UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

PREDICCIÓN Y DESCRIPCIÓN DE LA EXCLUSIÓN EDUCATIVA DEL SISTEMA ESCOLAR REGULAR CHILENO, CIENCIA DE DATOS PARA LA INNOVACIÓN PÚBLICA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

DIEGO ERNESTO IBÁÑEZ IRRIBARRA

PROFESOR GUÍA:

VÍCTOR LUIS PÉREZ VERA

PROFESOR CO-GUÍA:

FELIPE ARTURO TOBAR HENRÍQUEZ

MIEMBROS DE LA COMISIÓN:

PATRICIO ANDRÉS RODRÍGUEZ VALDÉS
HUGO ENRIQUE VÁSQUEZ GUARDAMAGNA

Este trabajo ha sido parcialmente financiado por
INSTITUTO SISTEMAS COMPLEJOS DE INGENIERÍA
(CONICYT – PIA – FB0816)

SANTIAGO DE CHILE

2018

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Diego Ernesto Ibáñez Irribarra
FECHA: 15/12/2018
PROFESOR GUÍA: Víctor Luis Pérez Vera

PREDICCIÓN Y DESCRIPCIÓN DE LA EXCLUSIÓN EDUCATIVA DEL SISTEMA ESCOLAR REGULAR CHILENO, CIENCIA DE DATOS PARA LA INNOVACIÓN PÚBLICA

La presente memoria de título desarrolla la predicción y descripción a nivel individual del potencial de exclusión educativa, también denominada deserción escolar, de les estudiantes de enseñanza básica y media del sistema escolar regular chileno, a través del prototipado de herramientas digitales. Con el fin de facilitar, mediante la detección temprana, y aportar entendimiento del fenómeno, a procesos de innovación pública para la prevención efectiva.

Como aporte a los antecedentes del fenómeno de la exclusión educativa escolar chilena, se evidencia que *del orden del 50% de les estudiantes excluides del sistema escolar año a año, asisten hasta diciembre a clases, con asistencia mes a mes en promedio superior al 80% para la enseñanza básica, y en promedio superior al 80% general anual para la enseñanza media*. Este nuevo antecedente cuantitativo del fenómeno plantea preguntas sobre la pertinencia del actual proceso de cierre anual escolar, y sobre el enfoque del actuar institucional para la prevención durante el año, y en vacaciones y cambios de ciclo.

Se utilizan 20 bases de datos de 4 instituciones distintas, disponibles transversalmente para cada estudiante. Para el aprendizaje de máquinas el conjunto de datos describe el caso de la exclusión educativa de la matrícula regular pública, municipal y particular subvencionada; de enseñanza básica, media científico humanista y técnico profesional. Para el aprendizaje supervisado el conjunto de entrenamiento (balanceado entre clases: excluides 50% y no excluides 50%) describe a cada estudiante en 2016-2017, y los conjuntos validación y test (balanceo natural de clases: 97%-3%) corresponden al 2017-2018. Para el aprendizaje no supervisado, se utilizan conjuntamente les excluides de ambos periodos.

Para predecir la exclusión educativa con un semestre de antelación (asistencia hasta julio) se obtiene *recall* 81.5% y *precision* 19.29% respecto a la clase excluida, y *accuracy* 89.215% del modelo general, para el conjunto de test con balanceo natural, con el meta-algoritmo XGBoost. También se proponen 9 perfiles de fenómeno de exclusión educativa, calculados con el modelo de agrupación Gaussian Mixture Model, y descritos cualitativamente en vínculo interdisciplinario con las ciencias sociales. Se utiliza la metodología CRISP-DM, y Python 3.6 para la programación.

"abramos todas las jaulas
pa' que vuelen como pájaros",

Víctor Jara.

AGRADECIMIENTOS

A mi mamá, Paola, mi tata, Oscar, y mi papá, Orlando. Con sus ejemplos me mostraron que la educación e ingeniería significan cimientos y herramientas de cambio.

A Canela. Por nuestro verano descalzo y multicolor. Tu compañerismo, enseñanzas, pasión y sonrisas le dieron la tinta a cada una de estas palabras. Y tanto más, infinitas gracias.

A Claudia. Por todo el apoyo y cariño.

A Constanza y Sergei. Por recorrer juntos estos extraños caminos.

Al Laboratorio de Gobierno, al Centro de Investigación Avanzada en Educación, y al Instituto Sistemas Complejos de Ingeniería. Por jugárselas en crear espacios que cambien futuros.

A mi familia, amigues y compañeres. Gracias a todes.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1.- Introducción

El presente capítulo tiene como finalidad entregar al lector el contexto en que se enmarca el trabajo de memoria.

En primer lugar, en el presente trabajo, el fenómeno en que les estudiantes no continúen su trayectoria en el sistema educativo, no presentando matricula escolar en tanto debiera, recibe dos denominaciones: "exclusión educativa" y "deserción escolar".

Esto se debe a que, de manera sostenida en la literatura y a nivel institucional, se ha utilizado el concepto "deserción escolar" para identificar el fenómeno, lo que etiqueta automáticamente a les estudiantes que lo viven como "desertores". Esto significa una diferencia con el enfoque de la inclusión educativa, ya que la palabra "desertar" atribuye la responsabilidad de abandonar a les estudiantes, lo que no sería el caso, ya que son múltiples factores, en gran parte de carácter socioestructural y ajenos a su poder de cambio, los que les empujan y condicionan para que se produzca la salida del sistema educativo: por eso el autor prefiere referirse al fenómeno como "exclusión educativa", ya que es el sistema quien finalmente excluye a les niñes y jóvenes, y no estos quienes libremente, y a conciencia, deciden abandonarlo.

A modo de respetar y exponer los enfoques de cada institución y/o autor, a lo largo del trabajo de memoria se presenta la terminología original de cada referencia según corresponda, pero el sentido a relevar por el autor es el de "exclusión educativa". Este tema es retomado también en el capítulo de Marco Conceptual.

1.1.- Motivación

A continuación, se describe tanto la motivación del autor, los objetivos del trabajo y los antecedentes generales.

1.1.1.- Prevención de la exclusión educativa

El presente trabajo nace de la importancia que le atribuye el autor a la problemática social que significa la exclusión educativa de les niñes y jóvenes del sistema escolar. Ya que se trata de un fenómeno de suma relevancia a la hora de proyectar sus perspectivas de vida, sociales, culturales y laborales. Repercutiendo tanto social, familiar e individualmente.

El dejar la escuela o liceo, para niñes y jóvenes, muchas veces en situación de vulnerabilidad, significa romper con la última cobertura institucional nativa y comunitaria, de contacto cotidiano y efectivo, abocada a satisfacer su derecho a un desarrollo integral a lo largo de su proceso de formación.

Por esto se propuso desarrollar un algoritmo predictivo de exclusión educativa, como propuesta de herramienta de política pública. Que permita una alerta temprana y mejor caracterización del fenómeno. Con el fin de aportar a una mejor y oportuna intervención preventiva por parte de las comunidades educativas e instituciones interventoras. Así como buscar y ofrecer nuevas formas de comprender el fenómeno de la exclusión educativa a través de la visualización de datos, con el fin de aportar a un mejor escenario para la innovación pública y diseño de políticas públicas.

