información sobre*

clima escolar, autopercepción del estudiante, prácticas educativas y uso de tecnología, entre otras informaciones.”

La encuesta se realiza anualmente desde el 2016 en adelante para nivel primario y secundario en todos los colegios del país, excepto durante el 2020 dado que se vieron postergadas por la situación sanitaria que atravesaba Argentina. Sus resultados son luego analizados en términos de (Secretaria de Evaluación e Información Educativa, 2019):

- Reporte por Escuela
- Reporte Nacional de Resultados
- Reportes jurisdiccionales y regionales
- Sistema Abierto de Consulta
- Presentación Interactiva de Datos

Hasta el año 2015, se realizaban los Operativos Nacional de Evaluación (ONE) que se pueden considerar como su antecesor directo. Estos operativos eran implementados por la entonces Dirección Nacional de Información y Evaluación de la Calidad Educativa (DINIECE) y tenían el objetivo de evaluar los logros de las y los estudiantes de 3° y 6° grados de primaria y de 2°/3° y 5°/6° años de secundaria, teniendo en cuenta la estructura vigente en cada jurisdicción (Secretaria de Evaluación e Información Educativa, 2019).

Los ONE consistieron en la administración de pruebas de Lengua, Matemática, Ciencias Sociales y Ciencias Naturales que, en un principio, se fundamentaron en los Contenidos Básicos Comunes (CBC). Posteriormente, se desarrollaron en base a los Núcleos de Aprendizaje Prioritarios (NAP) los diseños curriculares de cada jurisdicción, los acuerdos con las provincias, los resultados de estudios piloto y la literatura específica relativa a los dominios y temas evaluados. A su vez, en la mayoría de las instancias en las que fueron aplicados los ONE, se administraron cuestionarios complementarios cuyo objetivo era recabar información sobre factores escolares (ej. trayectoria escolar) y extraescolares (ej. nivel socioeconómico, nivel educativo de los padres) que permitiesen evaluar su asociación con el desempeño de los estudiantes. Estos cuestionarios fueron aplicados a estudiantes, docentes y equipos directivos.

Todas estas evaluaciones son una prueba estandarizada. Esto significa que se aplican los mismos instrumentos a todos/as los/as estudiantes del mismo grado o año, en las mismas condiciones, para luego valorar los resultados con los mismos criterios. Este tipo de evaluación se realiza de acuerdo a normas que garantizan la homogeneidad del proceso evaluativo y de los resultados obtenidos. Las pruebas estandarizadas suelen ser las referidas a normas o referidas a criterio (Richaud de Minzi, 2008):

- En las pruebas (o tests) referidas a normas (TRN) la medida de desempeño es relativa al conjunto evaluado. Esto significa que se enfocan en la comparación entre estudiantes o entre grupos de estudiantes. Por lo tanto, se omite la referencia a objetivos o logros esperados. La valoración es en comparación con el desempeño promedio del grupo.
- En cambio, los tests referidos a criterio (TRC) se definen en relación a la relevancia y representatividad de los ítems respecto al dominio específico, por lo que la validez del contenido es fundamental. Privilegian la comparación de los logros de las y los estudiantes con respecto a las metas de aprendizaje o a las competencias que el sistema educativo persigue que alcancen. Las puntuaciones tienen carácter absoluto y están en relación al dominio medido en la prueba. En este sentido, sirven para retroalimentar y monitorear el progreso de las y los estudiantes o del sistema. El propósito es conocer y tomar decisiones sobre: (a) si los individuos alcanzan o no el dominio o competencia evaluada; y (b) determinar la eficacia de programas y sistemas educativos.

En el caso de las encuestas Aprender es una prueba referida a criterio (TRC). (Secretaria de Evaluación e Información Educativa, 2019). A su vez, los ítems de las pruebas correspondientes a Aprender se analizaron con base en la Teoría de Respuesta al Ítem (TRI), modelo general sobre el cual se basan la mayoría de las evaluaciones estandarizadas internacionales, así como también los ONE desde 2005.

El modelo TRI asigna un puntaje de ‘competencia’ (θ) a cada estudiante, en base a las respuestas a un conjunto de ítems. Este puntaje θ es un número real (positivo, cero o negativo) y cada ítem se presupone que posee (por ejemplo, en el modelo a dos parámetros) dos números

reales que lo caracterizan, uno que mide su dificultad y otro, su discriminación. Los supuestos básicos del modelo de dos parámetros son:

- Para cada área/disciplina evaluada en cierto año, el/la estudiante posee una habilidad, rasgo latente (no observado), competencia, etc. que puede asociarse con un número real θ.
- La probabilidad de responder correctamente a un cierto ítem i (según su dificultad) es una función creciente y continua del valor θ.

2.2 Antecedentes

El ministerio de Educación realiza anualmente informes sobre los resultados de las encuestas Aprender (Ministerio de Educación, 2019). Estos informes se basan en análisis estadísticos, y tienen un enfoque integral del nivel secundario, considerando tanto las condiciones y los recursos con los que se desarrollan los procesos pedagógicos, así como las características de la población estudiantil, las trayectorias educativas y los resultados de los aprendizajes del último año del nivel.

Entonces, esta Encuesta pone a disposición variadas evidencias sobre la situación de la educación secundaria en Argentina, tomando como principales dimensiones la situación social y familiar de la población adolescente, los recursos y condiciones de las escuelas, el acceso a la educación, las trayectorias escolares y la graduación y los niveles de aprendizaje alcanzados (Secretaria de Evaluación e Información Educativa, 2019). Estos datos son útiles para construir una visión global sobre la educación, por lo que el ministerio de Educación disponible de múltiples análisis realizados a partir de los resultados de las encuestas Aprender (Ministerio de Educación).

Sin embargo, el proyecto propuesto en este trabajo se complementa con los informes que se mencionan dado que se considera pertinente la aplicación de técnicas de Explotación de Información para la generación de un Modelo Predictivo, a modo de obtener información más precisa, al detectar patrones y correlaciones entre los datos disponibles.

2.3 Bases de datos

Una base de datos se puede definir como:

> *"Una colección o depósito de datos integrados, almacenados en soporte secundario (no volátil) y con redundancia controlada. Los datos, que han de ser compartidos por diferentes usuarios y aplicaciones, deben mantenerse independientes de ellos y su definición, única y almacenada junto con los datos, se ha de apoyar en un modelo de datos, el cual ha de permitir captar las interrelaciones y restricciones existentes en el mundo real. Los procedimientos de actualización y recuperación, comunes y bien determinados, facilitarán la seguridad del conjunto de los datos*" (Piattini & De Miguel, 1999).

Las bases de datos permiten mejorar la calidad de las prestaciones de los sistemas informáticos y aumentar su rendimiento a través de las siguientes ventajas:

- Independencia de los datos y los programas y procesos. Esto permite modificar los datos sin modificar el código de las aplicaciones.
- Menor redundancia. No hace falta tanta repetición de datos. Sólo se indica la forma en la que se relacionan los datos.
- Integridad de los datos. Mayor dificultad de perder los datos o de realizar incoherencias con ellos.
- Coherencia de los resultados. Al recogerse y almacenarse la información una sola vez, en los tratamientos se utilizan siempre los mismos datos, por lo que los resultados son coherentes.
- Mayor seguridad en los datos. Al permitir limitar el acceso a los usuarios. Cada tipo de usuario podrá acceder a unas cosas.
- Datos más documentados. Gracias a los metadatos que permiten describir la información de la base de datos.
- Acceso a los datos más eficiente. La organización de los datos produce un resultado más óptimo en rendimiento.
- Reducción del espacio de almacenamiento. Gracias a una mejor estructuración de los datos.

- Acceso simultáneo a los datos. Es más fácil controlar el acceso de usuarios de forma concurrente.

2.4 Explotación de Información y Minería de Datos

Según Mitra & Acharya (2003), la revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir y transmitir. Con el importante progreso en informática y en las tecnologías relacionadas y la expansión de su uso en diferentes aspectos de la vida, se continúa recogiendo y almacenando en bases de datos gran cantidad de información. Por lo que descubrir conocimiento de este enorme volumen de datos es un reto en sí mismo.

La Explotación de Información es la subdisciplina informática que aporta a la Inteligencia de Negocio (Negash & Gray, 2008) las herramientas para la transformación de información en conocimiento (Langseth & Vivatrat, 2003). Por ello, se ha definido como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información (Grigori, y otros, 2004). Al hablar de Explotación de Información basada en sistemas inteligentes (Michalski, Bratko, & Kubat, 1998) se refiere específicamente a la aplicación de métodos de Minería de Datos, para descubrir y enumerar patrones presentes en la información. Dichos métodos (Kononenko & Cestnik, 1986) permiten obtener resultados de análisis de la masa de información que los métodos convencionales (Michalski R. , 1983) no logran.

La idea de la Minería de Datos no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como “Data Fishing”, “Data Mining” o “Data Archaeology” con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido.

La evolución de sus herramientas en el transcurso del tiempo puede dividirse en cuatro etapas principales:

- Colección de Datos (1960).
- Acceso de Datos (1980).

- Almacén de Datos y Apoyo a las Decisiones (principios de la década de 1990).
- Minería de Datos Inteligente (a partir de finales de la década de 1990).

De hecho, las técnicas de Ciencia de Datos ("Data Science" o "Data Analytics"), que tanto interés despiertan hoy en día, en realidad surgieron en la década de los 90, cuando se usaba el término KDD (acrónimo de "Knowledge Discovery in Databases" o "Descubrimiento de Conocimientos en Bases de Datos") para referirse al (amplio) concepto de hallar conocimiento en los datos. A principios de los años ochenta, Rakesh Agrawal, GioWiederhold, Robert Blum y Gregory Piatetsky-Shapiro entre otros, empezaron a consolidar los términos de Minería de Datos y KDD.

Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios Algunas de las tareas importantes de la minería de datos incluyen la identificación de aplicaciones para las técnicas existentes, y desarrollar nuevas técnicas para dominios tradicionales o de nueva aplicación, como el comercio electrónico y la bioinformática. Existen numerosas áreas donde la minería de datos se puede aplicar, siendo un gran referente en las siguientes especializaciones: (Riquelme, Ruiz, & Gilbert, 2006):

- Comercio y Banca: segmentación de clientes, previsión de ventas, análisis de riesgo.
- Medicina y Farmacia: diagnóstico de enfermedades y la efectividad de los tratamientos.
- Seguridad y detección de fraude: reconocimiento facial, identificaciones biométricas, accesos a redes no permitidos, etc.
- Recuperación de información no numérica: minería de texto, minería web, búsqueda e identificación de imagen, video, voz y texto de bases de datos multimedia.
- Astronomía: identificación de nuevas estrellas y galaxias.
- Geología, minería, agricultura y pesca: identificación de áreas de uso para distintos cultivos o de pesca o de explotación minera en bases de datos de imágenes de satélites

- Ciencias Ambientales: identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales (p.e. plantas depuradoras de aguas residuales) para mejorar su observación, gestión y/o control.
- Ciencias Sociales: Estudio de los flujos de la opinión pública. Planificación de ciudades: identificar barrios con conflicto en función de valores sociodemográficos.

En la actualidad se puede afirmar que la Minería de Datos ha demostrado la validez de una primera generación de algoritmos mediante diferentes aplicaciones al mundo real.

Existen cientos de productos de minería de datos y de compañías de consultoría. KDNuggets (www.kdnuggets.com) tiene una lista de estas compañías y sus productos en el campo de la minería de datos. Pueden resaltarse por su mayor expansión las siguientes: SAS con SAS Script y SAS Enterprise Miner; SPSS y el paquete de minería Clementine; IBM con Intelligent Miner; Microsoft incluye características de minería de datos en las bases de datos relacionales; otras compañías son Oracle, Angoss y Kxen. En la línea del software libre Weka (Witten & Frank, 2005) es un producto con mayor orientación a las técnicas provenientes de la IA (Inteligencia Artificial), pero de fuerte impacto.

Sin embargo, estas técnicas todavía están limitadas por bases de datos simples, donde los datos se describen mediante atributos numéricos o simbólicos, no conteniendo atributos de tipo texto o imágenes, y los datos se preparan con una tarea concreta en mente. Sobrepasar este límite será un reto a conseguir (Riquelme, Ruiz, & Gilbert, 2006).

2.5 Procesos y Metodologías para Minería de Datos

Como se ha mencionado anteriormente, la Minería de Datos surgió a principio de los años 70 y fue evolucionando y volviéndose más grande. En un intento de normalización de este proceso KDD, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de los 90 dos metodologías principales: “Sample, Explore, Modify, Model, and Assess” (SEMMA) y “Cross Industry Standard Process for Data Mining” (CRISP-DM). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase. De

esta manera, permiten contar con un marco de trabajo que permita planificar y guiar el proceso de desarrollo del proyecto.

A continuación se va a detallar los pasos de cada uno de estos proceso y metodologías.

2.5.1 Knowledge Discovery in Databases (KDD)

Un término común en la minería de datos es el Descubrimiento de Conocimiento en Bases de Datos (KDD), que es un proceso iterativo significativo que consta de una serie de fases para la generación de conocimiento y la toma de decisiones (Hasperué, 2013), como se puede apreciar en la figura 1.

Figura 1. Proceso KDD (Knowledge Discovery from Databases)

Las fases de KDD son:

- **Integración y recopilación:** Consiste en establecer un entendimiento del dominio de la aplicación y de los conocimientos previos relevantes. En esta fase se determina también la selección de un conjunto de datos que pueden ser obtenidos de diferentes fuentes, sobre los cuales se realiza el descubrimiento.

- **Selección, limpieza y transformación**: En esta etapa se seleccionan y preparan los datos que se van a minar. Sin embargo, existen factores como el ruido o valores atípicos que afectan la calidad de los datos, por lo que ante esta situación la limpieza es una de las tareas más importantes, puesto que permite la selección de la técnica que más se ajuste al problema a resolver.
- **Minería de datos:** Es la fase más representativa, se determina qué tipo de tarea es la más apropiada, ya sea agrupamiento, reglas de asociación, correlación, clasificación, regresión, entre otras. Los resultados obtenidos dependen de fases anteriores, por lo que existe la posibilidad de regresar a los pasos previos para requerir nuevos datos o para redefinir la solución al problema planteado.
- **Evaluación e interpretación:** Los patrones descubiertos deben cumplir con tres propiedades: precisión, comprensibles e interesantes. En esta fase se evalúan e interpretan los patrones obtenidos. Algunas validaciones pueden ser a través de índices de evaluación, validación cruzada, matrices de confusión, entre otras.

5. Difusión y uso: Como última fase, el conocimiento descubierto debe de ser incorporado en algún sistema o simplemente documentarlo para su difusión a las partes interesadas. Este proceso incluye también la revisión y resolución de posibles conflictos con los conocimientos que anteriormente se tenían.

2.5.2 SEMMA

SEMMA (Sample, Explore, Modify, Model, Assess) es una metodología creada por Statistical Analysis Systems Institute (SAS Institute), quien la define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de interés (SAS, 1998). Este proceso consta de cinco etapas necesarias para guiar el desarrollo de un proyecto de minería de datos. Sumathi y Sivananda (2006) mencionan que la metodología SEMMA permite aplicar la estadística exploratoria y técnicas de visualización de manera fácil, así como la selección y transformación de las variables más significativa, con el objetivo de crear modelos para predecir resultados y evaluarlos de manera que sirva de apoyo para la toma de decisiones.

Las etapas de SEMMA (figura 2) son:

- **Muestreo**: En esta etapa se toma una muestra del conjunto de datos disponible, que debe ser lo suficientemente grande para contener la información relevante, y lo suficientemente pequeña como para correr el proceso rápidamente. Esta etapa es aconsejable cuando el tamaño del conjunto de datos es demasiado extenso.
- **Exploración:** Consiste en explorar los datos en búsqueda de relaciones y tendencias desconocidas. Es una etapa especial para familiarizarse con los datos y formular nuevas hipótesis a partir de su análisis.
- **Modificación:** Etapa de preparación de datos que consiste en la limpieza de los valores atípico, se realiza un tratamiento de los datos faltantes y se seleccionan, crean y modifican las variables que servirán para la etapa del modelado.
- **Modelado:** Consiste en la creación del modelo para predecir las variables, utilizando algunas de las técnicas predictivas como árboles de decisión, redes neuronales, análisis discriminante o análisis de regresión.
- **Evaluación:** En esta fase se evalúan la utilidad y la exactitud de los modelos obtenidos en el proceso de minería de datos, por ejemplo analizando la capacidad predictiva de los mismos.

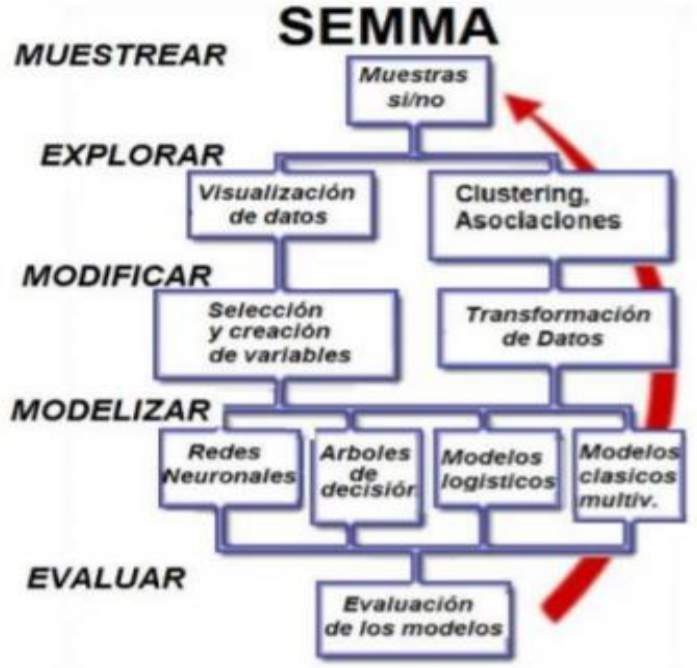

Figura 2. Diagrama de etapas de SEMMA

Una clara diferencia con respecto a otras metodologías es que en SEMMA la primera fase se inicia con el muestreo de datos. Por otra parte, SEMMA está relacionada particularmente con productos comerciales de SAS Institute.

2.5.3 CRISP-DM

La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico. (Chapman, y otros, 2000) afirma:

> *"Consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos. "*

Tal como se puede observar en la Figura 3, en el nivel superior, el proceso de minería de datos es organizado en un número de fases donde cada fase consiste en varias tareas genéricas de segundo nivel. Este segundo nivel lo denominan genérico porque está destinado a ser bastante

general para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas están destinadas a ser tan completas y estables como sea posible. Completo significa que cubre tanto al proceso entero de minería de datos como a todas las aplicaciones de minería de datos posibles. Estable significa que el modelo debería ser válido para acontecimientos normales y aún para desarrollos imprevistos como técnicas de modelado nuevo.

El tercer nivel, el nivel de tarea especializado, es el lugar para describir como las acciones en las tareas genéricas deberían ser realizadas en ciertas situaciones específicas. Por ejemplo, en el segundo nivel podría haber una tarea genérica llamada limpieza de datos. El tercer nivel describe cómo esta tarea se diferencia en situaciones diferentes, como la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es agrupamiento o el modelado predictivo.

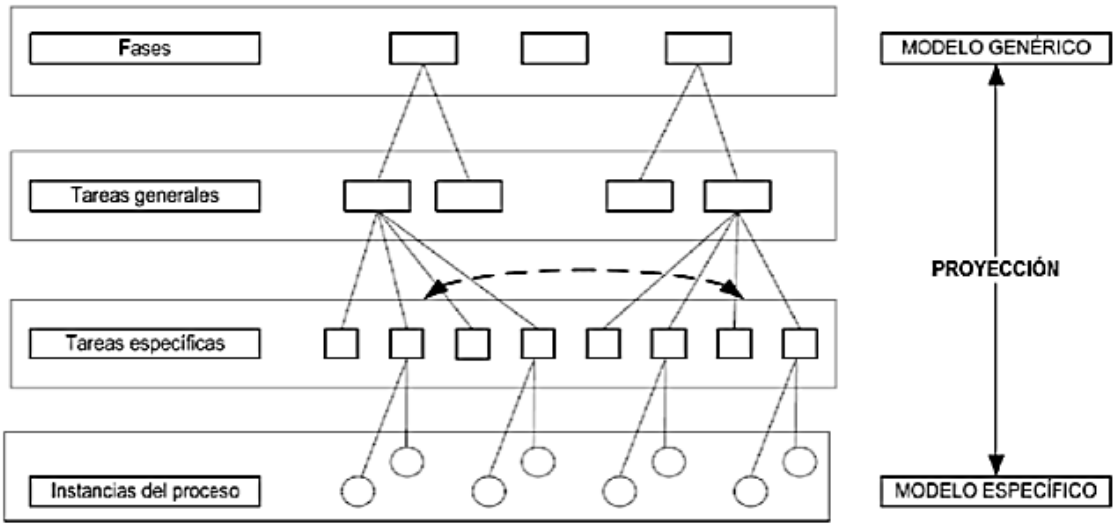

Figura 3 Desglose de 4 niveles en la metodología CRISP-DM (Chapman, y otros, 2000)

La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos.

En la práctica, muchas de las tareas pueden ser realizadas en una orden diferente, y a menudo será necesario volver a hacer tareas anteriores repetidamente y repetir ciertas acciones. Nuestro

modelo de proceso no intenta capturar todas estas posibles rutas del proceso de la minería de datos porque esto requeriría un modelo de proceso demasiado complejo.

El cuarto nivel, la instancia de proceso es un registro de las acciones, decisiones y de los resultados de una minería de datos real contratada.

Una instancia de proceso está organizado según las tareas definidas en los niveles más altos, pero representa lo que en realidad pasó en un contrato particular más bien que lo que pasa en general.