1.1.2.- Ciencia de datos para las políticas públicas

En la última década, de mano del advenimiento de la digitalización e hiperconexión, se han creado, almacenado y procesado inmensas cantidades de datos, virtualmente en todos los aspectos de la experiencia humana. Este fenómeno ha sido difundido fuertemente bajo el concepto de "Big Data". Simplemente para graficar, durante los últimos años se han creado más datos que en toda la historia de la humanidad. Este cambio de orden de magnitud, de los datos al alcance del ser humano, no puede ser pasado por alto, ya que, comprendiendo sus virtudes, cuestiones éticas, y limitaciones, posiciona a los datos como materia prima de una nueva generación de herramientas para la comprensión y transformación de las relaciones y realidades dentro de nuestras sociedades.

Junto a los avances de la capacidad de procesamiento y almacenamiento de datos (*hardware*, *cloud computing*), y desarrollo de arquitecturas y librerías

(*software*) para su procesamiento; la programación científica, la estadística y matemáticas; han convergido con la gran diversidad de ciencias en un nuevo paradigma científico: la Ciencia de Datos. Que básicamente se enfoca en el uso de colecciones masivas de datos, para enfoques retrospectivos, predictivos o prescriptivos, con el fin de generar nuevo conocimiento, y mejorar la toma de decisiones.

Este nuevo paradigma se ha iniciado y potenciado fuertemente en el ámbito privado y académico, pero de manera rezagada en el mundo de las políticas públicas. No obstante, gracias al posicionamiento de la Innovación Pública, los desafíos de modernización del Estado, y el enfoque de Estado Abierto; de a poco se ha logrado que la Ciencia de Datos para las Políticas Públicas se transforme en un nuevo horizonte a explorar por parte del Estado.

1.1.3.- Lenguaje inclusivo de género

El autor tiene la convicción de que el machismo, y la discriminación a las distintas identidades de género, en nuestra sociedad se pueden contrarrestar en parte creando, practicando e impulsando nuevas culturas, intentando modificar la imperante. Es por este sentido que se decide reemplazar el género aludido por pronombres personales masculinos y femeninos, referentes a personas, por neutros, al igual que terminaciones que quepan dentro del criterio. Mayoritariamente se verá expresado en el uso de “les”; al igual que en los sustantivos propios de personas, ejemplo: “les niñes”.

Si bien esto pudiera significar incomodidad inicial al lector, también es parte del objetivo que busca el autor: invitar al lector a salir de su zona de confort, experimentando otras formas del lenguaje, en pos de construir o soñar nuevas culturas que abracen la diversidad.

1.2.- Objetivos del trabajo

A continuación, se presenta el objetivo general del trabajo de memoria, seguido por los objetivos específicos.

1.2.1.- Objetivo general

Predecir de forma individual y temprana el potencial de exclusión educativa de les niñes y jóvenes estudiantes del sistema regular de enseñanza básica y media en Chile, utilizando el paradigma de la Ciencia de Datos, para facilitar la intervención y prevención oportuna por parte de las instituciones y comunidades incumbentes.

1.2.2.- Objetivos específicos

1. Generar predicciones probabilísticas para construir y asignar indicadores del potencial de exclusión educativa para cada estudiante.

2. Construir y asignar perfiles de fenómenos de potencial exclusión educativa del sistema escolar, para facilitar la eventual intervención preventiva al abordar de manera contextualizada el fenómeno particular que caracteriza al estudiante evaluado.

3. Generar visualizaciones de datos que ofrezcan mayor entendimiento del fenómeno de exclusión educativa del sistema escolar, con el fin de aportar a un mejor escenario para la innovación, diseño e implementación de políticas públicas.

1.3.- Antecedentes generales

El trabajo de memoria se ha desarrollado en vínculo, en distintos ámbitos, con instituciones relevantes para el ámbito de este: el Laboratorio de Gobierno, institución estatal mandatada al impulso de la Innovación Pública, participa en el marco de la exploración e incorporación de nuevas herramientas y metodologías que permitan crear mejores servicios y políticas públicas centradas en las personas, a través de su respaldo institucional y apoyo metodológico para la interpretación de los perfiles de exclusión educativa; el Centro de Investigación Avanzada en Educación (CIAE) de la Universidad de Chile se vincula en el marco del Proyecto de investigación: "Un sistema nacional de protección de trayectorias educativas: disminuyendo la exclusión educativa en la enseñanza escolar y previniendo la deserción en educación superior" (IT17i0006), FONDEF abocado a la creación del **Sistema Nacional de Protección de las Trayectorias Educativas (SNPTE)** que tiene el fin de poner a disposición de los Servicios Locales de Educación tanto la predicción de la exclusión educativa escolar como la sistematización de experiencias exitosas de retención. Su participación se expresa a través del metodológico en el procesamiento y análisis de datos, la facilitación de bases de datos relevantes para el problema, y su respaldo institucional; y, por último, el Instituto de Sistemas Complejos de Ingeniería, a través de su patrocinio en el marco del proyecto de investigación "Learning Analytics" (CONICYT – PIA – FB0816).

En lo que resta del presente capítulo se retratan los antecedentes generales que enmarcan el trabajo de memoria: experiencias similares previas atingente; la caracterización del *sistema educacional escolar regular chileno*; las caracterizaciones del Laboratorio de Gobierno y el Centro de Investigación Avanzada en Educación de la Universidad de Chile (CIAE); y el contexto y marco institucional de la exclusión educativa y la ciencia de datos en Chile.

1.3.1.- Experiencias previas o similares

A continuación, se abordan experiencias previas y/o similares que tratan la problemática de la exclusión educativa en Chile desde un enfoque de detección temprana para gatillar mecanismos de prevención:

- Sistema de Alerta Temprana: "Presente", Municipalidad de Peñalolén

Implementado a partir del 2010, el Sistema de Alerta Temprana de deserción escolar "Presente" de Peñalolén, es un programa que busca prevenir la deserción escolar al detectar comportamientos de ausentismo escolar y gatillas intervenciones profesionales en les estudiantes. Consta de un sistema informático al que se reporta la asistencia, y si esta es reiterada se activan equipos profesionales de apoyo para atender y acompañar individualmente a les niñes y sus familias. Así se busca superar esta situación y regularizar la asistencia. También cuenta con una línea de acción comunitaria, que con difusión y talleres busca prevenir la deserción al involucrar a las comunidades. Este sistema no cuenta con mecanismos o herramientas de aprendizaje de máquinas.

- "Aquí, Presente", SEREMI Metropolitana de Educación

Este programa comienza a ser implementado desde el 2015 en colegios de la Región Metropolitana. Se considera una suerte de escalamiento del SAT "Presente" de Peñalolén, ya que surge de una alianza entre el Ministerio de Educación y el Gobierno Regional Metropolitano bajo el mandato del Intendente Claudio Orrego, quien fuera Alcalde de la comuna de Peñalolén en la gestión del surgimiento del SAT "Presente". Cabe reiterar que estos sistemas no son compuestos por algoritmos de aprendizaje de máquinas, pero sí dicha experiencia da luces de la efectividad e importancia de la intervención temprana para la prevención de la deserción al aumentar la asistencia en un 11,7% de los estudiantes intervenidos (Vega & Grau, 2016).