Por otro lado, el modelo de procesos de CRISP-DM trae los siguientes beneficios que afirma Chapman (2000):

> *"Nos proporciona una descripción del ciclo de vida del proyecto de minería de datos. Este contiene las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción, no es posible identificar todas las relaciones. Las relaciones podrían existir entre cualquier tarea de minería de datos según los objetivos, el contexto, y –lo más importante- el interés del usuario sobre los datos."*

Respecto al ciclo de vida del modelo CRISP-DM, se divide en seis fases, mostradas en la Figura 4. La secuencia de las fases no debe ser necesariamente rígida. Se debe avanzar y retroceder entre fases. El resultado de cada fase determina que la fase o la tarea particular de una fase, tienen que ser realizadas después. Las flechas indican las más importantes y frecuentes dependencias entre fases (Chapman, y otros, 2000)

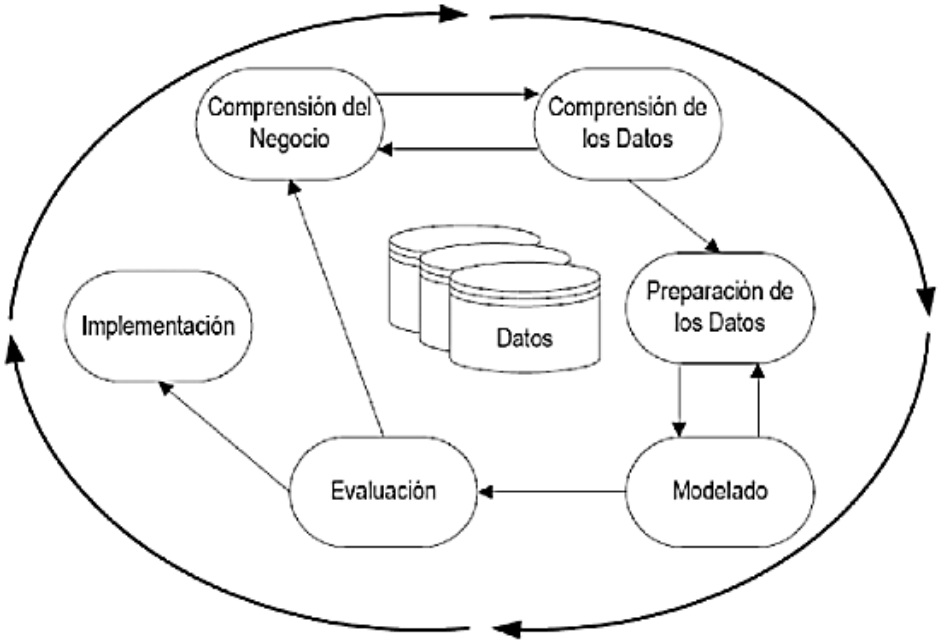

Figura 4 Fases del modelo de referencia CRISP-DM (Chapman, y otros, 2000)

El círculo externo en la Figura 4 simboliza la naturaleza cíclica de la minería de datos. La minería de datos no se termina una vez que la solución es desplegada. Las informaciones ocultas (lecciones cultas) durante el proceso y la solución desplegada pueden provocar nuevas, a menudo más - preguntas enfocadas en el negocio. Los procesos de minería subsecuentes se beneficiarán de las experiencias previas (Chapman, y otros, 2000)

A continuación, se perfilan las 6 fases previamente mencionadas, según Chapman (2000):

- **Comprensión del negocio.** Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos. Esta fase incluye las tareas que se indican a continuación en la Tabla 1:

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
1.1 Determinar los objetivos de negocio	• 1.1.1 Antecedentes • 1.1.2 Objetivos de negocio • 1.1.3 Criterios de éxito del negocio

1.2 Evaluar la situación	• 1.2.1 Evaluar la situación • 1.2.2 Requisitos, supuestos y limitaciones • 1.2.3 Riesgos y contingencias • 1.2.4 Terminología • 1.2.5 Costos y beneficios
1.3 Determinar objetivos explotación de información	• 1.3.1 Objetivos de explotación de información • 1.3.2 Criterios de éxito de la de explotación de información
1.4 Producir el plan del proyecto	• 1.4.1 Plan del proyecto • 1.4.2 Evaluación inicial de herramientas y técnicas

Tabla 1. Tareas Comprensión del negocio - CRISP-DM

- **Comprensión de los datos:** La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta. Esto se realiza aplicando las tareas que se indican en la Tabla 2:

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
2.1 Recolección inicial de datos	• 2.1.1 Informe inicial de recopilación de datos
2.2 Descripción de los datos	• 2.2.1 Informe de descripción de datos
2.3 Exploración de los datos	• 2.3.1 Informe de exploración de datos
2.4 Verificación de calidad de los datos	• 2.4.1 Informe de calidad de datos

Tabla 2. Tareas Comprensión de los datos - CRISP-DM

- **Preparación de datos:** La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [los datos que serán provistos en las herramientas de modelado] de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan. Las tareas de esta fase son las mostradas en la Tabla 3:

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
3.0 Tareas preparatorias	• 3.0.1 Conjunto de datos • 3.0.2 Descripción del conjunto de datos
3.1 Selección de datos	• 3.1.1 Justificación de la inclusión / exclusión
3.2 Limpieza de datos	• 3.2.1 Informe de limpieza de datos
3.3 Construcción de datos	• 3.3.1 Atributos derivados • 3.3.2 Registros generados
3.4 Integración de los datos	• 3.4.1 Datos combinados

Tabla 3. Tareas Preparación de datos - CRISP-DM

- **Modelado:** En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario. Esto se logra a través de las tareas indicadas en la Tabla 4:

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
4.1 Selección de la técnica de modelado	• 4.1.1 Técnica de Modelado • 4.1.2 Supuestos del modelado
4.2 Generación del diseño del ensayo	• 4.2.1 Prueba de diseño
4.3 Construcción del modelo	• 4.3.1 Configuración de parámetros • 4.3.2 Modelos • 4.3.3 Descripción del modelo
4.4 Evaluación del modelo	• 4.4.1 Evaluación del modelo • 4.4.2 Revisión de la configuración de parámetros

Tabla 4. Tareas Modelo - CRISP-DM

- **Evaluación:** En esta etapa en el proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos. Las tareas de la fase 'Evaluación' (Tabla 5) son:

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
5.1 Evaluar los resultados	• 5.1.1 Evaluación de los resultados de la explotación de información con respecto a los criterios de éxito del negocio • 5.1.2 Modelos aprobados
5.2 Proceso de revisión	• 5.2.1 Revisión del proceso
5.3 Determinación de los próximos pasos	• 5.3.1 Lista de posibles acciones • 5.3.2 Decisión

Tabla 5. Tareas Evaluación - CRISP-DM

- **Despliegue:** La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización, por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos de mercadeo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue,

esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente. En la Tabla 6 se pueden ver las tareas en esta fase:

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
6.1 Plan de implantación	• 6.1.1 Ejecución del plan de implantación
6.2 Plan de vigilancia y mantenimiento	• 6.2.1 Ejecución del plan de monitoreo y mantenimiento
6.3 Producción final	• 6.3.1 Informe final • 6.3.2 Presentación final
6.4 Revisión del proyecto	• 6.4.1 Documentación de la experiencia

Tabla 6. Tareas Despliegue - CRISP-DM

Si se comparan SEMMA con CRISP-DM, se puede destacar las conclusiones de Rodríguez Montequín, Álvarez Cabal, Mesa Fernández, González Valdés (2003):

> *"Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de Data Mining en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Data Mining en un proceso iterativo e interactivo. La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de Data Mining donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista más global se puede considera que la metodología CRISP-DM está más cercana al concepto real de proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas."*

Por otra parte, según la encuesta realizada por (KDnuggets, 2014), CRISP-DM, es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Minería de Datos, como se puede ver en el gráfico de la Figura 5. Esta figura, a su vez, compara los resultados con los realizados durante una encuesta de 2007, con la misma finalidad. Esta supremacía se mantiene

desde el año 2002 y se debe, entre otras razones, a que es de libre distribución (sin costo alguno) y se considera que es la metodología independiente del dominio más efectiva, dado que su alcance incluye todas las complejidades del proyecto a través de tareas fáciles de aplicar.

Sin embargo, aunque se la considera confiable y robusta, entre las principales críticas que se le han realizado, se destaca el hecho que CRISP-DM define qué hacer, pero no cómo hacerlo (Mariscal, Marbán, González, & Segovia, 2007). Es decir, indica los entregables a ser preparados por cada tarea, pero no formula ningún tipo de técnica o método específico para realizarla. Esto hecho tiene como consecuencia que muchos equipos de trabajo terminen utilizando adaptaciones y/o metodologías propias.

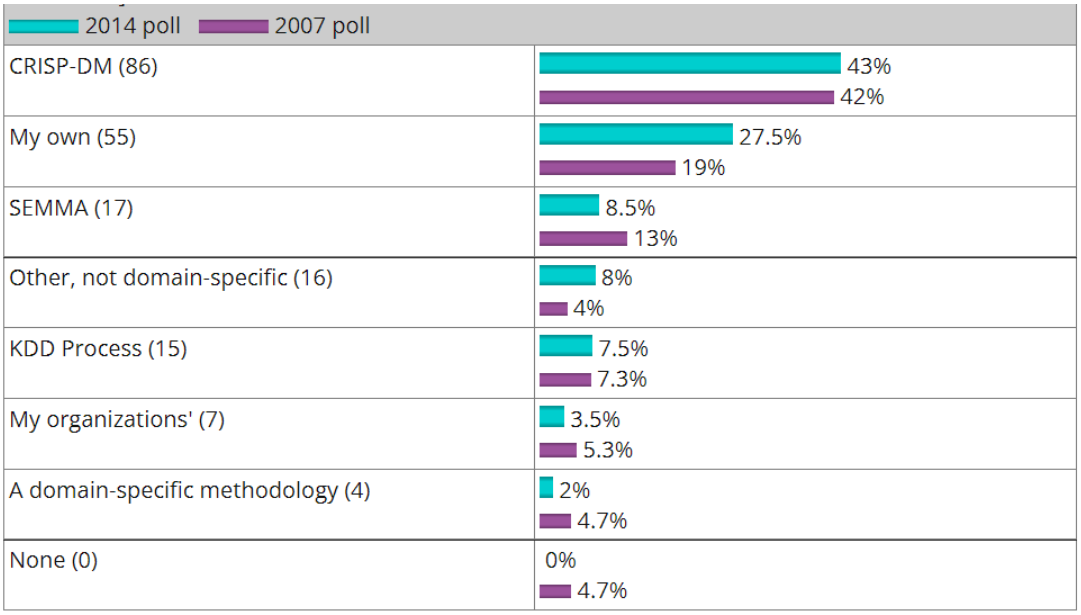

Figura 5. Metodologías más utilizadas para proyectos de Explotación de Información

2.6 Herramientas

En esta sección, se enumeran y describen brevemente las herramientas utilizadas durante el proyecto

2.6.1 Python

Python (Python Software Foundation. Python Language Reference, version 2.7, s.f.) es un lenguaje de escritura rápido, escalable, robusta y de código abierto, ventajas que hacen de Python un aliado perfecto para la Inteligencia Artificial (Soloaga, Principales Usos de Python, 2018)

Así, permite plasmar ideas complejas con unas pocas líneas de código, lo que no es posible con otros lenguajes. Existen bibliotecas como "Keras" (www.keras.io) y "TensorFlow" (www.tensorflow.org), que contienen mucha información sobre las funcionalidades del aprendizaje automático. Además, existen bibliotecas proporcionadas por Python, que se usan mucho en los algoritmos de Inteligencia Artificial como "Scikitlearn" (Pedregosa, y otros, 2011), una biblioteca gratuita de aprendizaje automático que presenta varios algoritmos de regresión, clasificación y agrupamiento. (Soloaga, 2018)

Pero, sobre todo, Python es un lenguaje gratuito de código abierto con una gran comunidad en activo, que proporciona soporte a cualquier programador. Todas estas razones combinadas hacen que aprender Python sea una opción fácil sobre otros lenguajes para aplicaciones de inteligencia artificial (Soloaga, Principales Usos de Python, 2018)

Dentro del lenguaje, se utilizaron las siguientes librerías:

- *Pandas* (The pandas development team, 2020)*:* es una de las librerías de python más útiles para los científicos de datos. Las estructuras de datos principales en pandas son Series para datos en una dimensión y DataFrame para datos en dos dimensiones. Estas son las estructuras de datos más usadas en muchos campos tales como finanzas, estadística, ciencias sociales y muchas áreas de ingeniería. Pandas destaca por lo fácil y flexible que hace la manipulación de datos y el análisis de datos.

- *NumPy* (Harris, Millman, van der Walt, & y otros, 2020)*:* proporciona una estructura de datos universal que posibilita el análisis de datos y el intercambio de datos entre distintos algoritmos. Las estructuras de datos que implementa son vectores multidimensionales y matrices con capacidad para gran cantidad de datos. Además,

esta librería proporciona funciones matemáticas de alto nivel que operan en estas estructuras de datos Seaborn

- *Matplotlib* (Hunter, 2007)*:* es una librería de gráficos, desde histogramas hasta gráficos de líneas o mapas de calor. También se pueden usar comandos de Latex para agregar expresiones matemáticas a tu gráfica.

- *Seaborn* (Waskom, 2021)*:* es una librería gráfica basada en matplotlib, especializada en la visualización de datos estadísticos. Se caracteriza por ofrecer un interfaz de alto nivel para crear gráficos estadísticos visualmente atractivos e informativos. Seaborn considera la visualización como un aspecto fundamental a la hora de explorar y entender los datos. Se integra muy bien con la librería de manipulación de datos pandas.

- *Sklearn* (Pedregosa, y otros, 2011)*:* Construida sobre NumPy, SciPy y matplotlib, esta librería contiene un gran número de eficientes herramientas para machine learning y modelado estadístico, como por ejemplo, algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad.

2.6.2 Jupyter Notebook

Jupyter Notebook (Kluyver, 2016) es una interfaz web de código abierto que permite la inclusión de texto, vídeo, audio, imágenes, así como la ejecución de código a través del navegador en múltiples lenguajes. Esta ejecución se realiza mediante la comunicación con un núcleo (Kernel) de cálculo. Aunque en principio, el equipo de desarrollo de Jupyter Notebook incluye por defecto únicamente el núcleo de cálculo Python, el carácter abierto del proyecto ha permitido aumentar el número de núcleos disponibles, incluyendo, por ejemplo, Octave, Julia, R, Haskell, Ruby, C/C++, Fortran, Java, SageMath, Scala, o también Matlab y Mathematica. Esta interfaz, agnóstica del lenguaje (de ahí su nombre al unir 3 de los lenguajes de programación de código abierto más utilizados en el ámbito científico: Julia, Python y R), puede suponer por tanto una estandarización para mostrar el contenido de cursos científicos, sin encontrarse limitado a la adopción de un único lenguaje. (Granado, s.f.)

2.7 Proceso de Modelado

Con el objetivo de convertir los datos disponibles en información útil y valiosa, la Minería de Datos puede utilizar dos tipos de procesos, los cuales se denominan como Analítica Descriptiva o Predictiva.

Por un lado, la Analítica Descriptiva examina los datos y analiza los acontecimientos pasados. De esta manera, se examina y entiende el rendimiento pasado al extraer datos históricos para buscar las razones detrás del éxito o el fracaso del pasado. Casi todos los informes de gestión, tales como ventas, marketing, operaciones y finanzas, utilizan este tipo de análisis post-mortem. (Witten & Frank, 2005).

Por otro lado, el Análisis Predictivo utiliza datos para determinar el resultado futuro probable de un evento o la probabilidad de que se produzca una situación. Es decir que considera situaciones pasadas para saber cómo abordar el futuro. En otras palabras, busca predecir las salidas y revelar relaciones en los datos (Mitra & Acharya, 2003). Para ello se utilizan herramientas automáticas que

- Emplean algoritmos sofisticados para descubrir principalmente patrones ocultos, asociaciones, anomalías, y/o estructuras de la gran cantidad de datos almacenados en los data warehouses u otros repositorios de información, y
- Filtran la información necesaria de las grandes bases de datos (Riquelme, Ruiz, & Gilbert, 2006)

Entonces, los Modelos Predictivos buscan predecir el comportamiento futuro de una entidad aprendiendo de los hechos ocurridos en el pasado. Un modelo predictivo se desarrolla con datos históricos para luego, al aplicar el algoritmo definido, intentar predecir lo que sucederá en el futuro. Por lo que el modelo contiene una variable a predecir (Variable Target) y un conjunto de variables predictoras (Variables Input).

En el caso del modelo que se está desarrollando en este trabajo se utilizarán bases de estudiantes que hayan desaprobado el examen de Matemáticas (Target 1) y clientes que lo hayan aprobado (Target 0). Una vez entrenado el modelo, y que este haya aprendido, se van a poder detectar con anticipación aquellos estudiantes con dificultades en Matemática, descubriendo los perfiles más propensos a desaprobar el examen.

El proceso de Modelado se divide en varias etapas a continuación detalladas:

- **Primera etapa:** Esta etapa es una de las más importantes en el proceso y consiste en el armado de la base necesaria para entrenar el modelo. Se construyen las variables consideradas necesarias para modelar del periodo de observación. El objetivo es generar la mayor cantidad de variables posibles para que el algoritmo pueda seleccionar las de mayor poder explicativo del fenómeno a predecir. En la sección 2.7 se detallan algunas técnicas de limpieza que se utilizarán en el proyecto.

- **Segunda etapa:** Consta de la generación de la Variable Target en el periodo de predicción. Para esto se definen los criterios de Aprobación y desaprobación. En el caso de desaprobación la Variable Target va a tomar el valor 1 y en el caso contrario, 0.

- **Tercera etapa:** Una vez construida la base y teniendo a cada estudiante con un valor en la Variable Target comienza la fase de entrenamiento donde se comienza con un tratamiento de los datos, una selección de variables para descartar algunas variables con poco poder predictivo y luego, después de probar distintos algoritmos de predicción, se selecciona el que mejor se adecua a nuestro objetivo. En este caso se van a probar los algoritmos de Regresión Logística, Redes Bayesianas, Clasificador KNN, Arboles de decisión, y Random Forest. En la sección 2.8 se detallan éstos para el modelo en análisis.

- **Cuarta Etapa:** Esta última es la llamada Etapa de Validación. Una vez elegido el modelo "Ganador" se realiza una validación atemporal, lo que significa aplicar el modelo a estudiantes diferentes a los utilizados para la etapa de entrenamiento y se analiza la estabilidad del mismo. Para dar por concluido el proceso, estas pruebas tienen que tener como resultado una performance parecida a la de entrenamiento.

2.8 Técnicas de Limpieza, transformación y construcción de Datos

En esta sección, se introducen las técnicas de limpieza y transformación que se utilizan a lo largo del proyecto.

2.8.1 Tratamiento de Datos Atípicos o Outliers

Se denominan casos atípicos u "outliers" a aquellas observaciones con características diferentes de las demás. Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos, sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar (Hair, Anderson, Tathman, & Black, 1999).

Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento valido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra (Hair, Anderson, Tathman, & Black, 1999).

En estos casos, se debe analizar primero la causa de generación del outlier. En caso de ser un error de formato, se pueden transformar para adaptar al resto de los valores. Por el contrario, si no se puede obtener la causa del outlier, se debe tratar similar a los datos ausentes o "missing" (Riquelme, Ruiz, & Gilbert, 2006), los cuales son explicados a continuación.

2.8.2 Tratamiento de Datos Ausentes o Missing

Los datos ausentes o “missing” son algo habitual, de hecho, rara es la investigación en la que no aparece este tipo de datos. En estos casos la ocupación primaria del investigador debe ser determinar las razones que subyacen en el dato ausente buscando entender el proceso principal de esta ausencia para seleccionar el curso de acción más apropiado. Para ello se debe determinar cuál es el motivo que generó la presencia datos ausentes, entendido como cualquier evento sistemático externo al encuestado (errores en la introducción de datos) o acción por parte del encuestado (tales como rehusar a contestar) que da lugar a la ausencia de datos (Hair, Anderson, Tathman, & Black, 1999). Una vez establecido esto, se pueden realizar sustituciones por métodos estadísticos. Como, por ejemplo, reemplazando estos valores por la media, la mediana, el mínimo o el máximo. Otra posible opción es borrar el registro o el atributo completo cuando la cantidad de valores faltantes es superior a la de los valores definidos.

2.8.3 Utilización de Variables Categóricas, Nominales o Dummies

Los datos categóricos o nominales, como su nombre lo indica, son usados para nombrar o categorizar información. Este tipo de dato se caracteriza por no ser ordenado, incluso si se usan números para representarlos. (Riquelme, Ruiz, & Gilbert, 2006)

El nombre de las diferentes provincias de Argentina es un dato categórico. Aunque puedes ordenar todos los nombres alfabéticamente, carece de sentido conceptual que “Buenos Aires” se encuentre antes que “Corrientes”, o que “Misiones” se encuentre después de “Chubut”. Esto no nos informa ni nos ayuda a entender sobre las características de estas provincias.

La manera más sencilla de transformar estos datos es crear variables “dummy” (falsas, en español)

Crear variables dummy implica transformar datos de un formato “alto”, en el que cada columna contiene la información de una variable, a datos con un formato “ancho”, en los que múltiples columnas contienen la información de las dos variables, codificada de manera binaria, esto es, con 0 y 1. (Hernandez, Ramirez Quintana, & Ferri Ramirez, 2004)

En el caso de las provincias, podríamos generar variables del estilo “Provincia_Bs_as” que tendrá valor 1 si el individuo pertenece a esta provincia, o 0 en caso de no pertenecer. Lo mismo se aplicaría para el resto de las provincias.

2.8.4 Estandarización de Valores Numéricos

Estandarizar un vector generalmente significa restar una medida de ubicación y dividir por una medida de escala. Por ejemplo, si el vector contiene valores aleatorios con una distribución gaussiana, puede restar la media y dividir por la desviación estándar, obteniendo así una variable aleatoria "normal estándar" con media 0 y desviación estándar 1. (Riquelme, Ruiz, & Gilbert, 2006)

La estandarización de las características alrededor del centro y 0 con una desviación estándar de 1 es importante cuando comparamos medidas que tienen diferentes unidades. Las variables que se miden a diferentes escalas no contribuyen por igual al análisis y podrían terminar creando un problema. (Riquelme, Ruiz, & Gilbert, 2006)

Por ejemplo, una variable que oscila entre 0 y 1000 tendrá más peso que una variable que oscila entre 0 y 1. El uso de estas variables sin estandarización dará a la variable con el rango más grande un peso de 1000 en el análisis. Transformar los datos a escalas comparables puede evitar este problema. Los procedimientos típicos de estandarización de datos igualan el rango y / o la variabilidad de los datos.

Python contiene múltiples librerías que nos permiten realizar este proceso de manera simple. Una de las más utilizadas es la función “StandardScaler” de Sklearn (Pedregosa, y otros, 2011). Esta librería nos permita realizar la estandarización eliminando la media y escalando a la varianza de la unidad.

La puntuación estándar de una muestra X se calcula como:

$$Z = \frac{X - U}{S}$$

Donde U es la media de las muestras de entrenamiento, y S es la desviación estándar de las muestras de entrenamiento.

2.9 Algoritmos de Clasificación

A continuación, se introducen y explican los distintos algoritmos que van a ser utilizados en el proyecto.

2.9.1 Regresión Logística

La Regresión Logística es una técnica analítica que nos permite relacionar funcionalmente una variable dicotómica con un conjunto de variables independientes. Puede considerarse una extensión de los modelos de regresión lineal, con la particularidad de que el dominio de salida de la función está acotado al intervalo [0,1] y que el proceso de estimación, en lugar de mínimos cuadrados, utiliza el procedimiento de estimación máximo-verosímil. Es esencial el uso de programas informáticos ya que este método de estimación es iterativo (Rioja, Llorente, & Ramirez, s.f.).

En otras palabras, un modelo de regresión logística es aquel en donde se asume que las variables explicativas multiplicadas por sus respectivos coeficientes tiene una relación lineal, no directamente con la variable de respuesta, sino con el logaritmo natural de las probabilidades (ODDS) de que el evento va a ocurrir. Esto, como explica Rioja (s.f.) se define con la fórmula final de la regresión logística:

$$Ln\left(\frac{p}{1-p}\right) = \alpha + \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Donde “p” es la probabilidad de que un cliente sea malo y “1-p” es la probabilidad de que un cliente sea bueno. El término (p/(1-p)) es llamado ODDS y se entiende como la proporción del número esperado de veces que un evento ocurra y el número esperado de veces que no ocurra.

Despejando “p” de la fórmula anterior se tiene, que la probabilidad de ser un cliente malo es:

$$p = \frac{1}{1 + e^{-(\alpha+\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}}$$

Uno de los problemas fundamentales cuando intervienen diversas variables en un fenómeno es determinar cuál es la contribución de cada una de ellas. El modelo de regresión logística a veces es el elegido ya que, a diferencia del resto de los algoritmos que veremos a continuación, es de mayor facilidad para interpretar los efectos que tienen las variables predictoras o independientes sobre la variable dependiente. Esto puede llevarse a cabo con la lectura de los coeficientes de ODDS.

2.9.2 Árboles de Decisión

Una de las técnicas más comunes de minería de datos son los Árboles de Decisión de tipo Top-Down Induction Decision Trees (TDIDT) utilizados para descubrir conocimiento en formato de regla que constituye un modelo que representa el dominio de conocimiento subyacente a los ejemplos disponibles del mismo.

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Así ayudan a tomar la decisión más “acertada”, desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a buscar subgrupos específicos y relaciones que tal vez no encontraríamos con estadísticos más tradicionales. (Silvente, Hurtado, & Baños, 2013)

Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discretización de variables continuas.

2.9.3 Clasificador kNN

El algoritmo los K-Vecinos más Cercanos (kNN) es un algoritmo simple y de alto rendimiento de clasificación supervisada. kNN pertenece al paradigma “perezoso de aprendizaje” (o “lazy learning”), donde el trabajo se retrasa todo lo posible, ya que no se construye ningún modelo. El modelo son los propios datos o conjunto de entrenamiento, y se

trabaja cuando llega un nuevo ejemplo a clasificar. Aunque kNN es una técnica simple, ha demostrado ser uno de los algoritmos más efectivos en la Minería de Datos (está considerado uno de los 10 algoritmos más importantes de la Minería de Datos (Abdelmalik , Inza, & Larrañaga)

El algoritmo kNN se basa en el cálculo de distancias entre el ejemplo a clasificar y un grupo de k ejemplos del conjunto de entrenamiento. Dichos ejemplos son los que se encuentran más próximos al ejemplo que queremos clasificar, de modo que lo clasificamos según las clases a las que pertenezcan los k ejemplos obtenidos

2.9.4 Redes Bayesianas

Una Red Bayesiana es un gráfico acíclico dirigido en el que cada nodo representa una variable y cada arco representa una dependencia probabilística que especifica el condicional de cada variable dados sus padres; la variable a la que apunta el arco depende (causa-efecto) de la variable en el origen de este (Felgaer, 2004)

Las redes bayesianas (Pearl, 1988) son utilizadas en diversas áreas de aplicación como por ejemplo medicina (Beinlich, Suermondt, Chavez, & Cooper, 1989), ciencias (Bickmore, 1994) y economía (Ezawa & Schuermann, 1995). Las mismas proveen una forma compacta de representar el conocimiento y métodos flexibles de razonamiento -basados en las teorías probabilísticas- capaces de predecir el valor de variables no observadas y explicar las observadas. Entre las características que poseen las redes bayesianas se puede destacar que permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos (Heckerman, Chickering, & Geiger, 1995) y pueden manejar bases de datos incompletas (Heckerman D. , 1995)

Las redes bayesianas están diseñadas para hallar las relaciones de dependencia e independencia entre todas las variables que conforman el dominio de estudio. Basado en ello, se utilizan métodos de razonamiento probabilístico que permiten realizar predicciones sobre el valor de cualquier variable desconocida basados en los valores de las conocidas.

Existen muchas tareas prácticas que pueden reducirse a problemas de clasificación: diagnóstico médico y reconocimiento de patrones son sólo dos ejemplos de ellas.

Las redes bayesianas pueden realizar la tarea de clasificación -caso particular de predicción- que se caracteriza por tener una sola de las variables de la base de datos (clasificador) que se desea predecir, mientras que todas las otras son los datos propios del caso que se desea clasificar. Pueden existir una gran cantidad de variables en la base de datos, algunas de las cuales estén directamente relacionadas con la variable clasificadora pero también otras variables que tienen una influencia directa sobre dicha clase (Felgaer, 2004)

2.9.5 Bosque Aleatorio (Random Forest)

El Random Forest es uno de los más poderosos algoritmos de aprendizaje de máquina disponibles en la actualidad. Es un algoritmo de aprendizaje automático supervisado que se puede utilizar para tareas de clasificación (predice una salida de valor discreto, es decir, una clase) y regresión (predice una salida de valor continuo). (Pramoditha)

Los dos conceptos principales detrás de los bosques aleatorios son:

- La sabiduría de la multitud: un gran grupo de personas es colectivamente más inteligente que los expertos individuales
- La diversificación - un conjunto de incorrelados árboles

Cuando entrena un bosque aleatorio para una tarea de clasificación, en realidad entrena un grupo de árboles de decisión. Luego, obtiene las predicciones de todos los árboles individuales y predice la clase que obtiene la mayor cantidad de votos. Aunque algunos árboles individuales producen predicciones incorrectas, muchos pueden producir predicciones precisas. Como grupo, pueden avanzar hacia predicciones precisas. A esto se le llama la sabiduría de la multitud.

En la siguiente figura 6, se puede ver gráficamente el funcionamiento de un Bosque Aleatorio (Random Forest):

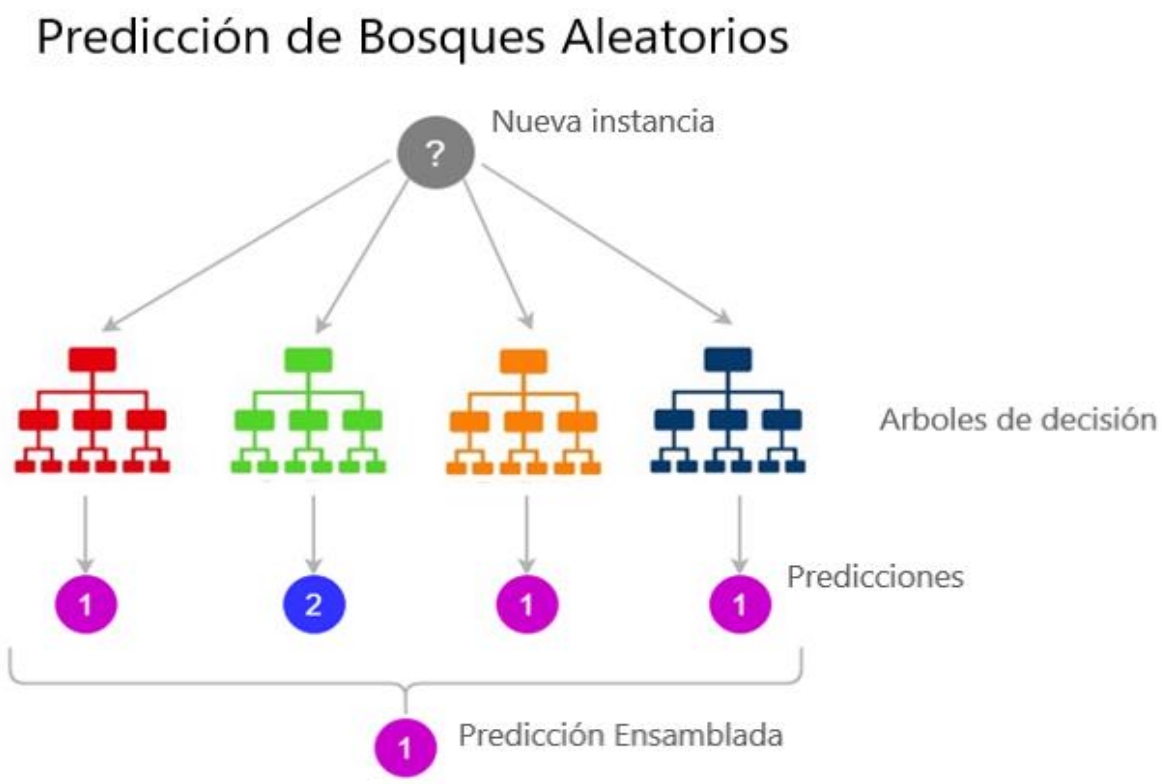

Figura 6 Diagrama del funcionamiento Random Forest (Pramoditha)

Para mantener una baja correlación (alta diversificación) entre árboles individuales, el algoritmo considera automáticamente las siguientes cosas.

- Aleatoriedad de características
- Embolsado (agregación bootstrap)

En un árbol de decisión normal, el algoritmo busca la mejor característica de todas las características cuando quiere dividir un nodo. Por el contrario, cada árbol en un bosque aleatorio busca la mejor característica de un subconjunto aleatorio de características. Esto crea una aleatoriedad adicional al hacer crecer los árboles dentro de un bosque aleatorio. Debido a la aleatoriedad de características, los árboles de decisión en un bosque aleatorio no están correlacionados.

2.9.6 Gradient Boosting

Gradient Boosting o Potenciación del Gradiente es una técnica de aprendizaje automático utilizado para el análisis de la regresión y para problemas de clasificación estadística, el cual produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión. Construye el modelo de forma escalonada como lo hacen otros métodos de boosting, y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable.

La idea de la potenciación del gradiente fue originada en la observación realizada por Leo Breiman (1997) en donde el Boosting puede ser interpretado como un algoritmo de optimización en una función de coste adecuada. Posteriormente Jerome H. Friedman (1999) desarrolló algoritmos de aumento de gradiente de regresión explícita, simultáneamente con la perspectiva más general de potenciación del gradiente funcional de Llew Mason, Jonathan Baxter, Peter Bartlett y Marcus Frean (1999). En sus últimos dos trabajos presentaron la visión abstracta de los algoritmos de potenciación como algoritmos iterativos de descenso de gradientes funcionales. Es decir, algoritmos que optimizan una función de coste sobre el espacio de función mediante la elección iterativa de una función (hipótesis débil) que apunta en la dirección del gradiente negativo. Esta visión de gradiente funcional de potenciación ha llevado al desarrollo de algoritmos de potenciación en muchas áreas del aprendizaje automático y estadísticas más allá de la regresión y la clasificación.

2.9.7 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) es un algoritmo de refuerzo (o también de potenciación) de gradientes (gradient boosting) basado en modelos de árboles de decisión, desarrollado por Microsoft. Puede ser utilizado para la categorización, clasificación y muchas otras tareas de aprendizaje automático, en las que es necesario maximizar o minimizar una función objetivo a través de la combinación de clasificadores sencillos, como por ejemplo árboles de decisión de profundidad limitada (Brownlee, 2020)

Entre sus principales ventajas podemos destacar las siguientes:

- Mayor velocidad de entrenamiento y mayor eficiencia
- Menor uso de memoria
- Mayor precisión
- Soporte de aprendizaje paralelo y soporte para GPUs
- Capacidad para manejar datos a gran escala

Los experimentos de comparación en conjuntos de datos públicos muestran que LightGBM puede superar los marcos de impulso existentes tanto en eficiencia como en precisión, con un consumo de memoria significativamente menor. Además, los experimentos de aprendizaje distribuido muestran que LightGBM puede lograr una aceleración lineal mediante el uso de varias máquinas para entrenar en entornos específicos.

2.10 Medidas de Ajuste de los Modelos

Frente a la multiplicidad de métodos de modelado, cada uno con sus propios indicadores estadísticos de calidad, se han tratado de encontrar métricas universales para el desempeño.

Para comprender esto, es necesario primero introducir el concepto de Matriz de Confusión. Esta es una tabla que se usa a menudo para describir el desempeño de un modelo de clasificación en un conjunto de datos de prueba para los que se conocen los valores verdaderos (Witten & Frank, 2005). Todas las medidas, excepto la curva ROC y KS, se pueden calcular utilizando los cuatro parámetros que presenta la Matriz, Entonces, hablemos primero de esos cuatro parámetros:

	Clase de Predicción		
Clase Real		**Clase: Si**	**Clase: No**
	Clase: Si	Verdadero Positivo	Falso Negativo
	Clase: No	Falso Positivo	Verdadero Negativo

Donde:

- **Verdadero Positivo (TP)**: estos son los valores positivos predichos correctamente, lo que significa que el valor de la clase real es sí y el valor de la clase pronosticada también es sí. P.ej. si el valor real de la clase indica que este estudiante aprobó y la clase prevista le dice lo mismo.
- **Verdaderos Negativos (TN):** estos son los valores negativos predichos correctamente, lo que significa que el valor de la clase real es no y el valor de la clase pronosticada también es no. P.ej. si la clase real dice que este no aprobó y la clase predicha le dice lo mismo.
- **Falsos positivos (FP):** cuando la clase real es no y la clase predicha es sí. P.ej. si la clase real dice que este pasajero no aprobó, pero la clase predicha le dice que este pasajero aprobó.
- **Falsos negativos (FN):** cuando la clase real es sí, pero la clase predicha es no. P.ej. si el valor real de la clase indica que este pasajero aprobó y la clase prevista le indica que el pasajero desaprobó

A continuación, describiremos los criterios mejor desarrollados y más utilizados para medir la efectividad de los modelos y así poder compararlos entre sí.

2.10.1 Exactitud o Accuracy

Es la medida de rendimiento más intuitiva y es simplemente una relación entre la observación predicha correctamente y el total de observaciones. Uno puede pensar que, si tenemos una alta precisión, nuestro modelo es el mejor. Sí, la precisión es una gran medida, pero solo cuando tiene conjuntos de datos simétricos donde los valores de falsos positivos y falsos negativos son casi iguales. Por lo tanto, debe observar otros parámetros para evaluar el rendimiento de su modelo.

$$\text{Formula Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

2.10.2 Precisión

La precisión es la relación entre las observaciones positivas predichas correctamente y el total de observaciones positivas predichas. La alta precisión se relaciona con la baja tasa de falsos positivos.

$$\text{Formula Precision} = \frac{TP}{TP+FP}$$

2.10.3 Recuperación o Recall

Es la proporción de observaciones positivas predichas correctamente con respecto a todas las observaciones en la clase real

$$\text{Formula Recall} = \frac{TP}{TP+FN}$$

2.10.4 Puntuación o Score F1

La Puntuación F1 es el promedio ponderado de precisión y recall. Por tanto, esta puntuación tiene en cuenta tanto los falsos positivos como los falsos negativos. F1 suele ser más útil que la precisión, especialmente si tiene una distribución de clases desigual. La precisión funciona mejor si los falsos positivos y los falsos negativos tienen un costo similar. Si el costo de los falsos positivos y los falsos negativos es muy diferente, es mejor mirar tanto Precision como Recall.

$$\text{Formula Score F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.10.5 Curva ROC

La curva ROC (Receiver Operating Characteristic) nos permite comparar modelos de diferentes tipos. Fogarty, Baker & Hudson (2005) indican que para entender este concepto primero es necesario introducir algunos conceptos.

- **Valor de corte (π_0)** es aquel valor que convierte a las predicciones en 1 cuando la probabilidad del evento estimada lo supera, y en 0 cuando la probabilidad del evento estimada es menor. El valor habitual que se suele tomar de corte es π_0=0.5

- **Sensibilidad y Especificidad**, la predicción de éxito cuándo es cierta se denomina sensibilidad, y la predicción de fracaso cuando es, a su vez cierta, se denomina especificidad.

Entonces, una Curva ROC es un gráfico en el que se representa la sensibilidad en función de (1-especificidad) y permite comparar modelos de diferentes tipos. Si se modifican los valores de corte π_0 y se representa la sensibilidad (en ordenadas) en función de (1-especificidad) (en abscisas), se obtiene la curva ROC. Ésta es una función cóncava que conecta los puntos (0,0) y (1,1).

- Cuando π_0 es cercano a 0, casi todas las predicciones serán 1, con lo cual la sensibilidad estará próxima a 1 y la especificidad estará cercana a 0. Así, el punto (1-especificidad, sensibilidad) tendrá coordenadas (1,1).

- Cuando π_0 es cercano a 1, casi todas las predicciones serán 0, con lo cual la sensibilidad estará próxima a 0 y la especificidad estará cerca de 1. Así, el punto (1-especificidad, sensibilidad) tendrá coordenadas (0,0).

- Cuanto mayor sea el área bajo la curva (AUC), mejores serán las predicciones. Un área igual a 0,5 representa el azar, mientras que un área igual a 1 representa al mejor

modelo.

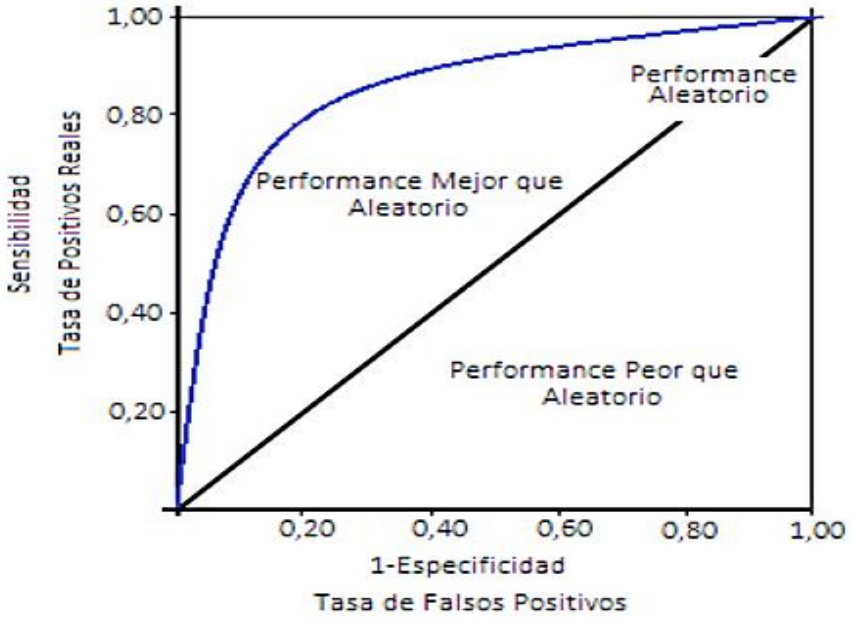

Figura 7 Curva ROC (Fogarty, Baker, & Hudson, 2005)

2.10.6 Kolmogorov-Smirnov o KS

La otra medida que vamos a tener en cuenta al momento de hacer nuestra comparación de modelos para así poder quedarnos con el de mejor performance es la prueba de Kolmogorov-Smirnov (KS), y a su vez es una de las medidas más importantes para el monitoreo de los modelos, o sea, para analizar si el modelo mantiene su performance en el tiempo.

El estadístico KS puede ser usado para medir la capacidad de clasificación de un modelo, ya consiste en medir cuan distintas son las funciones de distribución de buenos y malos clientes para cada valor de puntaje de score.