- Memoria de título, UAI 2015

La memoria de título "Desarrollo de un sistema prototipo para la detección temprana de la deserción escolar en escuelas públicas chilenas" realizada el 2015 por Camila Escobar y Felipe Lolas, para optar al título de Ingeniería Civil Industrial de la Universidad Adolfo Ibáñez, es un trabajo en que sí se utiliza el paradigma de la Ciencia de Datos, y en específico el Aprendizaje de Máquinas, para abordar la problemática de la deserción escolar. En dicha experiencia se contó con una muestra de 7.500 excluides y 250.000 no- excluides. Donde el algoritmo AdaBoost resultó ser el más efectivo, con un poder de predicción generalizado de exhaustividad (*recall*) del 88% respecto a los casos de deserción. Este trabajo dependía de fuentes no transversales de información, lo que explica la baja cantidad de muestras (10% de la muestra anual que utiliza el presente trabajo de memoria); también el modelo utilizaba toda la información de un año hacia el otro, sin posicionarse dentro del año escolar. También la evaluación del modelo fue hecha con los datos balanceados (50%

excluides - 50% no excluides), es decir no en situación de distribución real del fenómeno (3% excluides - 97% no excluides), lo cual no permite evaluar su desempeño en condiciones de vida real, ni evaluar su poder de precisión *(precision)*. Las principales diferencias con el presente trabajo de memoria, es que en este se construyen modelos basados sólo en fuentes de datos sostenidas en el tiempo, y transversales para todes les estudiantes, y validados en una muestra de datos con distribución real del fenómeno; también evalúa el potencial de exclusión educativa dentro del año, no sólo a su fin; y construye y asigna tipologías de fenómenos de exclusión educativa para cada caso.

1.3.2.- Sistema escolar regular de niñes y jóvenes

El sistema escolar regular de niñes y jóvenes se compone por los niveles de enseñanza básica, y media científico humanista y técnico profesional y artístico. A modo de generar una mayor comprensión de sus magnitudes, a continuación, se presentan datos que describen el periodo 2017 (MINEDUC, 2018):

- Matrícula 2017 (Total: 2,858,969 estudiantes):
    - Municipal: 1,064,187 (37.2%)
    - Particular subvencionado: 1,504,738 (52.6%)
    - Particular pagada: 244,441 (8.5%)
    - Corporación de administración delegada: 45,603 (1.6%)

Siendo el total de la matrícula escolar global, sin diferenciar la educación escolar especial y adultes, de la de niñes y jóvenes, es de 3,024,627 estudiantes. Por lo que la matrícula del sistema escolar regular representa un 94,5% del sistema escolar total.

- Establecimientos 2017:
    - Básica: 8,334
    - Media científico humanista: 2,612
    - Media técnico profesional: 949

Cabe destacar que los establecimientos que tienen básica y media fueron contabilizados cada una de las veces que correspondiera por separado.

- Reforma del sistema escolar

La evolución del sistema educacional chileno, desde los años 80', de manos de la dictadura cívico-militar, y por razones ideológicas ligadas a la concepción neoliberal, se ve parte de un plan mayor orientado al desmantelamiento de la influencia del Estado y de las autoridades políticas (Baeza y Fuentes 2004). Es desde ese momento, y de manos de la municipalización, entre otras medidas, y la posterior profundización de dichas lógicas por parte de la Concertación, es que el sistema educacional sufre un crudo proceso de profundización en su segregación y precarización. Dicho proceso finalmente cataliza una seguidilla de estallidos sociales, como el mochilazo del 2001, la revolución pingüina del 2006, y el mundialmente conocido movimiento estudiantil del 2011. Estos provocan un cambio histórico en las correlaciones de las fuerzas políticas imperantes, respecto a la concepción del modelo educativo y de derechos sociales, durante décadas en nuestro país. Generando así un ineludible proceso de reformas al sistema educacional en su conjunto.

A continuación, se describen los principales elementos del proceso de reformas ligadas al nivel escolar de la última década.

- Sistema Nacional de Aseguramiento de la Calidad de la Educación Escolar (SAC)

El año 2011, enmarcado en la mayor movilización social desde la vuelta a la democracia, se promulga, el 11 de agosto de 2011, la Ley Nº 20.529. Se crea así el Sistema Nacional de Aseguramiento de la Calidad de la Educación Escolar. Transformándose en la nueva Institucionalidad que coordina y redefine la interacción y desenvolvimiento de las instituciones a cargo de las políticas educativas. Este está compuesto por los siguientes elementos institucionales.

- *Ministerio de Educación (MINEDUC)*

El MINEDUC es el órgano rector del Sistema Nacional de Aseguramiento de la Calidad de la Educación Escolar. Está encargado de diseñar e implementar las políticas educacionales para todo el sistema educativo. Al igual que del diseño de instrumentos curriculares y de evaluación.

- *Superintendencia de Educación*

La Superintendencia de Educación es el órgano fiscalizador del sistema: Fiscaliza el cumplimiento de la normativa educacional y aplica sanciones; Fiscaliza la legalidad del uso de recursos y audita la rendición de cuentas; Investiga y resuelve denuncias; Y canaliza reclamos

.

- *Consejo Nacional de Educación*

El Consejo Nacional de Educación tiene por objetivo el aprobar e informar sobre: las Bases curriculares; Planes y programas de estudio; Estándares de Aprendizaje y Otros Indicadores de Calidad; Estándares Indicativos de Desempeño; Plan de evaluaciones nacionales e internacionales.

- *Agencia de la Calidad de la Educación*

La Agencia de la Calidad de la Educación tiene la labor de evaluar y orientar el desempeño de los establecimientos educacionales y sostenedores. Así como el evaluar los logros de aprendizaje de les estudiantes, y entregar información a las comunidades y centros educativos acerca su gestión, con el fin de apoyar la mejora progresiva del sistema.

Ilustración 1: esquema del Sistema Nacional de Aseguramiento de la Calidad Educacional Escolar, MINEDUC (2017)

- Ley de Inclusión Escolar

El 29 de mayo de 2015 se promulgó la Ley Nº 20.845 de "Inclusión Escolar que regula la admisión de los y las estudiantes, elimina el financiamiento compartido y prohíbe el lucro en establecimientos educacionales que reciben aportes del Estado", una de las leyes más importantes a la hora de igualar oportunidades educativas e ir contra la segregación instalada por décadas en el sistema educativo. Sus principales componentes son:

- *Gratuidad de la educación*

La ley de inclusión consagra la educación gratuita como un derecho social, por lo que elimina el copago de las familias gradualmente en los colegios, ya que identifica en éste un elemento de segregación.

- *Fin al lucro con recursos públicos*

La ley pone fin al lucro en la educación escolar con recursos público. Lo que significa que los establecimientos educacionales particulares subvencionados deben decidir si convertirse en entidades sin fines de lucro, o pasar a ser particulares pagados y por ende no afectos al financiamiento estatal.

- *Fin a la selección*

También pone fin a la selección en los colegios, igualando las oportunidades educativas, yendo en contra de la segregación y reproducción de desigualdades que este mecanismo generaba.

- Ley de Nueva Educación Pública

Promulgada el 16 de noviembre del 2017, la Ley de Nueva Educación Pública es una de las reformas más relevantes de las últimas décadas respecto a la institucionalidad educacional escolar. Con el objetivo de fortalecer la calidad, y revertir los tantos efectos negativos productos de la precarización de la educación pública, llevada a cabo por la dictadura cívico-militar en los años 80' y perpetuada por el proceso de transición a la democracia.

La Nueva Educación Pública crea una nueva institucionalidad educativa:

- *Dirección de Educación Pública*

La Dirección de Educación Pública, dependiente del MINEDUC, es un servicio público especializado que está a cargo de conducir estratégicamente, y coordinar al sistema. Debe velar por la calidad de la educación provista por los Servicios Locales en el territorio.