Se considera que un modelo con KS menor a 20% debe ser cuestionado, y si es mayor a 70% sea, probablemente, muy bueno para ser cierto (Kisbye, 2010)

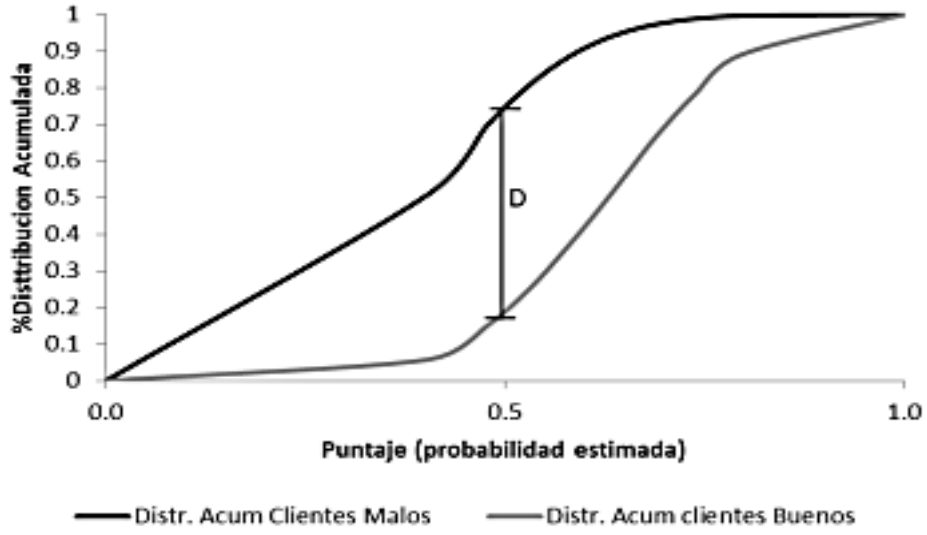

Figura 8 Test de Kolmogorov-Smirnov

2.11 Ajustes del Modelo e Hiperparámetros

Existen miles de formas y ajustes que se pueden realizar sobre el modelo para obtener mejores resultados. Entre estos se destacan estos dos métodos que permiten de manera simple, realizar distintas pruebas para obtener los mejores resultados de predicción.

2.11.1 Validación Cruzada

Los métodos de validación, son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso únicamente de los datos de entrenamiento. La idea en la que se basan todos ellos es la siguiente: el modelo se ajusta empleando un subconjunto de observaciones del conjunto de entrenamiento y se evalúa con las observaciones restantes. Por ejemplo, se puede calcular la métrica de accuracy usando datos diferentes a los de entrenamiento para medir que tan de bueno es el modelo. Este proceso se repite múltiples veces y los resultados se agregan y promedian. Gracias a las repeticiones, se compensan las posibles desviaciones que puedan surgir por el reparto aleatorio de las

observaciones. La diferencia entre métodos suele ser la forma en la que se generan los subconjuntos de entrenamiento/validación.

El método K-Fold Cross-Validation, o Validacion Cruzada, es un proceso de validación iterativo. Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración. El proceso genera k estimaciones del error cuyo promedio se emplea como estimación final.

2.11.2 Optimización de Hiperparámetros

Los hiperparámetros son parámetros externos al modelo, y que no pueden ser modificados desde los datos. Estos parámetros permiten controlar el proceso de entrenamiento de un modelo por lo que el rendimiento de un modelo depende en gran medida de los hiperparámetros (Rohan, 2018)

Un hiperparámetro de modelo es una característica de un modelo que es externa al mismo y cuyo valor no se puede estimar a partir de los datos. Por ejemplo, k en k-Vecinos más cercanos. El valor del hiperparámetro debe establecerse con cuidado antes de que comience el proceso de aprendizaje.

Por el contrario, un parámetro es una característica interna del modelo y su valor puede estimarse a partir de los datos. Ejemplo, coeficientes beta de regresión lineal / logística.

El módulo SKlearn contiene una herramienta llamada “GridSearchCV” (Pedregosa, y otros, 2011), que permite realizar esta tarea de manera simple y automática, a la vez que evalúa el modelo obtenido mediante validación cruzada

Entonces, la búsqueda en GridSearchCV se utiliza para encontrar los hiperparámetros óptimos de un modelo que dan como resultado las predicciones más "precisas". La herramienta realiza una prueba de modelo distinto por cada combinación de hiperparámetros posibles, y luego compara todos mediante validación cruzada para quedarse con el modelo con mejor poder de predicción.

3. Desarrollo del Modelo Predictivo

En este capítulo se desarrolla el proyecto que tiene como objetivo generar un Modelo Predictivo aplicado sobre las Encuestas Aprender, que logre detectar la probabilidad de un estudiante a desaprobar el examen de Matemática, para realizar acciones preventivas antes de que esto suceda, siguiendo las fases propuestas por la metodología CRISP-DM, que se explican en la sección 2.5.3.

3.1 Comprensión del negocio

Esta fase se enfoca en la comprensión de los objetivos de proyecto y las exigencias desde una perspectiva de negocio, para luego convertir este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos. Por lo tanto, en esta fase de la metodología, se llevan a cabo las siguientes actividades descriptas a continuación.

3.1.1 Determinación de los objetivos de negocio

El acompañamiento pedagógico a lo largo del año escolar es algo que viene incrementando año tras año, según indica la Dra. Pérez Guadalupe Maria en su conferencia “La importancia del acompañamiento para la transformación de la educación” (2012), luego de que se lograra comprobar que esto genera mejores resultados y experiencias en los estudiantes y los ayuda en su crecimiento personal. Hoy en día, gracias a toda la información disponible, sería posible trazar planes de estudio y seguimientos personalizados para cada estudiante.

Utilizando el resultado del examen de Matemática para los estudiantes de secundario dentro de las Encuestas Aprender, en este proyecto se trata de encontrar patrones de comportamiento en la población de estudiantes que tiene notas regulares o debajo del promedio, así como también de estudiantes con notas sobresalientes. Esto conlleva a analizar preguntas como.

- *¿Importa el acceso a internet para la performance del estudiante?*
- *¿El resultado del examen varía según los ingresos económicos del hogar?*

- *¿Importa el nivel de educación de los padres para que el estudiante incorpore correctamente los conceptos de Matemática?*

Por lo tanto, el criterio del éxito de este trabajo se basa en identificar a los atributos que poseen más influencia en las notas de los estudiantes, y a su vez generar un modelo de predicción para detectar a principio de año aquellos estudiantes que tendrán mayores dificultades en matemática, y a aquellos que se destacarán.

Por último, se establece una propuesta de implementación para poder mejorar los índices de aprobación del examen en los estudiantes mediante la utilización del modelo generado.

3.1.2 Evaluación de la situación

Las principales fuentes de información que se utilizan para el proyecto, son la encuesta Aprender 2019 **(**https://www.argentina.gob.ar/educacion/aprender2019). Para ello, se cuenta con información referente a cada encuesta mediante los sitios oficiales del ministerio de educación, así como también expertos en el área a disposición de la población.

En cuanto a herramientas de procesamiento de datos, se utiliza el lenguaje Python, a través de Jupyter Notebook para la preparación, modelado y graficado de la información, y una planilla de cálculo Excel para la representación gráfica de algunos atributos.

En cuanto a presupuesto, se dispone de los recursos disponibles por el estudiante y la universidad.

Por otra parte, como aporte de este proyecto, se intenta brindar una herramienta para ayudar a detectar el desempeño de los estudiantes en matemática de manera temprana.

3.1.3 Determinación de los objetivos de la minería de datos

El objetivo del proyecto es generar un algoritmo de predicción para determinar los estudiantes con más riesgo de tener notas bajas en matemática. Para esto, se utilizan algoritmos de clasificación descriptos en la sección 2.9, como Regresión Logística, Naive Bayes, KNN,

Arboles de Decisión y Random Forest. Asimismo, se genera un ranking de variables con mayor peso o más influyentes, y un análisis descriptivo de cada una.

Para ello, se evalúan distintos modelos en base a pruebas estadísticas como R2, Accuracy, entre otros. Además, por cada algoritmo se prueba con distintos hiper-parámetros, mediante la utilización de GridSearch. El objetivo final es obtener un modelo que tenga un Accuracy de al menos 0,70 y un área bajo la curva ROC mayor a 0,75.

3.1.4 Producir el plan del proyecto

En el proyecto se contemplan las siguientes etapas generales:

I. *Recolección de datos desde el Ministerio de Educación*: Se descargan desde el sitio web las encuestas de Aprender 2019, el diccionario de datos y el detalle de dicha instancia.

II. *Exploración y verificación de calidad de los datos:* Se verifica los datos mediante Jupyter Lab, utilizando las librerías de Numpy, Pandas y Matplotlib

III. *Preparación de los datos para el análisis:* Se imputan los datos que no se encuentren completos, se generan variables discretas a partir de las variables de texto, y por último se generan variables binarias categóricas (también conocidas como “dummies”) a partir de estas. Por otro lado, se normalizan las variables numéricas, y se genera la variable binaria target. También se divide el dataset en dos partes, una para entrenamiento y otra para evaluación.

IV. *Modelado*: En esta etapa se utilizan las técnicas de minería de datos mencionadas en la etapa de objetivo de minería de datos, mediante la librería Sklearn en Python

V. *Evaluación de los resultados*: Se evalúan los resultados de los modelos realizados, mediante diferentes técnicas de análisis y se valida la efectividad y precisión mediante la aplicación del modelo a un set de prueba.

En la figura 9, se muestra un diagrama de Gantt con la duración de cada etapa del proyecto

ACTIVIDAD	INICIO DEL PLAN	DURACIÓN DEL PLAN	INICIO REAL	DURACIÓN REAL	PERIODOS feb-21	mar-21	abr-21	may-21	jun-21	jul-21	ago-21
Recoleccion de datos desdse el Ministerio de Educación	feb-21	1	feb-21	1							
Exploración y verificación de calidad de los datos	mar-21	1	mar-21	1							
Preparación de los datos para el análisis	abr-21	2	abr-21	2							
Modelado	jun-21	2	jun-21	2							
Evaluación de los resultados	ago-21	1	ago-21	1							

Figura 9. Diagrama Gantt del proyecto

3.2 Comprensión de los datos

La fase de comprensión de los datos incluye la recolección, descripción, exploración y verificación de la calidad de los mismos

3.2.1 Recolección de datos iniciales

Los datos utilizados en este proyecto son datos referentes a la encuesta Aprender, la cual se explica según el Indec (Indec, s.f.) como:

> *"Es una encuesta de propósitos múltiples que releva información sobre los estudiantes en torno a distintos aspectos: situación familiar, características demográficas básicas (edad, sexo, situación conyugal, etc.), características migratorias, habitacionales, educacionales e ingresos. Más allá de su gran amplitud temática, los aspectos educativos y familiares adquieren una relevancia central. Entre los conceptos principales que permiten dar cuenta de la relación familiar con el desempeño del estudiante se encuentra el de condición del hogar, estudios del padre, madre, hermanos, y datos demográficos del barrio"*

Para este proyecto, además de los datos demográficos del estudiante, también se utiliza la nota del examen de Matemática, el cual incluye distintos aspectos sobre estadística, geometría, y

cálculos básicos. Todos estos datos fueron obtenidos del sitio **https://www.argentina.gob.ar/educacion/aprender2019**

Entonces, los atributos específicos disponibles para aplicar los algoritmos de minería de datos se detallan en el Anexo A.1 al final de este trabajo.

En cuanto a transformación de datos, se determina que los mismos deben ser convertidos todos a valores discretos, y luego a variables dummies. Además, se deben normalizar los mismos para poder ingresarlos de manera correcta a los algoritmos de predicción.

3.2.2 Descripción de los datos

Los datos se encuentran almacenados en un archivo .CSV, con delimitadores ";". En la figura 10 se puede ver una vista previa del dataset.

	ID1	cod_provincia	sector	ambito	claveseccion	idalumno	ap01_01	ap01_02	ap02	ap03	...
0	120003000120003	2	2	1	02SF00003	3	10	4	1	1	...
1	120003000120003	2	2	1	02SF00003	25	7	4	2	1	...
2	120003000120003	2	2	1	02SF00002	27	3	5	1	1	...
3	120003000120003	2	2	1	02SF00002	3	11	4	2	1	...
4	120003000120003	2	2	1	02SF00003	5	11	4	2	1	...

Figura 10. Vista previa del dataset Estudiantes Secundaria Aprender 2019

Todos los atributos se encuentran en formato numérico de punto flotante ("float"), a excepción de los atributos "Clave". Los valores nulos son representados, según el atributo, con valores negativos o 0.

En el Anexo A.2 al final de este trabajo se muestran los posibles valores que pueden tomar cada atributo y la referencia de los mismos.

3.2.3 Exploración de datos

Una vez que se han descrito los datos, se procede a explorarlos. Esto implica aplicar pruebas estadísticas básicas que revelan propiedades de los datos, así como también tablas de frecuencia

y gráficos de distribución de los datos. Este informe sirve principalmente para determinar la consistencia y completitud de los datos.

En la figura 11, se muestra la distribución del desempeño de los estudiantes en el examen de matemática, donde se puede apreciar que estuvieron “por debajo del nivel básico”, “básico” y “satisfactorio” (grupos 1, 2 y 3) tienen cantidades similares de estudiantes, mientras que el “Avanzado” (grupo 4) es el que tiene menor cantidad de estudiantes.

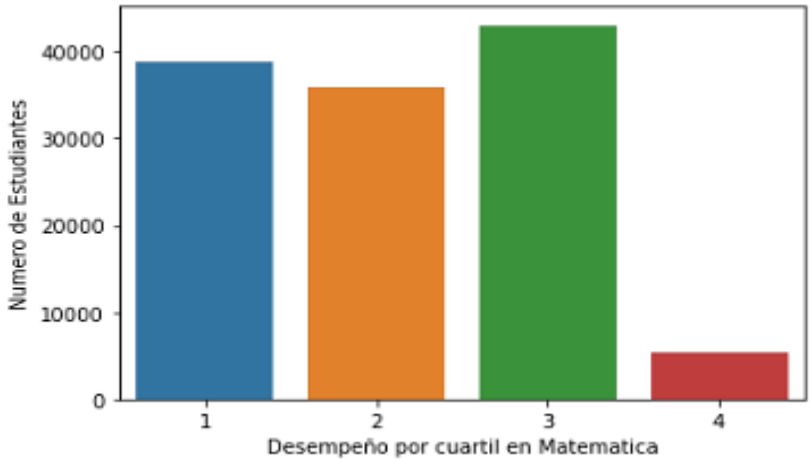

Figura 11 Distribución de desempeño de estudiantes en Matemática

Donde:

Valor	Descripción
1	Por debajo del nivel básico
2	Básico
3	Satisfactorio
4	Avanzado

Tabla 7. Valores Desmpeño de estudiantes en Matemática

En la figura 12, se puede ver la distribución de los estudiantes encuestados según el Sexo. Como se puede notar que hay casi un 40% más de Mujeres que de Varones encuestados.

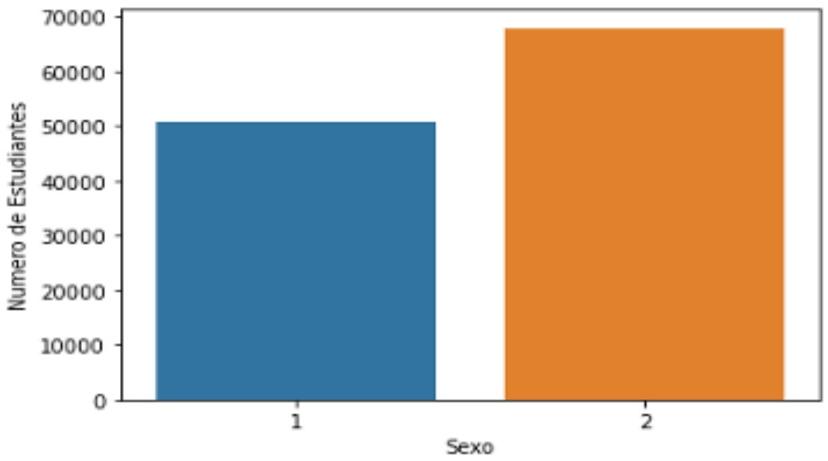

Figura 12. Distribución de Estudiantes según Sexo

Donde:

Valor	Descripción
1	Varón
2	Mujer

Tabla 8. Valores Distribución de Estudiantes segun Sexo

Por otro lado, en la figura 13 se muestran las distribuciones según el Sector del colegio. La misma nos demuestra que las cantidades de estudiantes encuestados entre colegio estatal y privado son similares. Pero, como se puede ver en la figura 14, la distribución del colegio según el Ámbito del Colegio nos muestra que la gran mayoría de estudiantes encuestados (más del 90%) pertenecen a sectores Urbanos.

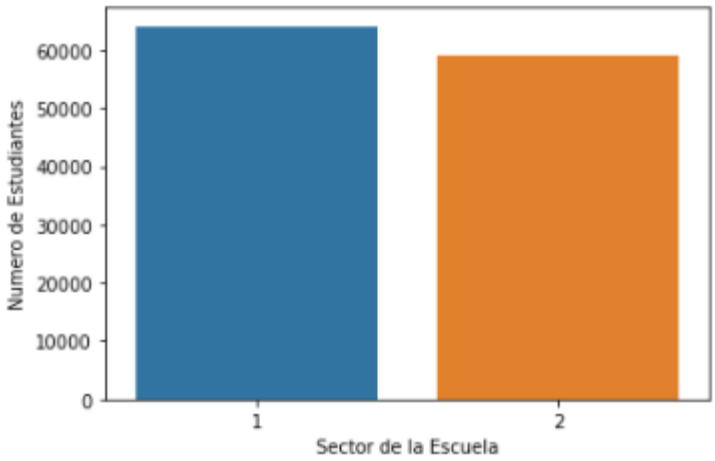

Figura 13. Distribución según el sector del Colegio

Donde:

Valor	Descripción
1	Estatal
2	Privado

Tabla 9. Valores Distribución según el sector del Colegio

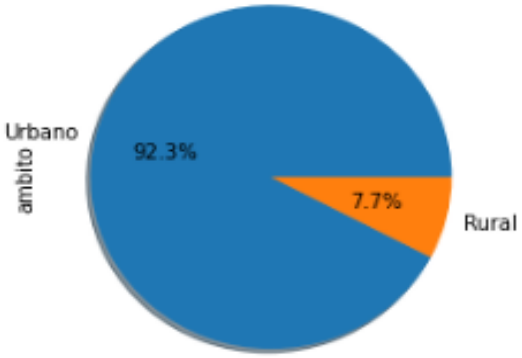

Figura 14. Distribución según Ámbito Escolar

Por último, también se puede observar la distribución según el índice Socioeconómico (figura 15), el cual ya viene integrado en los resultados de las encuestas Aprender, y está basado en el cálculo de la encuesta permanente de hogares realizada por el INDEC

(*www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos*). La misma contempla las siguientes variables:

- Nivel educativo de los padres.
- Hacinamiento en el hogar (relación entre la cantidad de habitaciones de la vivienda en la que habita el estudiante y el número de miembros del hogar).
- Recepción de la Asignación Universal por Hijo (AUH) en el hogar.
- Tenencia de equipamiento informático en el hogar (Internet, consolas de videojuegos, televisión y celular).

Como se puede apreciar, el nivel más representativo es el socioeconómico Medio, seguido por el Alto.

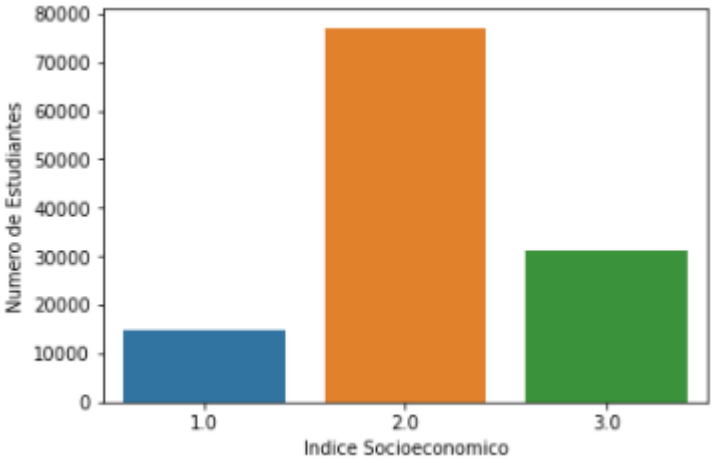

Figura 15 Distribución según índice Socioeconómico

Donde:

Valor	Descripción
1	Índice socioeconómico Bajo
2	Índice socioeconómico Medio
3	Índice socioeconómico Alto

Tabla 10. Valores Distribución según índice Socioeconómico

3.2.4 Verificación de la calidad de los datos.

El dataset tiene en total 343.750 registros, y 264 atributos. Luego de hacer la exploración inicial de los datos se puede afirmar que son completos, y aptos para el análisis. La cantidad de registros cubre los necesarios para aplicar los modelos. Se ha detectado la presencia de datos nulos y outliers (datos que vienen sin información y/o información fuera de sus niveles normales) por lo que se tendrá que trabajar en la preparación de los datos.

3.3 Preparación de los datos

En esta fase, se trata de preparar los datos para aplicar las técnicas de minería de datos detalladas en la sección 3.2.7 sobre ellos de manera eficaz y eficiente. A su vez se debe adaptar los datos a la necesidad, seleccionando un subset de datos con los que se trabajará posteriormente.

3.3.1 Seleccionar los Datos

En principio, se van a utilizar todos los registros (o filas) de la encuesta Aprender 2019. Pero, en términos de atributos (o columnas), a partir de la lista indicada en el anexo A.1 se seleccionan sólo los que tengan mayor correlación con el atributo target (aproximadamente 65 atributos). Los mismos son listados a continuación:

cod_provincia	ap25_03	ap50_05
sector	ap28_01	ap51_01
ambito	ap29_01	ap51_02
ap02	ap34	ap51_03
ap03	ap35_03	ap51_04
ap04	ap39_01	ap51_05
ap05	ap39_02	ap51_06
ap06	ap39_03	ap51_07
ap07	ap39_04	ap51_08
ap09	ap40_01	ap51_09
ap11_08	ap40_02	ap51_10
ap14_02	ap40_03	ldesemp
ap16	ap40_04	mdesemp

ap17	ap40_05	isocioa
ap23_01	ap40_06	ap26_rec
ap23_02	ap42_01	trabaja_fuera_hogar
ap23_03	ap43_a	trabaja_fuera_hogar_remunerado
ap23_04	ap49_01	migración
ap23_05	ap49_02	edadA_junio2019
ap23_06	ap50_01	ap32_e
ap24	ap50_02	ap33_01_a
ap25_01	ap50_03	ap33_01_b
ap25_02	ap50_04	

Tabla 11. Listado de atributos con mayor correlación al target

3.3.2 Limpiar los Datos

La base de datos con la que se cuenta para el proyecto contiene toda la información necesaria para poder cumplir los objetivos de la minería de datos.