Debe impulsar la Estrategia Nacional de Educación Pública a 8 años, de construcción participativa, que es aprobada por el Consejo Nacional de Educación. También apoya administrativamente a los Servicios Locales de Educación.

- *Servicio local de Educación*

La nueva institucionalidad crea 70 Servicios locales de Educación, descentralizados en funciones, y distribuidos por todo el territorio nacional. Integran lo técnico-pedagógico y lo administrativo-financiero del sistema, teniendo a su haber los jardines, escuelas y liceos públicos que les corresponden. Están encabezados por su Dirección Ejecutiva, y son los responsables de la provisión y gestión educativa en su territorio.

- *Consejo local y Comité directivo*

La gobernanza del sistema, en su nivel de Servicios locales de Educación, incluye un Consejo Local, de perfil educativo, y un Comité Directivo, que se dedica a la rendición de cuentas, y aprueba el Plan Estratégico a 6 años, y propondrá al Presidente de la República la terna de candidatos elegibles para la Dirección del Servicio Local de Educación respectivo, luego del proceso de Alta Dirección Pública.

- *Establecimientos educacionales públicos*

Lo liceos, escuelas, jardines Vía Transferencia de Fondos y salas cuna públicos representan la unidad fundamental y celular del nuevo sistema. Están integrados a la red de cada Servicio Local de Educación. Y tienen como objetivo el asegurar el aprendizaje y la formación integral de les estudiantes.

Ilustración 2: esquea organizacional de la Nueva educación Pública, MINEDUC (2017).

1.3.3.- Contexto y marco institucional

1.3.3.1- Exclusión educativa y deserción escolar

La exclusión educativa y deserción escolar del sistema chileno históricamente se ha tratado mayormente bajo la política de focalización, mayormente ligada a becas, subvenciones e intervenciones específicas, las más relevantes.

- Políticas públicas relevantes

A continuación, se retratan las políticas públicas que actualmente se abocan a la prevención de este fenómeno.

- Subvención Educacional Pro-Retención

Esta subvención se entrega a los sostenedores de establecimientos municipales y particulares subvencionados que acrediten haber retenido y matriculado a estudiantes que cursan entre séptimo y cuarto medio, o egresen de este nivel, que pertenezcan al Programa Chile Solidario. Esto de acuerdo con los resultados en su Calificación Socioeconómica (CSE), determinada por el Ministerio de Desarrollo Social. La subvención varía entre $90,000 y $215,000 aproximadamente (Ley N° 19.873, 2003). Esta subvención cubre cerca de 220,000 estudiantes anualmente.

- Programa de Apoyo a la Retención Escolar (PARE)

El PARE es un programa de apoyo sicosocial enfocado en entregar herramientas que generen factores protectores en les estudiantes que presentan mayor riesgo socioeducativo y/o en condición de embarazo, maternidad o paternidad, con el objetivo de prevenir la deserción o abandono escolar. El programa se instala en (JUNAEB, 2018a):

- Comunas y Establecimientos con alta concentración de estudiantes en condición de embarazo, maternidad o paternidad y alta concentración de estudiantes con riesgo socioeducativo.
- Establecimientos Educacionales priorizados por vulnerabilidad socioeducativa que cuenten con la Beca de Apoyo a la Retención Escolar BARE.

- Beca de Apoyo a la Retención Escolar (BARE)

La BARE es una beca que otorga la JUNAEB a les niñes y jóvenes que presentan alto riesgo de abandonar el sistema educacional. Esto se identifica al presentar al menos una o más de las siguientes situaciones (JUNAEB, 2018b):

- Presentar factores de riesgo de deserción: Sobre edad respecto al nivel cursado, baja asistencia el año inmediatamente anterior en los registros del Mineduc, condición de paternidad o maternidad, condición de embarazo o padres de hijos en gestación.
- Estar registrado en un programa de protección del Sename.
- Ser beneficiario del Programa de Apoyo a la Retención Escolar (PARE) del Departamento de Salud del Estudiante de Junaeb.
- Presentar una discapacidad acreditada en el Registro Social de Hogares.
- Pertenecer a una de las familias registradas en el Subsistema Seguridades y Oportunidades, ex Chile Solidario o Ingreso Ético Familiar.
- Estar incorporado en el primer tramo del Registro Social de Hogares (0-40%).
- Presentar otros factores de riesgo de deserción acreditados mediante informe socioeducativo emitido por el encargado de la red colaboradora u otro profesional del área social.

Consiste en un aporte de $196,000 anuales, que se distribuyen en cuatro cuotas de los siguientes montos a pagar durante el año; $40,000 (abril), $45,000 (julio), $50,000 (septiembre) y $61,600 (noviembre). Las cuotas se pagan en los primeros cinco días del mes, a través de BancoEstado. Esta beca cubre a aproximadamente a 25,000 estudiantes anualmente.

1.3.3.2.- Ciencia de datos y políticas públicas

La búsqueda de solución a problemáticas públicas por medio del paradigma de la ciencia de datos es un campo naciente a nivel global, y aún poco explorado en Chile.

A continuación, se describe el contexto y marco institucional en cual se encuentra lo referido a la ciencia de datos y políticas públicas.

1.3.3.2.1.- Actores nacionales relevantes

Entre los actores que destacan dentro del ecosistema, por estar desarrollando o aproximándose al uso del paradigma de la ciencia de datos para abordar problemas públicos, se encuentran:

- Ministerio de Hacienda
- Programa de Modernización del Estado
- SEGPRES
- Centro de Investigación Avanzada en Educación (CIAE) de la Universidad de Chile.
- Servicio de Impuestos Internos.
- Dirección del Trabajo.
- Laboratorio de Gobierno de la Universidad Adolfo Ibáñez.

1.3.3.2.2.- Regulaciones relevantes

Respecto a la regulación relevante a la hora de trabajar y analizar datos masivos, existen tres relevantes:

- *Ley de protección de la vida privada*

Ley N°19.628, de protección a la vida privada, es la principal legislación que rige el tratamiento de los datos privados de las personas ("LEY-19628 28-AGO-1999 MINISTERIO SECRETARÍA GENERAL DE LA PRESIDENCIA - Ley Chile - Biblioteca del Congreso Nacional", 2012). Esta legislación data de 1999, sufriendo modificación el 2012, y el autor considera que presenta altos grados de obsolescencia al concebirse en un contexto ajeno al actual. Actualmente se encuentra en trámite su modificación, tanto para una mayor protección de datos sensibles y la privacidad de las personas, y también para permitir mayor usabilidad en aspectos económicos y de políticas públicas (Tele13, 2017).

- *Interoperabilidad de datos*

La interoperabilidad de datos es el intercambio de información entre organismos públicos que lo requieran debido a la naturaleza de sus actividades. Esto requiere de un acuerdo de condiciones entre las instituciones incumbentes mediante un Convenio de Intercambio de Información.

Los intercambios de información pueden ser llevado a cabo mediante tres mecanismos (CETIUC, 2017):

- Transferencia de archivos de forma manual o semi automática.
- Interoperabilidad mediante la Plataforma de Interoperabilidad del Estado (PISEE).
- Interoperabilidad directa entre instituciones.

Existe un consenso de lo demoroso que resultan los convenios, y que debiera existir una normativa única para esto. Así también se identifica que la diferencia a nivel tecnológico y de falta prioridad por parte de los mandos directivos de las instituciones son aspectos que afectan el desenvolvimiento de este mecanismo (CETIUC, 2017).