Como se puede observar en A.2, la tabla cuenta con valores negativos, que, según el atributo, representa falta de contestación, o una opción. Una posible solución para aplicar la minería de datos en los datos nulos o sin contestar, podría ser descartarlos de la muestra, o imputarlos con la media o moda del atributo.

En este caso, para poder identificarlos, primero se transforman los atributos en los cuales el valor negativo aporta una respuesta válida usando los siguientes pasos:

1) Primero, se remplaza en los *atributos ap32_e, ap33_01_a, ap33_01_b y ap43_a* el valor “-9” por “2”.
2) Luego, el resto de valores negativos en el dataset se convierten a valores Nulos en Python (“NaN”). Esto permite cambiar el tipo de todas las columnas a formato de número entero (“integer”).
3) De esta manera se calcula el porcentaje de valores faltantes por atributo, el cual se puede ver en la siguiente tabla:

Atributo	**% Valores Faltantes**
cod_provincia	0%
sector	0%
ambito	0%
ap02	3%

Atributo	% Valores Faltantes
ap03	2%
ap04	2%
ap05	4%
ap06	3%
ap07	4%
ap09	3%
ap11_08	6%
ap14_02	15%
ap16	3%
ap17	4%
ap23_01	4%
ap23_02	7%
ap23_03	7%
ap23_04	6%
ap23_05	8%
ap23_06	7%
ap24	3%
ap25_01	9%
ap25_02	9%
ap25_03	10%
ap28_01	4%
ap29_01	4%
ap32_e	1%
ap33_01_a	1%
ap33_01_b	1%
ap34	6%
ap35_03	6%
ap39_01	6%
ap39_02	6%
ap39_03	6%
ap39_04	6%
ap40_01	6%
ap40_02	7%
ap40_03	7%
ap40_04	7%
ap40_05	7%
ap40_06	7%
ap42_01	9%
ap43_a	1%
ap49_01	27%
ap49_02	27%
ap50_01	25%

Atributo	% Valores Faltantes
ap50_02	26%
ap50_03	27%
ap50_04	27%
ap50_05	26%
ap51_01	26%
ap51_02	26%
ap51_03	27%
ap51_04	27%
ap51_05	27%
ap51_06	27%
ap51_07	28%
ap51_08	27%
ap51_09	27%
ap51_10	27%
ldesemp	4%
mdesemp	6%
isocioa	6%
ap26_rec	3%
trabaja_fuera_hogar	4%
trabaja_fuera_hogar_remunerado	4%
migración	2%
edadA_junio2019	5%

Tabla 12. Porcentaje de valores faltantes por atributo

Dado que se considera que se dispone de suficientes datos para poder aplicar los algoritmos, se decide borrar todos los registros donde haya uno o más atributos nulos.

Por lo tanto, como resultado, el dataset final queda de 123.018 registros, y 68 columnas.

3.3.3 Construir los Datos

En esta sección se generan las variables necesarias para el modelo, a partir de los datos existentes.

Específicamente se trabaja con la variable target “mdesemp”. La misma tiene de origen 4 posibles valores (1 - Por debajo del nivel básico, 2 - Básico, 3 – Satisfactorio, 4 – Avanzado). Como nuestro objetivo es detectar estudiantes con nota por debajo del promedio, se deben agrupar el resto de los valores para generar una variable binaria que nos permita tener mejores

niveles de predicción. Para esto, se va a transformar la variable agrupando los valores de la siguiente manera:

Valor Final	Valor Origen
1	1- Por debajo del nivel básico
0	2- Básico
	3- Satisfactorio
	4- Avanzado

Tabla 13. Valores de origen de la variable Target

Una vez generada la variable final de predicción, se evalúa la distribución de la variable target:

Valor de atributo Target	Cantidad
0	84.296
1	38.722

Tabla 14. Valores finales de la variable Target

3.3.4 Integrar los Datos

No ha sido necesario generar nuevas estructuras ni la fusión de distintas tablas, dado que, al ser una única tabla de origen, todos los datos ya se encuentran integrados.

3.3.5 Formateo de los Datos

Para poder adaptar el dataset a los modelos predictivos, es necesario generar variables "dummies" con todos los atributos discretos del dataset. Esto se realiza con la función "get_dummies" de Pandas. Además, para ahorrar recursos, se elimina la primera columna de cada "dummie" generada, ya que la misma es redundante al ser posible identificar el valor de la variable con el resto de las columnas.

Por otro lado, se normalizaron los atributos continuos, para mejorar las predicciones de los algoritmos. Para esto se utilizó la función "StandardScaler" de Sklearn.

Como resultado de estas operaciones, se puede ver en la tabla 16 que la distribución no es balanceada ya que los registros con el atributo clase o “target” igual a 1, son un 46% menos que tienen igual a 0.

Dado que, para mejorar la predicción de los modelos, es recomendable balancear las clases en la etapa de entrenamiento del modelo por lo que se decide aplicar los siguientes pasos para generar un balance, sobretodo de los datos utilizados para entrenar los algoritmos. Entonces, primero, para separar el dataset en datos de Entrenamiento y de Validación se aplica el método “Model_Selection” de SckitLearn. El mismo logra llevar a cabo una división de manera equitativa y estratificada del dataset en dos sub-datasets, usando un 70% del dataset para el entrenamiento, y el 30% restante para la validación o evaluación del modelo. Luego, se aplica la función de “under_sampling” del módulo “imblearn” con el objetivo de balancear la cantidad de ejemplos con respecto valores clase del atributo “target”, pero sólo en el dataset de Entrenamiento. Dicha función implica la selección aleatoria de ejemplos de la clase mayoritaria para eliminarlos del conjunto de datos, y así generar que ambas clases queden con la misma cantidad de registros

Luego de estas transformaciones, los dataset quedan definidos de la siguiente manera:

Dataset de Entrenamiento	
Valor de atributo Target	**Cantidad**
0	29.086
1	29.086

Tabla 15. Distribución de dataset entrenamiento

Dataset de Validacion	
Valor de atributo Target	**Cantidad**
0	21.122
1	9.633

Tabla 16. Distribucion de dataset validación

Finalmente, se separan tanto el dataset de entrenamiento como de validación, en dos dataframes distintos para separar los valores del atributo target de los atributos de entrada o features. Como resultado queda de la siguiente manera:

X_train = Dataset de entrenamiento con atributo Target binario, y balanceo de clases.

Y_train = Dataset de entrenamiento con features formateadas y balanceo de clases.

X_test = Dataset de validación con atributo Target binario.

Y_test = Dataset de validación con features formateadas.

3.4 Modelado

El modelado consiste en la implementación de los algoritmos de minería de datos necesarios para responder las preguntas surgidas en las fases de conocimiento del negocio y de exploración de los datos.

3.4.1 Escoger la Técnica de Modelado

Dado que el problema a resolver es de clasificación binaria se pueden usar múltiples algoritmos para dicha finalidad, donde cada algoritmo a su vez cuenta con diversos hiperparámetros que se pueden aplicar en cada uno.

Entonces, para poder realizar esta tarea de forma automatizada, se decide utilizar la función “Pipeline” del módulo Scikitlearn. Esta consiste en ensamblar distintos algoritmos que generan los modelos correspondientes usando los datos de Entrenamiento definidos en la sección anterior. Luego, cada modelo será validado a través de un operador de “cross validation” (explicado en la sección 2.11.1) usando los datos de Validación.

Por otro lado, se usa “GridSearch” (sección 2.11.2), que es otra herramienta del módulo ScikitLearn, que permite probar distintas combinaciones de hiper-parámetros para cada modelo, logrando así obtener el modelo más preciso para la predicción.

Finalmente, se miden con métodos estadísticos cada modelo realizado, para que los "pipelines" indiquen cuál es el mejor, y con qué hiperparámetros tuvo su mejor performance.

En este caso, dentro de la función "Pipeline" se desarrollan pruebas con los siguientes algoritmos que ya han sido explicados en la sección 2.9:

- Regresiones Logísticas
- Naive Bayes
- Clasificador KNN
- Arboles de Decision
- Random Forest
- LightGBM
- XGBoosting

En total, a partir de los distintos algoritmos seleccionados y su diferente combinación de hiperparámetros posible, se estarían probando alrededor de 300 modelos distintos para ver cual tiene mejor nivel de predicción.

3.4.2 Generar el plan de Prueba

El procedimiento que se emplea para probar la calidad y validez del modelo es, por un lado, la separación del dataset inicial en datos de entrenamiento y validación (usando 70% entrenamiento y 30% validación). Por otro lado, como se ha indicado en la sección anterior, se aplica la técnica de "CrossValidation" en la etapa de entrenamiento de todos los modelos, con separación de 5 sub-datasets. De esta manera, se va a probar el modelo primero con los sub-datasets dentro del entrenamiento, y posteriormente con el dataset de validación.

Por otro lado, se utilizan distintos métodos estadísticos para evaluar la calidad y performance de los modelos. En este caso, se aplican los siguientes:

- Exactitud o Accuracy

- Precisión o Precision
- Recuperación o Recall
- Puntuación o Score F1
- Area bajo la curva ROC (AUC)
- Kolmogorov-Smirnov o KS

3.4.3 Construir el Modelo

Antes de comenzar la ejecución de los algoritmos seleccionados sobre los datos de entrenamiento, es necesario determinar la combinación de hiperparámetros a utilizar. Por lo tanto, en este apartado se describen los ajustes de parámetros. Para mayor calidad, se indican los parámetros para cada algoritmo probado:

- **Regresiones Logísticas:**

  Para los modelos de regresión, se aplicaron pruebas con los siguientes posibles hiperparámetros:

  - Penalidad: 11 o 12
  - Regularizacion C: 0,001; 0,01; 0,10; 1.00; 10,00, 100,00; y 1000

- **Naive Bayes:**

  Para Bayes, se prueban con distintos valores sobre la porción de mayor variación de todas las características (smoothing), utilizando valores entre 0,00000001, 0,000000001 y 0,00000001

- **Clasificador KNN:**

  Para KNN, se prueban con distinto rango de distancia entre los vecinos, es decir con todos los valores enteros entre 1 y 50.

- **Árboles de Decisión:**

Para árboles de decisión, se prueba con toda la combinación de los siguientes hiperparámetros:

- Criterio de clasificación: Gini o Entropía
- Criterio de división: Mejor o Aleatorio
- Máxima profundidad: None, 5 o 10
- Mínimo de muestras para separación: 2 o 5
- Mínimo de muestras para hoja: 1, 2 o 3

- **Random Forest:**

  Para este algoritmo, se utilizaron la siguiente combinación de hiperpárametros:

  - Criterio de clasificación: Gini o Entropía
  - Cantidad de estimadores: 3, 5, 10 o 50
  - Máxima profundidad: None, 5 o 10
  - Mínimo de muestras para separación: 2 o 5
  - Peso de clases: None, o Balanceadas

- **LightGBM & XGBoosting:**

  Para el ensamble de árboles de decisión, se utilizan la siguiente combinación de hiperparámetros:

  - Detención temprana de rondas: 20
  - Métrica de evaluación: AUC
  - Evaluación de nombres: Valido
  - Verbose: 100
  - Máxima profundidad: Aleatorio entre 10 y 50
  - Número de abandonos: Aleatorio entre 6 y 50

- Ratio de aprendizaje: [0,5; 0,2; 0,1; 0,01; 0,001]
- Mínimo de hijos por muestra: Aleatorio entre 100 y 500
- Peso mínimo de hijo: [1e-5, 1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3, 1e4]
- Registro Alpha: [0, 1e-1, 1, 2, 5, 7, 10, 50, 100]
- Registro Lambda: [0, 1e-1, 1, 5, 10, 20, 50, 100]

3.4.4 Evaluar el Modelo

En esta sección, se evalúan los resultados de la ejecución de todos los modelos previamente mencionados. A continuación, se indican los resultados según el algoritmo probado:

- **Regresiones Logísticas:**

  El mejor modelo surge con los hiper-paráameros de C: 0.001 y Penalidad: 12. En los datos de validación se genera un accuracy de 0,711, y un área bajo la curva ROC de 0,80

Clase	precisión	recall	f1-score
0	0,755	0,624	0,784
1	0,550	0,633	0,629
Promedio	0,702	0,628	0,706

Accuracy:	**0,711**

Tabla 17. Resultados Regresiones Logísticas

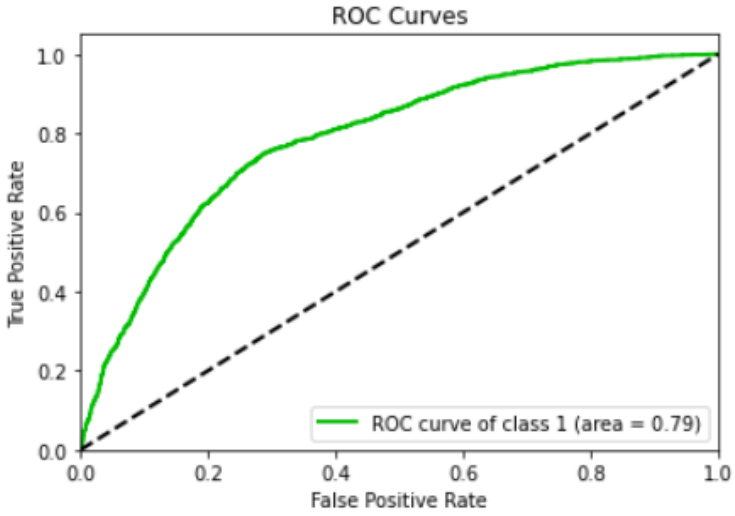

Figura 16 Curva ROC – Modelo Regresión Logistica

- **Naive Bayes:**

El mejor modelo surge con Smoothing = 0,00000001. En los datos de validación se genera un accuracy de 0,711, y un área bajo la curva ROC de 0,74

Clase	precisión	recall	f1-score
0	0,785	0,798	0,792
1	0,541	0,522	0,531
Promedio	0,663	0,660	0,661

Accuracy:	**0,711**

Tabla 18. Resultados Naive Bayes

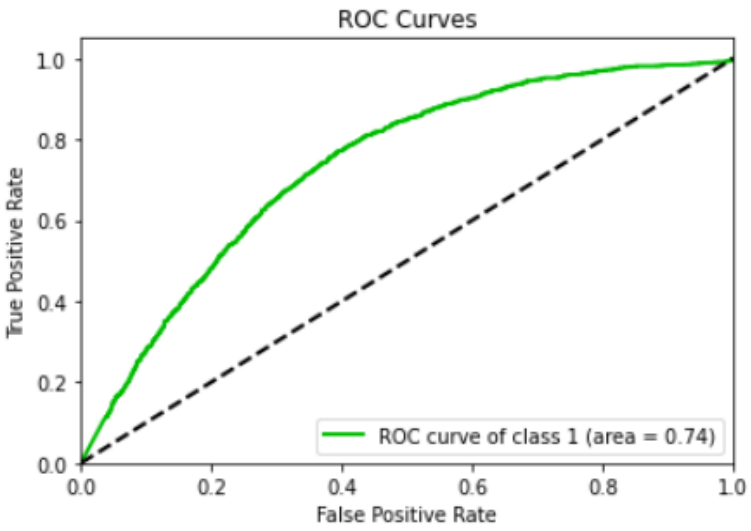

Figura 17 Curva ROC – Modelo Naive Bayes

- **Clasificador KNN:**

El mejor modelo surge con n_neighbors = 35. En los datos de validación se genera un accuracy de 0,706, y un área bajo la curva ROC de 0,75

Clase	precisión	recall	f1-score
0	0,789	0,810	0,799
1	0,558	0,524	0,540
Promedio	0,673	0,667	0,670

Accuracy:	**0,706**

Tabla 19. Resultados Clasificador KNN

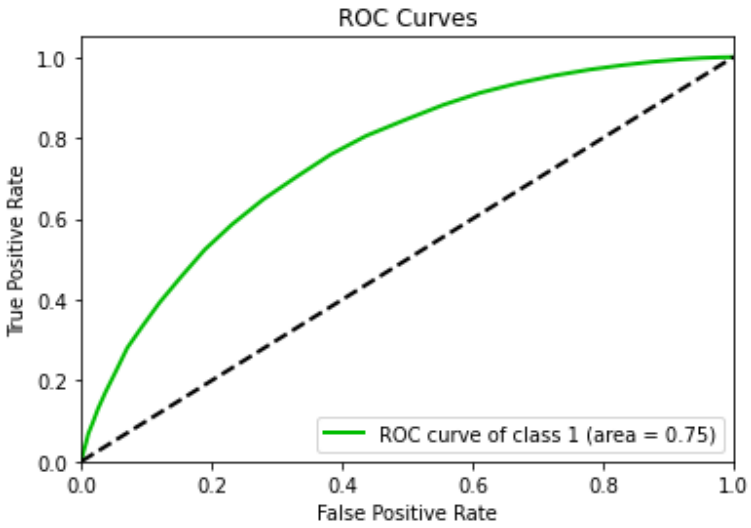

Figura 18 Curva ROC – Modelo clasificador KNN

- **Árbol de decisión:**

  El árbol con mejor predicción surge de combinar los siguientes hiperparámetros:

  - Criterion: 'entropy'
  - Max_depth: 5
  - Min_samples_leaf: 1
  - Min_samples_split: 2
  - Splitter: 'best'

  En los datos de validación se genera un accuracy de 0,671, y un área bajo la curva ROC de 0,73

Clase	precisión	recall	f1-score
0	0,822	0,665	0,735
1	0,482	0,684	0,566
Promedio	0,652	0,674	0,650

Accuracy:	**0,671**

Tabla 20. Resultados Árbol de Decisión

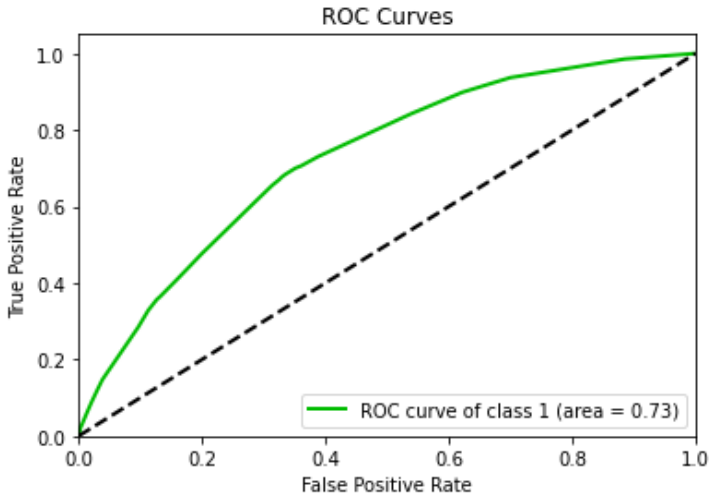

Figura 19 Curva ROC – Modelo de árbol de decisión

- **Random Forest:**

  El modelo con mejor predicción surge de combinar los siguientes hiperparámetros:

  - Criterion: 'entropy'
  - Max_depth: None
  - Class_Weight: ‘balanced’
  - Min_samples_split: 5
  - N_estimators: 50

  En los datos de validación se genera un accuracy de 0,694, y un área bajo la curva ROC de 0,76

Clase	precisión	recall	f1-score
0	0,844	0,680	0,753
1	0,508	0,724	0,597
Promedio	0,676	0,702	0,675

Accuracy:	**0,694**

Tabla 21. Resultados Random Forest

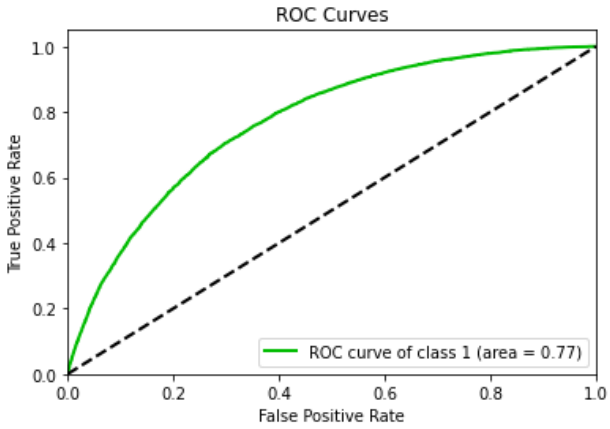

Figura 20 Curva ROC – Modelo Random Forest

- **Modelo LightGBM:**

  La mejor predicción surge con los hiperparámetros:

  - colsample_bytree: 0.40
  - learning_rate :0.1
  - max_depth: 14
  - min_child_samples: 462
  - min_child_weight: 10
  - num_leaves: 190
  - reg_alpha: 5
  - reg_lambda: 1

  En los datos de validación se genera un accuracy de 0,714, y un área bajo la curva ROC de 0,79

Clase	precisión	recall	f1-score
0	0,857	0,699	0,770
1	0,530	0,744	0,619
Promedio	0,693	0,722	0,694

Accuracy:	**0,714**

Tabla 22. Resultados LightGBM

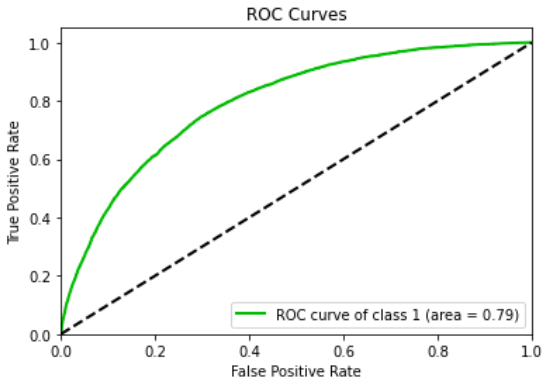

Figura 21 Curva ROC – Modelo LightGBM

- **Modelo XGBoosting:**

  La mejor predicción surge con los hiperparámetros:

  - colsample_bytree: 0.48
  - learning_rate :0.01
  - max_depth: 36
  - min_child_samples: 143
  - min_child_weight: 0.001
  - num_leaves: 12
  - reg_alpha: 10
  - reg_lambda: 100

  En los datos de validación se generan un accuracy de 0,711 y un área bajo la curva ROC de 0,79

Clase	precisión	recall	f1-score
0	0,861	0,690	0,766
1	0,527	0,756	0,621
Promedio	0,694	0,723	0,694

Accuracy:	**0,711**

Tabla 23. Resultados XGBoosting

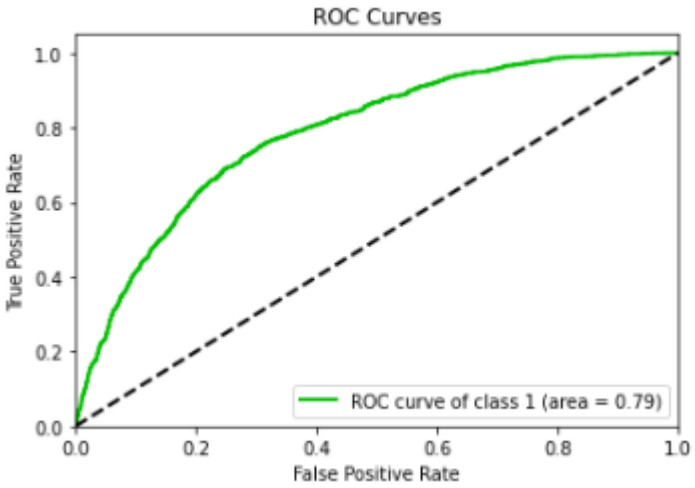

Figura 22 Curva ROC – Modelo XGBoosting

Como conclusión de esta sección, se presenta una tabla resumen con los resultados por cada algoritmo aplicado para su comparación:

Algoritmo	Accuracy	AUC	Resultado Final
Regresiones Logisticas	0,711	0,80	Resultado Aceptable
Naive Bayes	0,711	0,74	Resultado Aceptable
Clasificador KNN	0,706	0,75	Resultado Aceptable
Arbol de Decisión	0,671	0,73	Resultado No Aceptable
Random Forest	0,694	0,76	Resultado No Aceptable
Modelo LightGBM	0,714	0,79	Resultado Aceptable
Modelo XGBoosting	0,694	0,79	Resultado No Aceptable

Tabla 24. Comparativa de Accuracy según modelo

3.5 Evaluación

A continuación, se expone una evaluación de los resultados generales, así como una revisión de los aspectos por mejorar.