Actualmente la Unidad de Gobierno Digital del Ministerio Secretaría General de la Presidencia (SEGPRES), junto a un comité interinstitucional, ha desarrollado una propuesta de nuevo modelo de normativa para la interoperabilidad de las instituciones del Estado, el cual fue presentado a consulta pública entre el 1 y 15 de septiembre del 2017. Dicha propuesta de norma de interoperabilidad considera los aspectos destacados en el párrafo anterior (Ministerio Secretaría General de la Presidencia, 2017).

- *Ley sobre acceso a información pública*

La ley 20.285, sobre acceso a información pública es la normativa que provee mecanismos para que la sociedad civil, y el público en general, solicite y acceda a información contenida en actos, resoluciones, actas, expedientes, contratos y acuerdos, así como toda la información elaborada con presupuesto público ("LEY-20285 20-AGO-2008 MINISTERIO SECRETARÍA GENERAL DE LA PRESIDENCIA - Ley Chile - Biblioteca del Congreso Nacional", 2016).

1.3.3.2.3.- Tendencia

El creciente uso del análisis masivo de datos, en todo ámbito, es un fenómeno global. Constituyéndose interesantes experiencias relevantes su de aplicación a las políticas públicas, como el laboratorio de innovación tecnológica del Gobierno de Singapur (GovTech Singapur), el Laboratorio de Gobierno de la Universidad de Nueva York (GovLab NYU), el Laboratorio de Datos del Laboratorio para la Ciudad de la Ciudad de México, y el Centro de Ciencia de Datos y Políticas Públicas de la Universidad de Chicago, son ejemplos en que los laboratorios de innovación pública a nivel mundial están incorporando el paradigma de la ciencia de datos, para las políticas y servicios públicos, en sus campos de acción. Donde se presentan casos de desarrollo relevantes como:

optimización de recursos de servicios de emergencia, identificación de nivel de stress de funcionarios policiales, identificación de fraudes financieros, ciudades inteligentes, entre otros.

1.3.3.2.4.- Posicionamiento

A continuación, se analiza el posicionamiento desde el cual se plantea el trabajo de memoria, a través de las instituciones con que se vincula. Elemento considerado relevante, para efectos de la permeabilidad en la realidad chilena, del paradigma de la Ciencia de Datos para las Políticas Públicas, y en particular para abordar la exclusión educativa.

El Laboratorio de Gobierno ha logrado posicionarse fuertemente como una institución pública que de forma efectiva y eficiente impulsa una nueva cultura de Estado Innovador. Se ha enfocado en la transferencia de capacidades de innovación a funcionarios, en la innovación de servicios públicos y el involucrando nuevos actores de fuera del Estado para encontrar soluciones a problemáticas públicas. Ha logrado movilizar a miles de personas, entablar fuertes relaciones y dinamizar al ecosistema nacional e internacional de innovación. Esto lo deja en una posición de validación institucional ante la exploración de este nuevo campo.

El Centro de Investigación Avanzada en Educación, cuenta con una importante experiencia en el enfoque del uso de nuevas tecnologías para aportar al entendimiento, análisis y desarrollo de políticas públicas educativas. Proyectos destacados son: "Modelamiento y optimización de recursos educacionales MORE" (CORFO); "Enseñanza y evaluación de la escritura mediante una plataforma tecnológica colaborativa" (FONDEF IT 15I10002); "The geography of education: a territorial intelligence platform to support the implementation and management of new public policies in education" (FONDEF IT 15I10010). Es de suma relevancia recalcar la participación del presente trabajo en el proyecto de investigación "Un sistema nacional de protección de trayectorias educativas: disminuyendo la exclusión educativa en la enseñanza escolar y previniendo la deserción en educación superior" (FONDEF IT17i0006), que constituye el hogar natural del presente trabajo de memoria, ya que está abocado a la creación del **Sistema Nacional de Protección de las Trayectorias Educativas (SNPTE)** que tiene la finalidad de disponer para los Servicios Locales de Educación tanto la predicción de la exclusión educativa del sistema escolar como la sistematización de experiencias exitosas de retención.

1.3.4.- Descripción organizacional

A continuación, se describen los aspectos más relevantes de las organizaciones con que se realiza el trabajo de memoria, con el objetivo de caracterizar de mejor manera su labor para con la innovación, ciencia de datos y políticas públicas.

1.3.4.1.- Laboratorio de Gobierno

El Laboratorio de Gobierno es un ente Público, constituido como "Comité de Innovación en el Sector Pública" mediante el acuerdo N°2.826, de 2014, del Consejo de CORFO. Está liderado por un Directorio multisectorial compuesto por representantes del Ministerio de Economía, Fomento y Turismo; Ministerio Secretaría General de la Presidencia; Ministerio del Interior y Seguridad Pública; Ministerio de Desarrollo Social; Ministerio de Hacienda; CORFO; Servicio Civil; y expertos y académicos de la Sociedad Civil. Su proceso de compra de bienes y contratación de servicios se produce conforme a la Ley N°19.886, a través del mecanismo de compra del sector público Chile Compra (Corporación de Fomento de la Producción, 2014). El Laboratorio de Gobierno ejecutó durante el 2016 un presupuesto de 2.804.841 mil pesos chileno (Ministerio de Economía y Turismo, 2017). Las líneas de acción que desarrolla son la transferencias de capacidades para innovar de funcionarios públicos a través del programa Experimenta, creando una red de más de 2200 innovadores público (Red de Innovadores Públicos, octubre 2017); la incorporación de nuevas ideas desde fuera del Estado para abordar problemas públicas a través de los Concursos de innovación abierta AULAB e IMPACTA, financiando, coordinando y entregando tutela metodológica de procesos de diseño para la innovación pública; y apoyando y participando del fortalecimiento del ecosistema nacional a través del programa de apoyos institucionales (PAI), por el cual diversas iniciativas de innovación pública se ven beneficiadas del apoyo y respaldo del Laboratorio.

- Misión

Impulsar, apoyar y fomentar la innovación en el sector público, dando soluciones a las problemáticas y desafíos de sus distintos niveles y ámbitos de trabajo, de modo que éstas generen valor público en el entorno y la sociedad.

- Objetivo general

Promover, coordinar y fomentar la gestión de la innovación en el sector público, dando soluciones a las problemáticas y desafíos de sus distintos niveles y ámbitos de trabajo, de modo que éstas generen valor público incidente en la productividad del país, dentro de la esfera de competencia de cada uno de los órganos.

- Estructura organizacional

El Laboratorio de Gobierno contó en su primer periodo 2014-2018 con una estructura organizacional de tres niveles. En la actualidad este modelo se encuentra en proceso de iteración, propio del desarrollo de un nuevo espacio organizacional, donde sólo se conserva los componentes del primer y segundo nivel (Directorio y Dirección Ejecutiva). No obstante, se destaca la efectividad y capacidad de adecuación de la estructura inicial para el cumplimiento de sus objetivos, y ponerlo en conocimiento de lector resulta de bastante utilidad para hacerse una idea de las dinámicas organizacionales del Laboratorio. Por lo que se describe a continuación.

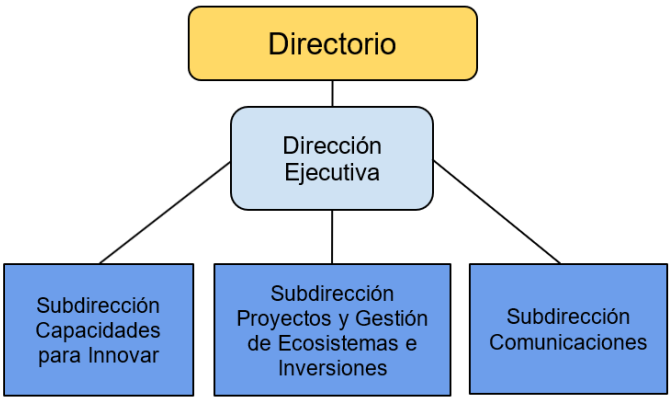

Ilustración 3: esquema organizacional Laboratorio de Gobierno, elaboración propia (2018)

En su nivel superior se encuentra el Directorio, seguido por la Dirección Ejecutiva, que tiene a su haber a tres subdirecciones que desarrollan las líneas de trabajo del Laboratorio. La Subdirección de Capacidades para Innovar; Proyectos y Gestión de Ecosistemas e Inversiones; Comunicaciones y producción.

- *Directorio (Consejo Estratégico)*

El Directorio o Consejo Estratégico del Laboratorio de Gobierno está compuesto de manera intersectorial por representantes del Ministerio de

Economía, Fomento y Turismo; Ministerio Secretaría General de la Presidencia; Ministerio del Interior y Seguridad Pública; Ministerio de Desarrollo Social; Ministerio de Hacienda; CORFO; Servicio Civil; y expertos y académicos de la Sociedad Civil.

El Directorio tiene la misión de darle lineamientos estratégicos al Laboratorio de Gobierno y aprobar sus metas y objetivos. Asimismo, producto de su composición transversal e interministerial, le permite dar foco estratégico a los distintos programas e intervenciones que el Laboratorio de Gobierno realice. También selecciona al Director Ejecutivo, de una terna presentada por el Vicepresidente Ejecutivo de CORFO, previo concurso público. ("El Lab", s. f.)

- *Dirección Ejecutiva*

La Dirección Ejecutiva encabeza las relaciones institucionales y la administración del Laboratorio Gobierno; responde directamente al Directorio (que es quien lo elige); y dirige la plana de subdirectores. Resalta el profundo seguimiento y supervisión del Director respecto a cada uno de los proyectos de Laboratorio, siendo una pieza clave en el dinamismo y estándares profesionales.

- *Subdirección de Proyectos y Gestión de Ecosistemas e Inversiones (PROGEI)*

La Subdirección de Proyectos y Gestión de Ecosistemas e Inversiones (PROGEI) está a cargo de gestionar y desarrollar la cartera de proyectos del Laboratorio de Gobierno, compuesta por AULAB, IMPACTA y PROYECTOS. También de la inversión que se hace en los distintos actores que participan de estos. También está a cargo del Programa de Apoyos Institucionales (PAI), que da apoyo a iniciativas de impacto público e innovación. ("Gestión de Ecosistemas e Inversiones", s. f.)

- *Subdirección de Capacidades para Innovar (Capacidades)*

La Subdirección de Capacidades para Innovar tiene por objetivo el promover el desarrollo de habilidades, motivaciones y oportunidades de innovación pública en funcionarios del Estado que a través de experiencias del aprender

haciendo, puedan relevar su rol como fuente de innovación y agentes de cambio en el sector público. Está a cargo de EXPERIMENTA, ESTADO FUTURO, ESTUDIO OCDE y FUNCIONA!. ("Capacidades para Innovar", s. f.)

- *Subdirección de Comunicaciones (COM)*

La Subdirección de Comunicaciones se encarga de la documentación, difusión y articulación con los medios de comunicación. Así también como de la producción y puesta en escena de las distintas actividades a realizar en cada proyecto o programa. También está a cargo de toda la línea gráfica del Laboratorio de Gobierno y la producción de su material multimedia.

- Actividades claves del Laboratorio de Gobierno

A continuación, se describen las actividades claves del Laboratorio de Gobierno:

a. Concursos de innovación abierta: Tanto el AULAB como el IMPACTA, son concursos en que previa definición de desafíos con los servicios públicos correspondiente, se convoca abiertamente a la postulación de ideas que luego serán financiadas y aceleradas, para su posterior pilotaje y eventual escalamiento.

b. Talleres de transferencia de capacidades para innovar: Experimenta y la Red de Innovadores Públicos, son las líneas de trabajo con que se transfieren e incentivan las capacidades para innovar en los funcionarios e instituciones públicas. Su componente principal son los distintos talleres o dinámicas participativas, que a través del aprender haciendo, impulsan estos cambios al interior del Estado.

c. Utilización de metodologías no incorporadas previamente al diseño de servicios y políticas públicas en nuestro país: El estudiante considera que una de las innovaciones en el Estado más importantes que representa al Laboratorio de Gobierno, es la incorporación de nuevas concepciones del diseño de servicios y políticas públicas a la cultura organizacional del Estado (Ejemplo de esto es el Doble Diamante diseñado por el Design Council del Reino Unido).

d. Proyectos de innovación pública de desarrollo interno: Se considera muy relevante el que el equipo del Laboratorio, constituido como un colectivo experto en el diseño e innovación de servicios y políticas públicas, a través de la línea de Proyectos, se aboque al desarrollo interno de proyectos de innovación pública.

e. Convocatorias abiertas para ayudas: El Programa de Apoyo Institucional es la línea de convocatoria abierta encargada de apoyar iniciativas que promuevan la innovación, a través de ayuda en logístico, metodología, financiamiento y organización en general. Esta actividad permite otra forma de interacción entre el Laboratorio con el ecosistema, lo que el estudiante considera positivo.

f. Reconocimientos de iniciativas de innovación dentro del Estado: Funciona! juega un rol de reconocimiento a las iniciativas ya realizadas dentro del Estado, en un claro escenario de adversidad cultural organizacional. Aporta directamente a la dimensión relacional del Estado, donde sus funcionarios ven fortalecida y estimulada, a través del reconocimiento institucional, sus valiosas iniciativas innovadoras.

- Propuesta de valor del Laboratorio de Gobierno

  En consecuencia, con lo recién descrito, la propuesta de valor del Laboratorio de puede describir siguiente manera:

  - Alcanzar nuevas y mejores soluciones para problemáticas públicas, a través de metodologías de cocreación e innovación pública, en un espacio de experimentación para finalmente evaluar su escalabilidad.

  - Transferir capacidades de innovación a los funcionarios públicos para la construcción de una nueva cultura del qué hacer dentro del servicio público.

  - Apoyar y fortalecer el ecosistema nacional de innovación y emprendimiento.

1.3.4.2.- Centro de Investigación Avanzada en Educación CIAE

El Centro de Investigación Avanzada en Educación (CIAE) fue inaugurado a fines del 2008, siendo una iniciativa conjunta entre la Universidad de Chile, la Universidad de Concepción y la Pontificia Universidad Católica de Valparaíso ("CIAE - Universidad de Chile", 2018).

- Misión

Su misión es producir y diseminar conocimiento en el ámbito de la educación; dar soporte científico a la discusión y diseño de políticas públicas en el sector educación, para que las políticas nacionales, la gestión educativa local y la docencia en el aula estén basadas en la evidencia que genera la investigación; y contribuir a la formación de una nueva generación de jóvenes investigadores en el campo de la educación ("CIAE - Universidad de Chile", 2018).