3.5.1 Evaluar los resultados

Luego de analizar los resultados arrojados por los distintos modelos e hiperparámetros probados, se concluye que el modelo con mejores niveles de predicción es el resultante del algoritmo **LightGBM** con los hiperparámetros configurados de la siguiente manera:

- colsample_bytree: 0.40
- learning_rate: 0,1
- max_depth: 14
- min_child_samples: 462
- min_child_weight: 10
- num_leaves: 190
- reg_alpha: 5
- reg_lambda: 1

De esta manera, el dataset de entrenamiento genera un accuracy de 0,780, mientras que en con datos de validación arroja un accuracy de 0,714. Dichos resultados son levemente mayores al resto de los modelos probados y alcanzando el objetivo propuesto (es decir, resultados mayores a 0,7 de accuracy).

Por otro lado, el área bajo la curva ROC (AUC) se encuentra entre las mejores sólo siendo superado por el valor de Regresiones Logisticas e igualado por el Modelo XGBoosting. No obstante en ambos casos el valor del accuracy es levemente menor.

Por último, para el modelo ganador se mide el test KS (Kolmogorov-Smirnov) lo cual arroja un resultado de 0,44. De esta manera se confirma que el modelo es aceptable según lo dicho en la sección 2.10.6.

Por último, para facilitar la comprensión de los resultados, se analiza el peso que tienen las variables dentro del modelo mediante la librería “eli5” en Python. La lista de las 15 variables por importancia queda ordenada de la siguiente manera:

Peso	Variable	Descripción
0,1793	sector_2	¿Es sector privado o público?
0,0419	ap40_01_1.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Disfruto estudiando Matemática - Nada de Acuerdo
0,0315	ap40_04_4.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Si me lo propongo, puedo ser bueno en Matemática
0,0298	isocioa_3.0	¿Índice socioeconómico Alto del estudiante?
0,0279	edadA_junio2019_17.0	¿Edad de 17 años?
0,0257	ap40_02_1.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Me interesan las clases de Matemática en mi escuela
0,0213	ap24_1.0	¿Asististe a nivel inicial? - Sí, antes de los cuatro años
0,0191	ap39_04_3.0	¿cómo te resulta Resolver problemas y ejercicios? - Fácil
0,0171	ap39_04_4.0	¿cómo te resulta Resolver problemas y ejercicios? - Muy Fácil
0,0171	ap40_01_4.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Si me lo propongo, puedo ser bueno en Matemática
0,0153	ap16_6.0	Máximo nivel educativo de la madre - Educación Superior universitaria
0,0137	ap40_01_2.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Disfruto estudiando Matemática - Poco de Acuerdo
0,0129	ap40_01_3.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Disfruto estudiando Matemática - De Acuerdo
0,0119	ap23_02_1.0	En la última semana, ¿cuántas veces aprendiste idiomas en tu tiempo libre, fuera de la escuela? Una o dos veces
0,0118	ap34_2.0	¿Vas a Seguir estudiando en educación universitaria cuando termines el secundario? Seguir estudiando en educación universitaria

Tabla 25. Peso de variables más importantes

Si se observa la distribución de algunas de estas variables con el atributo target, se nota que existe una fuerte correlación entre ellas, es decir, la medida estadística que expresa la medida en que dos variables están relacionadas linealmente.

En la Figura 23, podemos ver como la desaprobación tiene gran correlación con el sector del colegio del estudiante. El índice de desaprobados es muy bajo en colegios de sector privado.

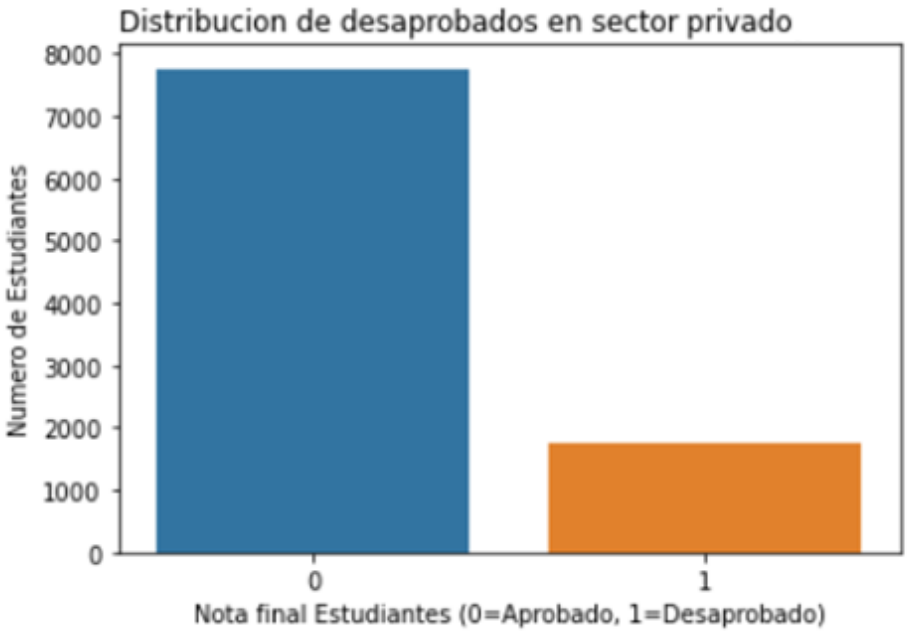

Figura 23. Distribución de desaprobados según atributo Sector

Por otro lado, en la figura 24 se puede detectar que los estudiantes con índice socioeconómico Alto tienen una tasa de desaprobación muy baja.

Figura 24. Distribución de desaprobados según atributo isocioa = 3

Siguiendo el análisis, se ve en la Figura 25 como, con menor correlación, hay una tendencia marcada en aquellos estudiantes con 17 años de edad, teniendo una relación un poco más pareja entre desaprobados y aprobados.

Figura 25. Distribución de desaprobados según atributo edadA_junio2019 = 17

Por otro lado, en las figuras 26 y 27, vemos como aquellos estudiantes con interés en matemática, o en seguir una educación universitaria relacionada, no siempre tienen tendencia a la aprobación del examen.

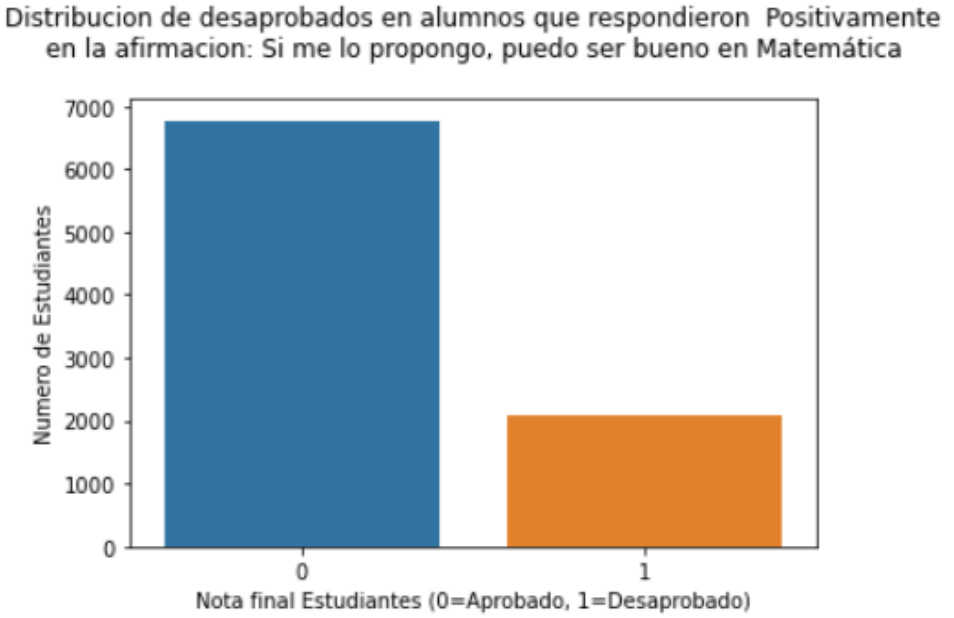

Figura 26. Distribución según atributo ap40_04 = 4

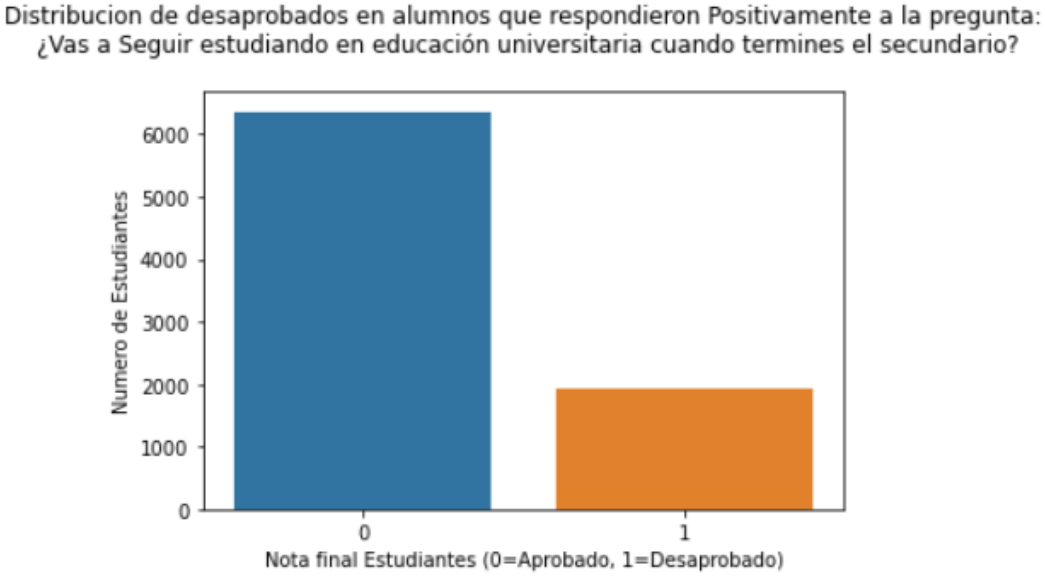

Figura 27. Distribución según atributo ap34 = 2

3.5.2 Revisión del proceso

Haciendo una revisión global del proceso de modelado, se determina que se han utilizado todos los algoritmos relevantes, y se han obtenido resultados satisfactorios con el dataset de pruebas.

3.5.3 Próximos pasos

Como próximos pasos del proyecto, está estipulado probar la efectividad del modelo ganador en las futuras encuestas de Aprender, para ver la estabilidad del modelo a lo largo de los años posteriores

3.6 Implementación

A partir de los resultados del modelo, surgen recomendaciones de aplicación de los resultados, ejecución del modelo en otros conjuntos de datos y preguntas que podrían generar nuevas investigaciones. A continuación, se presentan dichas recomendaciones.

3.6.1 Plan de implementación

La implementación que se recomienda, en base a los resultados y lo propuesto en la sección 3.1.1, es separar a los estudiantes según su riesgo de desaprobación en el examen de matemáticas. De esta manera, se definen 10 grupos de riesgos de acuerdo al resultado de probabilidad que arroja el modelo **LightGBM** generado.

Utilizando los datos de validación (datasets X_test y Y_test) se divide a la población en 10 partes iguales (o deciles) de igual cantidad de estudiantes, quedando organizados de la siguiente manera:

Ranking Riesgo	Estudiantes totales	Desaprobados	Tasa de desaprobación	Captura acumulada de desaprobados
1	3076	2230	72%	23%

2	3075	1840	60%	42%
3	3076	1483	48%	58%
4	3075	1201	39%	70%
5	3075	972	32%	80%
6	3076	721	23%	88%
7	3075	551	18%	93%
8	3076	332	11%	97%
9	3075	214	7%	99%
10	3076	89	3%	100%
Total	**30755**	**9633**	**31%**	**-**

Tabla 26. Ranking de estudiantes segun riesgo de desaprobación

Teniendo en cuenta la población separada en grupos según su riesgo de desaprobación, los docentes deberían aplicar acciones de corrección temprana especiales sobre los estudiantes que se encuentran en los deciles más críticos 1, 2, 3 y 4. De esta manera, se estaría dando apoyo anticipado al 70% de los estudiantes que desaprueban el examen (12.302 estudiantes, sobre la muestra total de 30.755).

En el resto de los grupos del rankings, al no ser de riesgo, no es necesario realizar ninguna acción diferente, aunque se recomienda ser monitoreados de forma normal.

Por último, se responden las preguntas planteadas en la sección 3.1.1:

- *¿Importa el acceso a internet para la performance del estudiante?*

  Si bien el atributo tiene una correlación marcada con el modelo, no es de los atributos más críticos al momento de determinar el éxito del estudiante en el examen

- *¿El resultado del examen varía según los ingresos económicos del hogar?*

  Efectivamente, es la 4ta variable con más influencia (o correlación) con respecto a la nota del examen, por lo que es determinante al momento de evaluar al estudiante.

- *¿Importa el nivel de educación de los padres para que el estudiante incorpore correctamente los conceptos de Matemática?*

Los resultados muestran que el nivel de la madre (Variable N°11 en la tabla de correlaciones) es más importante que el nivel del padre al momento de determinar el éxito del estudiante en el examen.

3.6.2 Plan de monitoreo y mantención

El plan de monitoreo para el modelo, se establece según lo indican los pasos futuros del proyecto, aplicando el modelo de predicción en futuras encuestas de Aprender, para validar que siga siendo efectivo y con un ordenamiento estable.

3.6.3 Informe final

Como informe final del modelado, se puede establecer que se alcanzaron las metas proyectadas, y pudimos resolver los conflictos de limpieza y estructura de datos de manera positiva. Como resultado final, se obtiene un modelo con buenos niveles de predicción, y estabilidad, para poder utilizar con las futuras encuestas de Aprender, o para aplicar en otro ámbito similar.

3.6.4 Revisión de proyecto

En la revisión del proyecto, hacemos una validación sobre los objetivos previamente mencionados:

- ✓ Realizar un análisis de datos sobre la encuesta Aprender
- ✓ Adaptar los datos para poder ser procesados en un modelo predictivo
- ✓ Aplicar un algoritmo de predicción con un Accuracy mayor a 0,7
- ✓ Definir variables de mayor peso con respecto a la nota del estudiante
- ✓ Establecer un plan de implementación para ayudar a los estudiantes más críticos

4. Líneas Finales

En el siguiente capítulo, se presentan las conclusiones del trabajo, en conjunto con información adicional referente al mismo.

4.1 Conclusión

La aplicación de técnicas de minería de datos sobre distintas fuentes de datos puede revelar información que no se conocía hasta el momento. La misma le otorga un foco de mayor detalle al análisis, y puede mejorar, improvisar, y adelantarse a problemas sobre cualquier contexto.

En nuestro estudio se observa que, utilizando las técnicas de minería, se ha logrado revelar información innovadora en el campo de la educación nacional. Es decir que, en base al modelo generado, ha sido posible tener una ganancia real y demostrada con los planes de implementación propuestos.

El aprendizaje automático (machine learning) no es una herramienta que de soluciones por sí sola, sino que debe estar acompañada en todo el proceso, por un grupo de personas involucradas y con distintos roles. Pero, si al conocimiento del experto, se lo complementa con modelos de datos, puede dar como resultado una experiencia superior y más eficiente.

Por otro lado, se han probado soluciones con distintos algoritmos, y ver que todos arrojan resultados positivos, aunque algunos demuestran mayor poder de predicción que otros.

El lenguaje Python, junto con sus librerías y herramientas, brinda una gran ayuda para encontrar soluciones a los problemas encontrados, y facilita el procesamiento y limpieza de la información. Por otro lado, permite graficar de manera analítica los distintos aspectos del dataset.

La metodología CRISP-DM ha facilitado la organización del proyecto, así como establecer un marco práctico y metodológico apto para desarrollar y optimizar los modelos de machine learning en un entorno controlable y escalable

4.2 Líneas futuras

Todo el trabajo fue realizado con las encuestas realizadas durante el 2019 a los estudiantes a nivel nacional. Por limitaciones en el hardware disponible para el procesamiento de datos, se realizaron las pruebas con una muestra aleatoria del total de la población. En el futuro, nuestra expectativa es poder probar el modelo final con datos del 2021 sobre las encuestas de Aprender y evaluar si el nivel de predicción se mantiene estable.

Bibliografía

Abdelmalik , M., Inza, I., & Larrañaga, P. (s.f.). *Clasificadores K-NN.* Universidad del Paıs Vasco–Euskal Herriko Unibertsitatea.

Azevedo, A., & Santos, M. (2008). KDD, semma and CRISP-DM: A parallel overview. *IADIS European Conference on Data Mining* . Amsterdam, The Netherlands.

Beinlich, I., Suermondt, H., Chavez, R., & Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *In proceedings of the 2nd European Conference on Artificial Intelligence in Medicine.* .

Bickmore, T. W. (1994). Real-Time Sensor Data Validation. *NASA Contractor Report 195295, National Aeronautics and Space Administration.*

Breiman, L. (1997). *Arcing The Edge.* California, US.

Britos, P., & García-Martínez, R. (2009). *Propuesta de Procesos de Explotación de Información.*

Britos, P., Felgaer, P., & García-Martínez, R. (2008). Bayesian Networks Optimization Based on Induction Learning Techniques. *In Artificial Intelligence in Theory and Practice II.* Boston: Springer: M. Bramer.

Britos, P., Jiménez Rey, E., & García-Martínez, E. (2008). Work in Progress: Programming Misunderstandings Discovering Process Based On Intelligent Data Mining Tools. *Proceedings 38th ASEE/IEEE Frontiers. Education Conference.*

Brownlee, J. (2020). *Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost.*

Chapman, P., Clinton, J., R., K., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0.*

Ezawa, K. J., & Schuermann, T. (1995). Fraud/Uncollectible Debt Detection Using a Bayesian Network Based Learning System: A Rare Binary Outcome with Mixed Data Structures. *Proceedings of the Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA,*, 157-166.

Fayyad, U. M. (1996). From data mining to knowledge discovery: an overview. *Advances in knowledge discovery and data mining*.

Felgaer, P. (2004). Optimización de Redes Bayesianas basado en Técnicas de Aprendizaje por Inducción. *Reportes Técnicos en Ingeniería del Software*, págs. 64-69.

Fogarty, J., Baker, R., & Hudson, S. (2005). Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *ACM International Conference Proceeding Series, Proceedings of Graphics Interface 2005.* Waterloo, Ontario, Canada.

Granado, E. C. (s.f.). *Manual de uso de Jupyter notebook para aplicaciones docentes.*

Grigori, D., Casati, F., Castellanos, M., Dayal, u., Sayal, M., & Shan, M. (2004). Business Process Intelligence. *Computers in Industry 53*, págs. 321-343.

Hair, J., Anderson, R., Tathman, R., & Black, W. (1999). *Análisis Multivariante.* Prentice Hall.

Harris, C., Millman, K., van der Walt, S., & y otros. (2020). Array programming with NumPy. *Springer Science and Business Media* (págs. 357-362). Nature 585.

Hasperué, W. (2013). *Extracción de Conocimiento en Grandes Bases de Datos Utilizando Estrategias Adaptativas.* Edulp.

Heckerman, D. (1995). A tutorial on learning bayesian networks. *Technical report MSR-TR-95-06, Microsoft research, Redmond, WA*.

Heckerman, D., Chickering, M., & Geiger, D. (1995). Learning bayesian networks, the combination of knowledge and statistical data. *Machine learning 20*, 197-243.

Hernandez, J., Ramirez Quintana, J., & Ferri Ramirez, C. (2004). *Introducción a la Minería de Datos.* Pearson Educación.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* (págs. 90-95). IEEE COMPUTER SOC.

Indec. (s.f.). *Quiénes somos: Instituto Nacional de Estadistica y Censos*. Obtenido de Instituto Nacional de Estadistica y Censos: https://www.indec.gob.ar/indec/web/Institucional-Indec-QuienesSomos-1

KDnuggets. (2014). *Encuesta sobre metodologías utilizadas en Data Mining.* Obtenido de https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Kisbye, P. (2010). *Test de Kolmogorov-Smirnov.* FaMAF.

Kluyver, T. R.-K. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (págs. 87-90). F. Loizides & B. Schmidt (Eds.).

Kononenko, I., & Cestnik, B. (1986). *Lymphography Data Set*. Obtenido de UCI Machine Learning: http://archive.ics.uci.edu/ml/datasets/Lymphography

Langseth, J., & Vivatrat, N. (2003). Why Proactive Business Intelligence is a Hallmark of the Real-Time Enterprise: Outward Bound. *Intelligent Enterprise 5*, págs. 34-41.

Mariscal, G., Marbán, Ó., González, Á., & Segovia, J. (2007). Hacia la Ingeniería de Data Mining: Un modelo de proceso para el desarrollo de proyectos. *II Congreso Español de Informática*, (págs. 139-148).

Mata, M. C., & Macassi, S. (1997). *Cómo elaborar muestras para los sondeos de audiencias.* Quito.

Michalski, R. (1983). A Theory and Methodology of Inductive Learning. *Artificial Intelligence*, *20*, págs. 111-161.

Michalski, R., Bratko, I., & Kubat, M. (1998). Machine Learning and Data Mining, Methods. *John Wiley & Sons*.

Microsoft. (2021). Obtenido de https://docs.microsoft.com/es-es/azure/machine-learning/how-to-tune-hyperparameters

Ministerio de Educación. (s.f.). *Análisis de respuestas a ítems de Matemática.* Obtenido de https://www.argentina.gob.ar/sites/default/files/analisis_de_respuestas_a_items_de_matematica.pdf

Ministerio de Educación. (2019). *Evaluación de la educación secundaria en Argentina 2019*. Obtenido de https://www.argentina.gob.ar/educacion/evaluacion-informacion-educativa/evaluacion-de-la-educacion-secundaria-en-argentina-2019

Ministerio de Educación. (s.f.). *Aprender*. Obtenido de www.argentina.gob.ar/educacion/evaluacion-informacion-educativa/aprender

Ministerio de Educación. (s.f.). *CUESTIONARIO DE ENCUESTAS APRENDER 2019.* Obtenido de https://drive.google.com/file/d/1UwFOUxcPabof0TxA-Bzt2PbhOX4bWAuH/view

Mitra, S., & Acharya, T. (2003). *Data mining: multimedia, soft computing and bioinformatics.* John Wiley & Sons.

Negash, S., & Gray, P. (2008). Business Intelligence. *En Handbook on Decision Support Systems 2, ed. F. Burstein y C. Holsapple*.

Pearl, J. (1988). Probabilisic reasoning in intelligent systems. *Morgan Kaufmann, San Mateo, CA*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, (págs. 2825-2830).

Perez, J., Henriques, M. F., Pazos, R., Cruz, L., Reyes, G., Salinas, J., & Mexicano, A. (2007). La IO Aplicada a la solucion de problemas regionales. *2do Taller Latino Liberoamericano de Investigacion de Operaciones*.

Piattini , M., & De Miguel, A. (1999). Diseño de Bases de Datos Relacionales. *Ra-Ma*.

Pramoditha, R. (s.f.). *blog Data Science 365.* Obtenido de https://medium.com/data-science-365

Python Software Foundation. Python Language Reference, version 2.7. (s.f.). Obtenido de http://www.python.org

Richaud de Minzi, M. C. (2008). Nuevas tendencias en psicometría. *Evaluar*, 1-19.

Rioja, L., Llorente, A., & Ramirez, B. (s.f.). *Regresión Logística: Fundamentos y aplicación a la investgación sociológa.*

Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias . *Revista Iberoamericana de Inteligencia Artificial.*

Rodríguez Montequín, M., Álvarez Cabal, V., Mesa Fernández, J., & González Valdés, A. (2003). *Metodologías para la realización de proyectos de data mining.*

Rohan, J. (2018). *Towards Data Science.* Obtenido de https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

SAS Institute. (s.f.). Obtenido de https://web.archive.org/web/20120308165638/http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html/

Secretaria de Evaluación e Información Educativa. (2019). *Aprender: Documento metodológico.* Ministerio de educacion Argentina.

Silvente, Hurtado, & Baños. (2013). Cómo aplicar árboles de decisión en SPSS. *Reire*.

Soloaga, A. (2018). *Akademus.* Obtenido de www.akademus.es

Soloaga, A. (2018). *Principales Usos de Python.* Akademus.

Sumathi, & Sivanandam. (2006). *Introduction to Data Mining and its Applications.*

The pandas development team. (2020). pandas-dev/pandas: Pandas. *Zenodo*, 10.5281/zenodo.3509134.

Violi, P. E. (s.f.). *Repositorio de proyecto TFI.* Obtenido de https://github.com/pevioli/TFI

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software* (pág. 3021). The Open Journal.

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques 2nd Edition.* Morgan Kaufmann.

Anexo A

DICCIONARIO DE DATOS DE ENCUESTAS APRENDER 2019

A.1 Lista de atributos Encuesta Aprender:

Variable	Descripción
cod_provincia	Número de jurisdicción
sector	Sector de gestión
ambito	Ámbito
claveseccion	Identificador de la sección
idestudiante	Identificador del caso
ap01_01	¿En qué mes y año naciste? mes
ap01_02	¿En qué mes y año naciste? año
ap02	Sexo
ap03	¿En qué país naciste?
ap04	¿En qué país nació tu mamá?
ap05	¿En qué país nació tu papá?
ap06	¿Vivís solo/a?
ap07	Si no vivís solo/a, ¿con cuántas personas vivís sin contarte a vos?
ap08_a	Marcá en el cuadro con quien/es vivís: Madre
ap08_b	Marcá en el cuadro con quien/es vivís: Padre
ap08_c	Marcá en el cuadro con quien/es vivís: Hermano/a
ap08_d	Marcá en el cuadro con quien/es vivís: Hijo/a
ap08_e	Marcá en el cuadro con quien/es vivís: Tío/a
ap08_f	Marcá en el cuadro con quien/es vivís: Abuelo/a
ap08_g	Marcá en el cuadro con quien/es vivís: Pareja
ap08_h	Marcá en el cuadro con quien/es vivís: Amigo/s o amiga/s
ap08_i	Marcá en el cuadro con quien/es vivís: Otros/as
ap09	¿Tenés hija/s o hijo/s?
ap10	En total ¿cuántas habitaciones hay en el lugar donde vivís, sin contar baño y cocina?
ap11_01	¿Cuántas de estas cosas hay en el lugar donde vivís? Televisor
ap11_02	¿Cuántas de estas cosas hay en el lugar donde vivís? Auto
ap11_03	¿Cuántas de estas cosas hay en el lugar donde vivís? Baño
ap11_04	¿Cuántas de estas cosas hay en el lugar donde vivís? Microondas
ap11_05	¿Cuántas de estas cosas hay en el lugar donde vivís? Aire acondicionado
ap11_06	¿Cuántas de estas cosas hay en el lugar donde vivís? Computadora de escritorio, laptop o notebook
ap11_07	¿Cuántas de estas cosas hay en el lugar donde vivís? Tablet

Variable	Descripción
ap11_08	¿Cuántas de estas cosas hay en el lugar donde vivís? Conexión a internet
ap12	¿Tenés celular propio?
ap13	¿Podés acceder a internet a través de tu celular?
ap14_01	En tu familia, ¿hay alguien que esté cursando estudios universitarios?
ap14_02	En tu familia, ¿hay alguien que haya finalizado estudios universitarios?
ap15	Aproximadamente, ¿cuántos libros hay donde vivís?
ap16	Máximo nivel educativo de la madre
ap17	Máximo nivel educativo del padre
ap18_01	En un día típico, ¿cuánto tiempo dedicás a las siguientes tareas? Cuidar a un/a hermano/a
ap18_02	En un día típico, ¿cuánto tiempo dedicás a las siguientes tareas? Cuidar de otro familiar
ap18_03	En un día típico, ¿cuánto tiempo dedicás a las siguientes tareas? Realizar tareas del hogar (cocinar, limpiar, lavar la ropa, hacer las compras, etc.)
ap19	En las últimas dos semanas, ¿ayudaste a tus padres o familiares en su trabajo?
ap21	En las últimas dos semanas ¿cuántos días trabajaste fuera de tu casa?
ap22	¿Recibís un pago por realizar ese trabajo fuera de tu casa?
ap23_01	En la última semana, ¿cuántas veces realizaste las siguientes actividades en tu tiempo libre, fuera de la escuela? Juntarme con mis amigos/as
ap23_02	En la última semana, ¿cuántas veces realizaste las siguientes actividades en tu tiempo libre, fuera de la escuela? Aprender idiomas
ap23_03	En la última semana, ¿cuántas veces realizaste las siguientes actividades en tu tiempo libre, fuera de la escuela? Leer libros diferentes de los de la escuela
ap23_04	En la última semana, ¿cuántas veces realizaste las siguientes actividades en tu tiempo libre, fuera de la escuela? Hacer deportes
ap23_05	En la última semana, ¿cuántas veces realizaste las siguientes actividades en tu tiempo libre, fuera de la escuela? Realizar actividades artísticas (clases de pintura, danzas, música, etc.)
ap23_06	En la última semana, ¿cuántas veces realizaste las siguientes actividades en tu tiempo libre, fuera de la escuela? Mirar televisión
ap24	¿Asististe a nivel inicial?
ap25_01	¿Repetiste algún año durante tu escolaridad? Primaria
ap25_02	¿Repetiste algún año durante tu escolaridad? Secundaria: ciclo básico (1° y 2° año si tu secundaria es 5 años; 1°,2° y 3° año si tu secundaria es de 6 años.)
ap25_03	¿Repetiste algún año durante tu escolaridad? Secundaria: ciclo

Variable	Descripción
	orientado (3°, 4° y 5°/6° año si tu secundaria es de 5 años; 4°, 5° y 6°/7°año si tu secundaria es de 6 años.)
ap26	En lo que va del año, ¿cuántas veces faltaste a la escuela?
ap27_a	En general, ¿cuáles son los principales motivos por los que faltaste a la escuela? Por enfermedad
ap27_b	En general, ¿cuáles son los principales motivos por los que faltaste a la escuela? Porque no tenía ganas de ir a la escuela
ap27_c	En general, ¿cuáles son los principales motivos por los que faltaste a la escuela? Por ayudar con las tareas en mi casa
ap27_d	En general, ¿cuáles son los principales motivos por los que faltaste a la escuela? Por problemas de acceso a la escuela por temas climáticos o de transporte
ap27_e	En general, ¿cuáles son los principales motivos por los que faltaste a la escuela? Por responsabilidades laborales
ap27_f	En general, ¿cuáles son los principales motivos por los que faltaste a la escuela? Otros
ap28_01	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre la convivencia en tu escuela? Hay un ambiente de buena convivencia
ap28_02	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre la convivencia en tu escuela? Los estudiantes nos llevamos bien
ap28_03	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre la convivencia en tu escuela? Los docentes se llevan bien con nosotros
ap28_04	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre la convivencia en tu escuela? Yo me siento bien de venir a esta escuela
ap29_01	¿Con qué frecuencia en tu escuela pasa lo siguiente? Discriminan por alguna característica personal o familiar (religión, nacionalidad, condición de género, discapacidad.)
ap29_02	¿Con qué frecuencia en tu escuela pasa lo siguiente? Discriminan por aspectos físicos
ap29_03	¿Con qué frecuencia en tu escuela pasa lo siguiente? Amenazan o agreden a otros compañeros
ap29_04	¿Con qué frecuencia en tu escuela pasa lo siguiente? Amenazan o agreden a docentes
ap29_05	¿Con qué frecuencia en tu escuela pasa lo siguiente? Amenazan o agreden a otros compañeros por redes sociales
ap29_06	¿Con qué frecuencia en tu escuela pasa lo siguiente? Amenazan o agreden a docentes por redes sociales
ap29_07	¿Con qué frecuencia en tu escuela pasa lo siguiente? Dañan las cosas de la escuela
ap30	¿Tu escuela cuenta con normas de convivencia?
ap31_01	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre las normas de convivencia de tu escuela? Conozco las normas de convivencia de mi escuela
ap31_02	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre las normas de convivencia de tu escuela? Las normas de convivencia

Variable	Descripción
	son respetadas por los estudiantes
ap31_03	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre las normas de convivencia de tu escuela? Las normas de convivencia son respetadas por los docentes
ap31_04	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre las normas de convivencia de tu escuela? Los estudiantes hemos participado en la definición de las normas de convivencia de esta escuela
ap31_05	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre las normas de convivencia de tu escuela? En casos de faltas a las normas de convivencia se aplican las medidas establecidas por ellas
ap31_06	¿Cuán de acuerdo estás con las siguientes afirmaciones sobre las normas de convivencia de tu escuela? Hemos elaborado normas de convivencia para nuestra aula
ap32_a	Cuando ocurre un conflicto en la escuela, ¿cuál es la forma más frecuente para resolverlo? Tratando el conflicto con los adultos de la escuela
ap32_b	Cuando ocurre un conflicto en la escuela, ¿cuál es la forma más frecuente para resolverlo? Tratando el conflicto con todos los involucrados
ap32_c	Cuando ocurre un conflicto en la escuela, ¿cuál es la forma más frecuente para resolverlo? Tratando el conflicto con el equipo de orientación escolar
ap32_d	Cuando ocurre un conflicto en la escuela, ¿cuál es la forma más frecuente para resolverlo? No dándole importancia y dejándolo pasar
ap32_e	Cuando ocurre un conflicto en la escuela, ¿cuál es la forma más frecuente para resolverlo? Con sanciones
ap32_f	Cuando ocurre un conflicto en la escuela, ¿cuál es la forma más frecuente para resolverlo? Otro
ap33_01_a	¿Ofreció tu escuela los siguientes espacios a los estudiantes? Orientación vocacional para la elección de la carrera
ap33_01_b	Si los ofreció, ¿asististe? Orientación vocacional para la elección de la carrera
ap33_02_a	¿Ofreció tu escuela los siguientes espacios a los estudiantes? Charlas informativas sobre oferta educativa en universidades
ap33_02_b	Si los ofreció, ¿asististe? Charlas informativas sobre oferta educativa en universidades
ap33_03_a	¿Ofreció tu escuela los siguientes espacios a los estudiantes? Charlas informativas sobre oferta educativa terciaria
ap33_03_b	Si los ofreció, ¿asististe? Charlas informativas sobre oferta educativa terciaria
ap33_04_a	¿Ofreció tu escuela los siguientes espacios a los estudiantes? Charlas informativas sobre la oferta educativa en tu localidad y zonas cercanas
ap33_04_b	Si los ofreció, ¿asististe? Charlas informativas sobre la oferta educativa en tu localidad y zonas cercanas

Variable	Descripción
ap33_01bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Estoy a gusto en esta clase.
ap33_02bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: En esta clase suele haber orden.
ap33_03bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: En esta clase puedo participar.
ap33_04bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Hay buen ambiente en esta clase.
ap33_05bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Se tarda poco en empezar a trabajar.
ap33_06bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Los docentes coordinan la clase hasta el final.
ap33_07bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Respetamos a los docentes.
ap33_08bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Sabemos siempre lo que tenemos que hacer..
ap33_09bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Nuestra relación con los docentes es buena.
ap33_10bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Los estudiantes nos ayudamos unos a otros.
ap33_11bis	Por favor, leé cada una de las frases atentamente, pensá cómo es el ambiente de tu clase y contestá según la siguiente escala: “Nunca”, “A veces”; “A menudo” o “Siempre”: Hay el silencio suficiente para aprovechar cada materia.
ap34	¿Qué vas a hacer cuando termines el secundario?
ap35_01	Te pedimos que marques cuán de acuerdo estás con las siguientes afirmaciones sobre tu experiencia en la escuela secundaria: Me llevo vínculos y amistades muy importantes en mi vida

Variable	Descripción
ap35_02	Te pedimos que marques cuán de acuerdo estás con las siguientes afirmaciones sobre tu experiencia en la escuela secundaria: Estoy preparado/a para empezar mi búsqueda laboral
ap35_03	Te pedimos que marques cuán de acuerdo estás con las siguientes afirmaciones sobre tu experiencia en la escuela secundaria: Estoy preparado/a para seguir estudiando
ap35_04	Te pedimos que marques cuán de acuerdo estás con las siguientes afirmaciones sobre tu experiencia en la escuela secundaria: Estoy preparado/a para seguir estudiando
ap35_05	Te pedimos que marques cuán de acuerdo estás con las siguientes afirmaciones sobre tu experiencia en la escuela secundaria: Me ayudó a formarme como un ciudadano responsable con mi comunidad
ap36_a	¿Cómo se abordan en tu escuela los temas vinculados a Educación Sexual Integral (ESI)? Como una materia más
ap36_b	¿Cómo se abordan en tu escuela los temas vinculados a Educación Sexual Integral (ESI)? De manera transversal a todas las materias
ap36_c	¿Cómo se abordan en tu escuela los temas vinculados a Educación Sexual Integral (ESI)? En clases especiales con especialistas invitados/as
ap36_d	¿Cómo se abordan en tu escuela los temas vinculados a Educación Sexual Integral (ESI)? En clases especiales con equipo de la escuela (directivo, docentes, orientadores, preceptores etc.)
ap36_e	¿Cómo se abordan en tu escuela los temas vinculados a Educación Sexual Integral (ESI)? No se ven estos temas en mi escuela
ap36_f	¿Cómo se abordan en tu escuela los temas vinculados a Educación Sexual Integral (ESI)? Otros
ap37_a	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? El cuerpo que cambia, la autonomía y su construcción progresiva
ap37_b	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? Las distintas formas de ser joven según los contextos y las experiencias de vida
ap37_c	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? Construcción de identidad y de proyecto de vida
ap37_d	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? Los patrones hegemónicos de belleza y su relación con el consumo
ap37_e	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? Reproducción, embarazo, parto, maternidad y paternidad desde un abordaje integral
ap37_f	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? El embarazo no intencional en la adolescencia: los métodos anticonceptivos
ap37_g	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? La prevención de infecciones de transmisión sexual
ap37_h	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? Los marcos legales para el acceso a los servicios de salud sexual

Variable	Descripción
ap37_i	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? La pareja, el amor y el cuidado mutuo en las relaciones afectivas
ap37_j	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? Mirada hacia la violencia de género en el noviazgo
ap37_k	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? El derecho de las personas a vivir su sexualidad de acuerdo a sus convicciones y preferencias en el marco del respeto y cuidado por uno mismo y por los/as otros/as
ap37_l	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? La vulneración de derechos sexuales: La discriminación, la violencia, el acoso, el abuso, el maltrato, la explotación sexual y trata
ap37_m	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? La violencia de género en la adolescencia
ap37_n	Durante tu secundaria, ¿trabajaste estos temas en tu escuela? Prevención del grooming. Redes sociales y sexualidad
ap38_a	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? El cuerpo que cambia, la autonomía y su construcción progresiva
ap38_b	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? Las distintas formas de ser joven según los contextos y las experiencias de vida
ap38_c	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? Construcción de identidad y de proyecto de vida
ap38_d	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? Los patrones hegemónicos de belleza y su relación con el consumo
ap38_e	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? Reproducción, embarazo, parto, maternidad y paternidad desde un abordaje integral
ap38_f	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? El embarazo no intencional en la adolescencia
ap38_g	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? La prevención de infecciones de transmisión sexual
ap38_h	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? Los marcos legales para el acceso a los servicios de salud sexual
ap38_i	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? La pareja, el amor y el cuidado mutuo en las relaciones afectivas
ap38_j	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? Mirada hacia la violencia de género en el noviazgo
ap38_k	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? El derecho de las

Variable	Descripción
	personas a vivir su sexualidad de acuerdo a sus convicciones y preferencias en el marco del respeto y cuidado por uno mismo y (...)
ap38_l	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? La vulneración de derechos sexuales
ap38_m	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? La violencia de género en la adolescencia
ap38_n	¿En cuáles de estos temas relacionados con la ESI, según tu opinión, tu escuela debería profundizar? Prevención del grooming. Redes sociales y sexualidad
ap39_01	En general, ¿cómo te resultan las siguientes actividades? Comprender un texto
ap39_02	En general, ¿cómo te resultan las siguientes actividades? Escribir un texto
ap39_03	En general, ¿cómo te resultan las siguientes actividades? Exponer oralmente
ap39_04	En general, ¿cómo te resultan las siguientes actividades? Resolver problemas y ejercicios
ap40_01	¿En qué medida estás de acuerdo con las siguientes afirmaciones? Disfruto estudiando Matemática
ap40_02	¿En qué medida estás de acuerdo con las siguientes afirmaciones? Me interesan las clases de Matemática en mi escuela
ap40_03	¿En qué medida estás de acuerdo con las siguientes afirmaciones? Me esfuerzo para que me vaya bien en Matemática
ap40_04	¿En qué medida estás de acuerdo con las siguientes afirmaciones? Si me lo propongo, puedo ser bueno en Matemática
ap40_05	¿En qué medida estás de acuerdo con las siguientes afirmaciones? Aprender Matemática es importante, porque la voy a necesitar en estudios futuros
ap40_06	¿En qué medida estás de acuerdo con las siguientes afirmaciones? Aprendí muchas cosas en Matemática que me ayudarán a conseguir un trabajo
ap41_01	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A investigar y a redactar un informe sobre un tema
ap41_02	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A tomar apuntes de las clases y de los libros
ap41_03	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A tomar una posición propia y fundada respecto a diferentes temas
ap41_04	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A resolver situaciones que representan un desafío

Variable	Descripción
ap41_05	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A hacer presentaciones orales
ap41_06	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A identificar mis dificultades en el aprendizaje y a organizarme para trabajarlas
ap41_07	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A expresar mis ideas
ap41_08	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A escuchar a los demás, respetando sus puntos de vista
ap41_09	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A analizar las consecuencias que tienen mis acciones sobre los demás
ap41_10	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A asumir compromisos con mis tareas
ap41_11	Pensando en tu experiencia en la escuela secundaria, te pedimos que marques tu nivel de acuerdo con las siguientes afirmaciones: A trabajar con otros
ap42_01	¿En tu escuela con qué frecuencia se realizan las siguientes acciones frente a las dificultades en el aprendizaje de los estudiantes? Se proporciona clases de apoyo a los estudiantes con dificultades en alguna materia
ap42_02	¿En tu escuela con qué frecuencia se realizan las siguientes acciones frente a las dificultades en el aprendizaje de los estudiantes? Se hace seguimiento personalizado a los estudiantes con dificultades para comprender un tema
ap42_03	¿En tu escuela con qué frecuencia se realizan las siguientes acciones frente a las dificultades en el aprendizaje de los estudiantes? Se adapta la clase a las necesidades y conocimientos de los estudiantes con dificultades
ap42_04	¿En tu escuela con qué frecuencia se realizan las siguientes acciones frente a las dificultades en el aprendizaje de los estudiantes? Se consideran diferentes formas de evaluar los aprendizajes según las necesidades de los estudiantes
ap43_a	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Robótica y programación
ap43_b	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Cambio climático y calentamiento global
ap43_c	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Salud mundial (epidemias, desnutrición, etc.)
ap43_d	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Conflictos internacionales y sus consecuencias (migraciones, refugiados, etc.)