- Áreas de investigación

El trabajo del CIAE se organiza en seis áreas temáticas. Realizando en cada una de ellas actividades de investigación fundamental y aplicada, y desarrollo y transferencia de sus resultados. Dentro de cada área de investigación los académicos del CIAE colaboran en la realización de proyectos conjuntos. Las áreas son ("CIAE - Universidad de Chile", 2018):

  - Políticas Educacionales
  - Profesión Docente
  - Enseñanza y Aprendizaje
  - Neurociencia y Aprendizaje
  - Tecnología de la información y educación
  - Educación inicial

Siendo el área de Tecnología de la información y educación donde se desarrolla el presente trabajo de memoria.

- Iniciativas

El CIAE, a parte de los proyectos de investigación que lleva a cabo, desarrolla iniciativas de impacto en las áreas de investigación e impacto. Estos se describen a continuación ("CIAE - Universidad de Chile", 2018).

- *ARPA*

Es una iniciativa de investigación y desarrollo entre el CIAE y el Centro de Modelamiento Matemático (CMM) de la Universidad de Chile. Busca implementar estrategias de desarrollo profesional docente que promueva la resolución de problemas matemáticos en el aula.

- Mejor Matemática

Es un programa de conjunta iniciativa entre el MINEDUC y la Universidad de Chile. Trabaja con escuelas públicas para la mejora continua de los procesos de enseñanza-aprendizaje de la matemática, fortaleciendo las capacidades docentes y las comunidades escolares.

- Mejoramiento Escolar

Es un área del CIAE que como objetivo tiene el comprender cómo las escuelas cambian para ofrecer a sus estudiantes mejor y más oportunidades de aprendizaje. Para ello se investiga y difunden los resultados de sus estudios a través de libros y artículos, conferencias, seminarios y talleres, y cursos.

- Observatorio Docente

Iniciativa del CIAE, es integrada por las Facultades de Educación de las Universidades del Consejo de Rectores (CRUCH). Busca ofrecer información actualizada sobre los programas de Formación Inicial Docente, permitiendo así conocer los aspectos sustantivos de los programas; y aspectos cuantitativos, como la variación de matrícula, egreso y titulación.

- Test de Aprendizaje y Desarrollo Infantil (TADI)

El TADI es una escala que permite evaluar de manera continua el desarrollo y aprendizaje de niñes entre 3 meses y 6 años de edad, diseñado y estandarizado en Chile. Evalúa dimensiones como la cognición, motricidad, lenguaje y socioemocionalidad.

2.- Justificación del proyecto

En el presente capítulo se expone las distintas aristas que componen la justificación del presente trabajo de memoria. Desde la realidad actual de la exclusión educativa en Chile, los efectos que este fenómeno significa, las alternativas de solución y la propuesta de valor.

2.1.- La exclusión educativa escolar en Chile

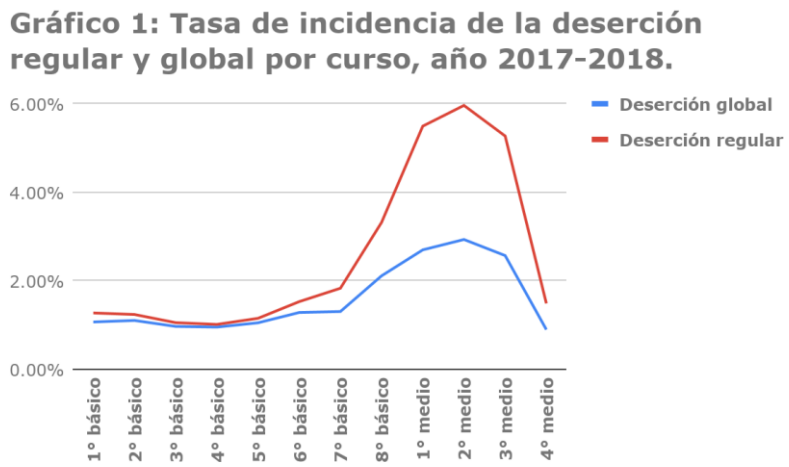

Ilustración 4: Tasa de incidencia de la deserción regular y global por curso 2017-2018, elaboración propia (2018) – Fuente datos: MINEDUC

La exclusión educativa escolar en Chile es un problema que acarrea efectos negativos a nivel social, como a nivel privado de les niñes, jóvenes y de sus familias.

En el período 2017-2018, es decir los estudiantes que no se matricularon el 2018 debiendo haberlo hecho, la tasa de incidencia de deserción global alcanzó un 1.81%, lo que equivale a 51,882 estudiantes, donde la matrícula del año 2017 del sistema regular fue de 2,859,064 estudiantes. A su vez, la tasa de incidencia del sistema regular alcanzó el 2.87%, equivalente a 82,092 estudiantes (elaboración propia desarrollada ampliada más adelante). Esto significa que el 63.2% de les estudiantes que desertan del sistema regular lo hacen de todo el sistema, y que 30,210 estudiantes optaron por migrar del sistema regular al sistema escolar para adultos.

Como se muestra en la *ilustración 4*, se identifica que la deserción se manifiesta de manera predominante en el tramo de enseñanza media. Siendo primero y tercero medio los momentos en que más ocurre.

La explicación predominante de este fenómeno en primero medio se refiere a que, muchos colegios que no cuentan con enseñanza media obligan a les estudiantes al cambio y adaptación a nuevas condiciones en nuevos establecimientos, donde se acentúan las condiciones que facilitan la exclusión

educativa. En tercero medio, el ingreso prematuro al mundo laboral se presenta como un factor importante (MINEDUC, 2013).

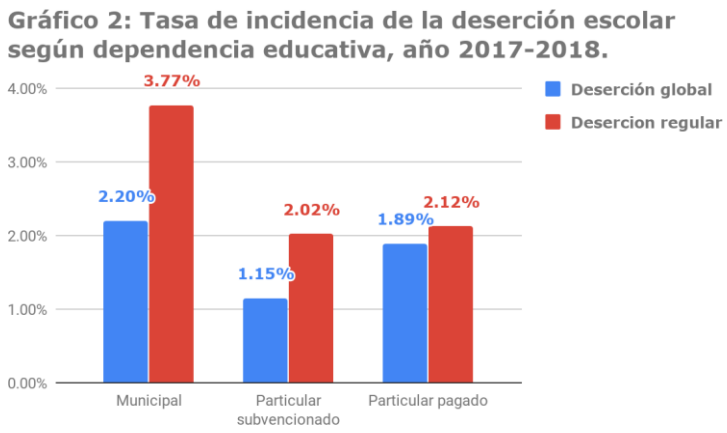

Ilustración 5: tasa de incidencia de la deserción escolar según dependencia educativa 2017-2018, elaboración propia (2018) – Fuente datos: MINEDUC

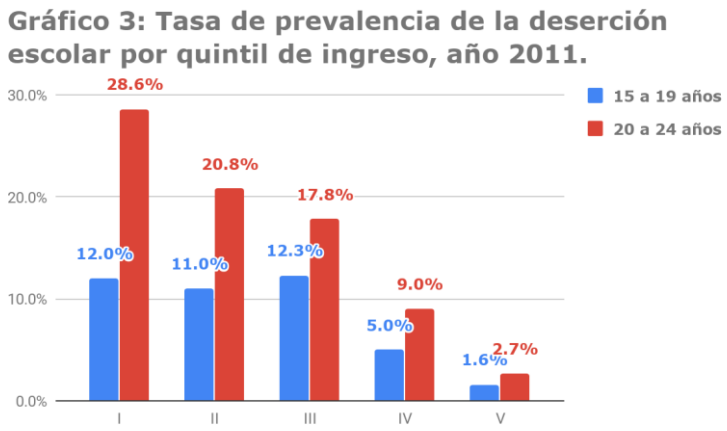

Ilustración 6: tasa de prevalencia de la deserción escolar por quintil de ingreso, año 2011, MINEDUC (2013) - Fuente datos: CASEN

También es importante señalar la diferencia evidenciada en los niveles de deserción escolar, dependiendo de la condición de dependencia administrativa de los establecimientos. Alcanzando en el sistema regular un 3.77% en los establecimientos municipales, y del orden el 2% en los establecimientos particulares subvencionados y pagados. Esto, junto con que la distribución por quintil de la tasa de prevalencia de deserción se ve drásticamente inclinada a los quintiles más vulnerables, es que se evidencia que la expresión del fenómeno de la deserción está fuertemente marcada por la procedencia de clase de las y los estudiantes, dadas por las mejores condiciones tanto extraescolares como intraescolares (MINEDUC, 2013).