Variable	Descripción
ap43_e	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Desarrollo sustentable
ap43_f	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Oferta educativa terciaria y universitaria local y provincial
ap43_g	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Temas vinculados a la demanda laboral local y regional
ap43_h	De los siguientes temas, ¿en cuáles te parece que la escuela debería profundizar? Otros
ap44_01	¿Con qué frecuencia trabajás con estos dispositivos TIC en clase? Computadora de escritorio
ap44_02	¿Con qué frecuencia trabajás con estos dispositivos TIC en clase? Notebook, Netbook
ap44_03	¿Con qué frecuencia trabajás con estos dispositivos TIC en clase? Tablet
ap44_04	¿Con qué frecuencia trabajás con estos dispositivos TIC en clase? Carro digital
ap44_05	¿Con qué frecuencia trabajás con estos dispositivos TIC en clase? Celular
ap45	¿Los docentes permiten usar el celular en el aula?
ap46_01	¿En alguna materia, trabajaron con los docentes sobre los siguientes temas relacionados con el uso de internet? Cómo buscar información en internet y evaluar si es confiable
ap46_02	¿En alguna materia, trabajaron con los docentes sobre los siguientes temas relacionados con el uso de internet? Cómo cuidar nuestros datos personales y nuestra privacidad en línea
ap46_03	¿En alguna materia, trabajaron con los docentes sobre los siguientes temas relacionados con el uso de internet? Cómo funcionan los motores de búsqueda
ap46_04	¿En alguna materia, trabajaron con los docentes sobre los siguientes temas relacionados con el uso de internet? Cómo citar la información que encontramos en internet en nuestros trabajos
ap46_05	¿En alguna materia, trabajaron con los docentes sobre los siguientes temas relacionados con el uso de internet? Identificar el propósito de cada página (informar, comunicar, persuadir, entretener.)
ap47_01	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Escribir textos
ap47_02	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Leer textos en pantalla
ap47_03	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Responder cuestionarios
ap47_04	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Buscar información en internet
ap47_05	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Usar simulaciones

Variable	Descripción
ap47_06	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Hacer animaciones
ap47_07	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Producir/editar fotos, videos o audios
ap47_08	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Usar juegos educativos
ap47_09	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Desarrollar páginas web y/o aplicaciones
ap47_10	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Armar robots
ap47_11	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Hacer cálculos y usar funciones de planillas de cálculo
ap47_12	¿Con qué frecuencia realizás este tipo de actividades con la computadora, celular o tablet en clase? Programar/escribir en lenguaje de programación como scratch u otro
ap48_a	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Lengua y Literatura.
ap48_b	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Historia y Geografía
ap48_c	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Matemática
ap48_d	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Artes (música, plástica, teatro)
ap48_e	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Química y Física
ap48_f	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Biología
ap48_g	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Informática/computación
ap48_h	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Otros
ap48_i	Te pedimos que selecciones las materias en las que más usas la computadora, celular y/o tablet: Ninguna
ap49_01	¿A qué edad empezaste a usar computadora o tablet?
ap49_02	¿A qué edad empezaste a usar celular?
ap50_01	Cuando estás en tu casa ¿con qué frecuencia usás la computadora, celular o tablet para hacer actividades escolares?: Buscar información en internet para alguna materia
ap50_02	Cuando estás en tu casa ¿con qué frecuencia usás la computadora, celular o tablet para hacer actividades escolares?: Escribir trabajos prácticos
ap50_03	Cuando estás en tu casa ¿con qué frecuencia usás la computadora, celular o tablet para hacer actividades escolares?: Contactarte con tus compañeros y resolver tareas

Variable	Descripción
ap50_04	Cuando estás en tu casa ¿con qué frecuencia usás la computadora, celular o tablet para hacer actividades escolares?: Hacer videos o sacar fotos para trabajos de la escuela
ap50_05	Cuando estás en tu casa ¿con qué frecuencia usás la computadora, celular o tablet para hacer actividades escolares?: Ver tutoriales sobre temas de la escuela
ap51_01	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Chatear o mandar mensajes
ap51_02	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Buscar información en internet
ap51_03	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Hacer música y videos (Grabar, editar o mezclar.)
ap51_04	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Ver una película, serie o videos online
ap51_05	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Utilizar redes sociales (Instagram, Snapchat, YouTube, Twitter, entre otras.)
ap51_06	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Leer correos electrónicos
ap51_07	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Participar de un grupo de debate o foro en internet
ap51_08	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Aprender idiomas
ap51_09	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Leer noticias en internet
ap51_10	En la última semana, ¿con qué frecuencia te dedicaste a las siguientes actividades?: Jugar juegos de consola o computadora online
ponder	Factor de expansión cuestionario complementario
lpondera	Factor de expansión prueba de Lengua (censo)
mpondera	Factor de expansión prueba de Matemática (censo)
ldesemp	Nivel de desempeño en Lengua
mdesemp	Nivel de desempeño en Matemática
TEL	Puntaje en Lengua
TEM	Puntaje en Matemática
modelo	Modelo de prueba
isocioa_puntaje	Puntaje NSE Pondera
isocioa	Indice socioeconómico del estudiante
isocioal_puntaje	Puntaje NSE lpondera
isocioal	Indice socioeconómico del estudiante ponderador Lengua
isocioam_puntaje	Puntaje NSE mpondera
isocioam	Indice socioeconómico del estudiante ponderador Matemática
repitencia_dicotomica	Repitencia dicotómica
jardín	Asistencia al nivel inicial recodificada

Variable	Descripción
ap37_dicotomica	Se trataron temas ESI - Dicotomica (Ap37)
ap29_01.dico	Discriminan por alguna característica personal o familiar (religión, nacionalidad, condición de género, discapacidad.)
ap29_02.dico	Discriminan por aspectos físicos
ap29_03.dico	Amenazan o agreden a otros compañeros
ap29_04.dico	Amenazan o agreden a docentes
ap29_05.dico	Amenazan o agreden a otros compañeros por redes sociales
ap29_06.dico	Amenazan o agreden a docentes por redes sociales
ap29_07.dico	Dañan las cosas de la escuela
ap26_rec	Inasistencias en el año
trabaja_fuera_hogar	Trabaja fuera del hogar
trabaja_fuera_hogar_remunerado	Trabajo remunerado fuera del hogar
migración	Condición migratoria
edadA_junio2019	Edad a Junio 2019
sobreedad	Sobreedad
infraestructura	Índice de Infraestructura (de la escuela) - Puntaje
iinfraestructura	Indice de Infraestructura escolar (de la escuela) - Categorizado
ap42_01rec	Se proporciona clases de apoyo a los estudiantes con dificultades en alguna materia
ap42_02rec	Se hace seguimiento personalizado a los estudiantes con dificultades para comprender un tema
ap42_03rec	Se adapta la clase a las necesidades y conocimientos de los estudiantes con dificultades
ap42_04rec	Se consideran diferentes formas de evaluar los aprendizajes según las necesidades de los estudiantes

A.2 Respuestas posibles Encuestas Aprender:

Variable	Valores	Referencia
cod_provincia	02	Ciudad Autónoma de Buenos Aires
	06	Buenos aires
	10	Catamarca
	14	Córdoba
	18	Corrientes
	22	Chaco
	26	Chubut
	30	Entre Ríos
	34	Formosa
	38	Jujuy
	42	La Pampa

Variable	Valores	Referencia
	46	La Rioja
	50	Mendoza
	54	Misiones
	58	Neuquén
	62	Río Negro
	66	Salta
	70	San Juan
	74	San Luis
	78	Santa Cruz
	82	Santa Fe
	86	Santiago del Estero
	90	Tucumán
	94	Tierra del Fuego
sector	1	Estatal
	2	Privado
ambito	1	Urbano
	2	Rural
ap01_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	-1	No Corresponde
	1	Enero
	2	Febrero
	3	Marzo
	4	Abril
	5	Mayo
	6	Junio
	7	Julio
	8	Agosto
	9	Septiembre
	10	Octubre
	11	Noviembre
	12	Diciembre
ap01_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	-1	No Corresponde
	1	1998
	2	1999
	3	2000
	4	2001
	5	2002

Variable	Valores	Referencia
	6	2003
ap02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Varón
	2	Mujer
ap03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Argentina
	2	Bolivia
	3	Brasil
	4	Chile
	5	Paraguay
	6	Perú
	7	Uruguay
	8	Colombia
	9	Venezuela
	10	México
	11	Otro país de América
	12	País de Europa
	13	País de Asia
	14	País de África
	15	País de Oceanía
ap04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Argentina
	2	Bolivia
	3	Brasil
	4	Chile
	5	Paraguay
	6	Perú
	7	Uruguay
	8	Colombia
	9	Venezuela
	10	México
	11	Otro país de América
	12	País de Europa
	13	País de Asia
	14	País de África
	15	País de Oceanía

Variable	Valores	Referencia
ap05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Argentina
	2	Bolivia
	3	Brasil
	4	Chile
	5	Paraguay
	6	Perú
	7	Uruguay
	8	Colombia
	9	Venezuela
	10	México
	11	Otro país de América
	12	País de Europa
	13	País de Asia
	14	País de África
	15	País de Oceanía
ap06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap07	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	1
	2	2
	3	3
	4	4
	5	5
	6	6
	7	7 o más
ap08_a	-9	Blanco
	1	Sí
ap08_b	-9	Blanco
	1	Sí
ap08_c	-9	Blanco
	1	Sí
ap08_d	-9	Blanco
	1	Sí
ap08_e	-9	Blanco

Variable	Valores	Referencia
	1	Sí
ap08_f	-9	Blanco
	1	Sí
ap08_g	-9	Blanco
	1	Sí
ap08_h	-9	Blanco
	1	Sí
ap08_i	-9	Blanco
	1	Sí
ap09	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap10	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	1
	2	2
	3	3
	4	4
	5	5
	6	6
	7	7 o más
ap11_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno
	2	Uno
	3	Dos
	4	Tres o más
ap11_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno
	2	Uno
	3	Dos
	4	Tres o más
ap11_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno

Variable	Valores	Referencia
	2	Uno
	3	Dos
	4	Tres o más
ap11_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno
	2	Uno
	3	Dos
	4	Tres o más
ap11_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno
	2	Uno
	3	Dos
	4	Tres o más
ap11_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno
	2	Uno
	3	Dos
	4	Tres o más
ap11_07	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno
	2	Uno
	3	Dos
	4	Tres o más
ap11_08	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap12	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap13	-9	Blanco

Variable	Valores	Referencia
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap14_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap14_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap15	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	No hay libros
	2	De 1 a 50 libros
	3	De 51 a 100 libros
	4	Más de 100 libros
	5	No sé
ap16	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Primaria incompleta
	2	Primaria Completa
	3	Secundaria incompleta
	4	Secundaria completa
	5	Educación Superior Técnica
	6	Educación Superior universitaria
	7	Posgrado
	8	No se
ap17	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Primaria incompleta
	2	Primaria Completa
	3	Secundaria incompleta
	4	Secundaria completa
	5	Educación Superior Técnica
	6	Educación Superior universitaria

Variable	Valores	Referencia
	7	Posgrado
	8	No se
ap18_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Menos de una hora
	2	Entre una y tres horas
	3	Más de tres horas
	4	No le dedico tiempo
ap18_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Menos de una hora
	2	Entre una y tres horas
	3	Más de tres horas
	4	No le dedico tiempo
ap18_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Menos de una hora
	2	Entre una y tres horas
	3	Más de tres horas
	4	No le dedico tiempo
ap19	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap21	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguno
	2	1 día
	3	Entre 2 y 4 días
	4	Entre 5 y 7 días
	5	Entre 8 y 10 días
	6	Más de 10 días
ap22	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No

Variable	Valores	Referencia
ap23_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Una o dos veces
	3	Tres o cuatro veces
	4	Cinco o más veces
ap23_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Una o dos veces
	3	Tres o cuatro veces
	4	Cinco o más veces
ap23_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Una o dos veces
	3	Tres o cuatro veces
	4	Cinco o más veces
ap23_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Una o dos veces
	3	Tres o cuatro veces
	4	Cinco o más veces
ap23_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Una o dos veces
	3	Tres o cuatro veces
	4	Cinco o más veces
ap23_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Una o dos veces
	3	Tres o cuatro veces
	4	Cinco o más veces

Variable	Valores	Referencia
ap24	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí, antes de los cuatro años
	2	Sí, desde sala de 4
	3	Sí, en sala de 5
	4	No asistí
ap25_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	No
	2	Sí, una vez
	3	Sí, dos o más veces
ap25_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	No
	2	Sí, una vez
	3	Sí, dos o más veces
ap25_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	No
	2	Sí, una vez
	3	Sí, dos o más veces
ap26	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Ninguna vez
	2	Hasta 5 veces
	3	Entre 10 y 15 veces
	4	Entre 16 y 24 veces
	5	Más de 24 veces
ap27_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap27_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap27_c	-9	Blanco

Variable	Valores	Referencia
	-8	No disponible
	-6	Multimarca
	1	Sí
ap27_d	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap27_e	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap27_f	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap28_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap28_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap28_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap28_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo

Variable	Valores	Referencia
	4	Muy de acuerdo
ap29_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap29_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap29_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap29_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap29_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap29_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces

Variable	Valores	Referencia
	4	Siempre
ap29_07	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap30	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap31_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap31_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap31_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap31_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap31_05	-9	Blanco

Variable	Valores	Referencia
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap31_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap32_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap32_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap32_c	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap32_d	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap32_e	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap32_f	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_01_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_01_b	-9	Blanco

Variable	Valores	Referencia
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_02_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_02_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_03_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_03_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_04_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_04_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap33_01bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_02bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_03bis	-9	Blanco

Variable	Valores	Referencia
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_04bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_05bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_06bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_07bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_08bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_09bis	-9	Blanco

Variable	Valores	Referencia
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_10bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap33_11bis	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	A menudo
	4	Siempre
ap34	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Seguir estudiando en educación terciaria
	2	Seguir estudiando en educación universitaria
	3	Trabajar y seguir estudiando en educación terciaria
	4	Trabajar y seguir estudiando en educación universitaria
	5	Trabajar
	6	Aún no lo sé
ap35_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap35_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo

Variable	Valores	Referencia
	4	Muy de acuerdo
ap35_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap35_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap35_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap36_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap36_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap36_c	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap36_d	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap36_e	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí

Variable	Valores	Referencia
ap36_f	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_c	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_d	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_e	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_f	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_g	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_h	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_i	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_j	-9	Blanco
	-8	No disponible

Variable	Valores	Referencia
	-6	Multimarca
	1	Sí
ap37_k	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_l	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_m	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap37_n	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_c	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_d	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_e	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_f	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí

Variable	Valores	Referencia
ap38_g	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_h	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_i	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_j	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_k	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_l	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_m	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap38_n	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap39_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Muy difícil
	2	Difícil
	3	Fácil
	4	Muy fácil
ap39_02	-9	Blanco
	-8	No disponible
	-6	Multimarca

Variable	Valores	Referencia
	1	Muy difícil
	2	Difícil
	3	Fácil
	4	Muy fácil
ap39_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Muy difícil
	2	Difícil
	3	Fácil
	4	Muy fácil
ap39_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Muy difícil
	2	Difícil
	3	Fácil
	4	Muy fácil
ap40_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap40_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap40_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap40_04	-9	Blanco
	-8	No disponible
	-6	Multimarca

Variable	Valores	Referencia
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap40_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap40_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_04	-9	Blanco
	-8	No disponible
	-6	Multimarca

Variable	Valores	Referencia
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_07	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_08	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_09	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_10	-9	Blanco
	-8	No disponible
	-6	Multimarca

Variable	Valores	Referencia
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap41_11	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nada de acuerdo
	2	Poco de acuerdo
	3	De acuerdo
	4	Muy de acuerdo
ap42_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap42_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap42_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap42_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap43_a	-9	Blanco
	-8	No disponible
	-6	Multimarca

Variable	Valores	Referencia
	1	Sí
ap43_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap43_c	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap43_d	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap43_e	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap43_f	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap43_g	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap43_h	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap44_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	La mayoría de las veces
	4	Siempre
ap44_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	La mayoría de las veces

Variable	Valores	Referencia
	4	Siempre
ap44_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	La mayoría de las veces
	4	Siempre
ap44_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	La mayoría de las veces
	4	Siempre
ap44_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	A veces
	3	La mayoría de las veces
	4	Siempre
ap45	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap46_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap46_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap46_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No

Variable	Valores	Referencia
ap46_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap46_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
	2	No
ap47_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca

Variable	Valores	Referencia
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_07	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_08	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_09	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_10	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_11	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca

Variable	Valores	Referencia
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap47_12	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap48_a	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap48_b	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap48_c	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap48_d	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap48_e	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap48_f	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap48_g	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap48_h	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí

Variable	Valores	Referencia
ap48_i	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sí
ap49_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Antes de los 6 años
	2	Entre los 6 y los 9 años
	3	Desde los 10 años
	4	Nunca usé
ap49_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Antes de los 6 años
	2	Entre los 6 y los 9 años
	3	Desde los 10 años
	4	Nunca usé
ap50_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap50_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap50_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap50_04	-9	Blanco
	-8	No disponible
	-6	Multimarca

Variable	Valores	Referencia
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap50_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Algunas veces
	3	La mayoría de las veces
	4	Siempre
ap51_01	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_02	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_03	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_04	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día

Variable	Valores	Referencia
	5	Todo el tiempo
ap51_05	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_06	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_07	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_08	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_09	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ap51_10	-9	Blanco

Variable	Valores	Referencia
	-8	No disponible
	-6	Multimarca
	1	Nunca
	2	Por lo menos una vez por semana
	3	Por lo menos una vez por día
	4	Más de una vez por día
	5	Todo el tiempo
ldesemp	1	Por debajo del nivel básico
	2	Básico
	3	Satisfactorio
	4	Avanzado
mdesemp	1	Por debajo del nivel básico
	2	Básico
	3	Satisfactorio
	4	Avanzado
isocioa	1	Bajo
	2	Medio
	3	Alto
isocioal	1	Bajo
	2	Medio
	3	Alto
isocioam	1	Bajo
	2	Medio
	3	Alto
repitencia_dicotomica	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Repitió en primaria y/o secundaria
	2	No repitió en primaria ni en secundaria
	9	No sabe/No contesta
jardín	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Sala de 3 o anterior
	2	Sala de 4 o 5
	3	No asistió
ap37_dicotomica	1	Si (al menos un tema)
	2	No (Ningún tema)
ap29_01.dico	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap29_02.dico	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca

Variable	Valores	Referencia
ap29_03.dico	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap29_04.dico	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap29_05.dico	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap29_06.dico	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap29_07.dico	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap26_rec	1	Ninguna inasistencia
	2	Hasta 5 veces
	3	Entre 10 y 24 veces
	4	Más de 24 veces
trabaja_fuera_hogar	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Si
	2	No
trabaja_fuera_hogar_remunerado	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Trabaja remuneradamente fuera del hogar
	2	Trabaja fuera del hogar sin remuneración
		No trabaja fuera del hogar
migración	-9	Blanco
	-8	No disponible
	-6	Multimarca
	1	Hogar migrante
	2	Hogar no migrante
	9	No sabe/No contesta
sobreedad	0	Menos de 17 años
	1	Edad teórica para el grado (17 años al 30 de junio)
	2	1 año de sobreedad (18 años al 30 de junio)
	3	2 años de sobreedad (19 años al 30 de junio)
	4	3 años o más de sobreedad (20 años o más al 30 de junio)
iinfraestructura	-99	Estudiantes de CUES no participantes en directivos
	-1	Estudiantes de CUES sin índice por no tener respuesta en todas las variables en directivos
	1	Bajo
	2	Medio
	3	Alto

Variable	Valores	Referencia
ap42_01rec	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap42_02rec	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap42_03rec	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca
ap42_04rec	1	Siempre/La mayoría de las veces
	2	Algunas veces/Nunca

A.3 Repositorio con Script y Bases utilizadas

GitHub: https://github.com/pevioli/TFI