2.1.1.- Efectos de la exclusión educativa

La exclusión educativa escolar trae consigo efectos personales, sociales y económicos en lo público y privado.

Se evidencia que este fenómeno impacta el nivel de capital humano de la fuerza de trabajo de los países, teniendo efecto en el crecimiento y nivel de productividad de las economías de los países (Barro, 1991).

En el caso chileno de les jóvenes infractores de ley, se pueden caracterizar con que han atravesado dos procesos que parecieran indisolubles entre sí: previo a su institucionalización en recintos de privación de libertad, se ha hecho presente la desescolarización y la exclusión social. La deserción escolar, junto con la repitencia, son las características fundamentales de la desescolarización (SENAME, 2013), y recordar, como evidencia de la trayectoria de la exclusión social en las personas, que el 47% de los reos en Chile pasaron por instituciones de SENAME (Fundación San Carlos de Maipo, 2017).

Lo anterior plantea una relación crítica entre la exclusión educativa escolar con problemas de cohesión y desigualdad social, ya que este fenómeno reproduce y profundiza las condiciones de vulnerabilidad de las y los jóvenes que la sufren, y de sus familias.

Es relevante mencionar el impacto que significa el no finalizar la enseñanza media para la vida adulta. La exclusión educativa escolar tiene como consecuencia impactos claros en el desenvolvimiento de las personas, que condicionan de forma negativa sus prospectivas de vidas y laborales, situándolas en una posición de clara desventaja respecto a quienes completan sus estudios. Este fenómeno se manifiesta en seis importantes dimensiones de la vida adulta (González, 2017):

1. Habilidades y competencias
2. Uso de habilidades y competencias en contextos laborales
3. Uso de habilidades y competencias en la vida cotidiana
4. Condiciones laborales
5. Oportunidades de educación continua
6. Bienestar subjetivo

2.2.- Hipótesis y alternativas de solución

La hipótesis principal del presente trabajo de memoria es que el fenómeno de la exclusión educativa escolar puede ser predicho de manera temprana con cierta precisión. Donde el criterio de "temprana" se define en otorgar margen de acción a las instituciones y/o comunidades cercanas al estudiante para el efectivo ejercicio de la prevención.

Se plantean dos alternativas de solución para un sistema de alerta temprana: sistemas basados en aprendizaje de máquinas, o sistemas basados en alerta arbitraria.

Ejemplo de sistemas de alerta arbitraria es la experiencia del Sistema de Alerta Temprana "Presente" de la Municipalidad de Peñalolén. Y su suerte de escalamiento, el programa "Aquí, Presente" implementado por la SEREMI Metropolitana de Educación. Que en su evaluación destaca el logro de un aumento del 11.7% en la asistencia de las y los estudiantes intervenidos (Vega & Grau, 2016). Aún así se evidenció una tardía aproximación de la intervención al estudiante, ya que las alertas se encendían cuando ya se inicia un proceso de abandono o intermitencia del proceso educativo. En esto es relevante el aspecto de "temprano", ya que es más difícil intervenir a un estudiante que aún no deja de asistir, que a uno que ya comenzó a dejar de asistir clases. Y también

Ejemplo de aproximación desde el aprendizaje de máquinas, para el caso de la exclusión educativa escolar chilena, es la memoria de título ingeniería de Lolas y Vargas (UAI, 2015), donde se mostró capacidad de predicción dentro de una muestra de estudiantes reducida de excluides educacionales, pero donde no se pudo evaluar la efectividad de predicción entre datos de años pasados hacia set de datos más adelantados en el tiempo, ni la precisión en contexto de distribución natural del fenómeno en los datos.

Motivado también porque la exclusión educativa escolar es un fenómeno multicausal, y resulta casi imposible intentar definir arbitrariamente los criterios efectivos que envuelvan o permitan identificar al fenómeno de la exclusión educativa, es que se opta por una aproximación del aprendizaje de máquinas. Esto debido a que el aprendizaje de máquinas automático es una forma de reconocimiento de patrones de alta complejidad y dimensionalidad, lo que lo hace una herramienta adecuada para intentar abordar un problema tan complejo como lo son el fracaso en las trayectorias educativas de les estudiantes.

2.3.- Propuesta de valor

El potencial de agregación de valor que brinda el uso del paradigma de Ciencia de Datos para abordar problemáticas públicas, a través de técnicas del aprendizaje de máquinas, se ha convertido en más evidente en los últimos años. Esto ya que permite identificar y analizar patrones de alta complejidad, en grandes volúmenes de datos, y de diferentes vertientes, para la implementación y evaluación de mejores servicios y políticas públicas, e intentar combatir problemáticas públicas y sociales. En este caso específico, de la exclusión educativa escolar, el predecir tempranamente, y de forma más precisa, le permita al Estado llevar a cabo políticas de intervención que logren prevenir este fenómeno de manera más eficaz.

En términos económicos, el impacto de la exclusión educativa escolar sustancial. Un estudio del CIAE, que analizó una cohorte de 142,918 jóvenes entre 15 y 19 años que no completaron la educación media, estimó que la exclusión educativa escolar tiene un costo para el país de más de 5,000 millones de dólares en valor presente (CIAE, 2018). Por lo que los esfuerzos por disminuir la deserción de manera efectiva, y sostenida en el tiempo, están asociadas a dichas magnitudes.

3.-Marco conceptual

El presente capítulo sienta las bases conceptuales en que se desenvuelve el presente trabajo de memoria. Abordando las temáticas de la exclusión educativa, la ciencia de datos, y la política de datos abiertos.

3.1.- Exclusión educativa y deserción escolar

El concepto generalizado por el MINEDUC, pero también transversalmente a nivel institucional y en sus fundamentos académicos, para describir el fenómeno en que les estudiantes ven truncada su trayectoria educativa escolar es: "deserción escolar".

Este concepto atribuye a le estudiante no dar continuidad a sus estudios escolares, lo que, desde el enfoque de la inclusión, es considerado una aproximación errónea ya los factores que gatillan la ocurrencia del fenómeno son en su gran mayoría de carácter estructurales y externos, que empujan y condicionan el accionar de les estudiantes. Por lo que el autor considera correcto referirse al concepto como: "exclusión educativa".

A modo de respetar el concepto original utilizado por cada institución y/o autor, a lo largo del trabajo se presenta la terminología utilizada originalmente en cada caso, pero para efectos de unificación de sentido semántico es que se comprenderán todos los casos como "exclusión educativa".

A continuación, se describen los aspectos teóricos con que se aborda el fenómeno de la exclusión educativa o deserción escolar en el presente trabajo de memoria.

3.1.1.- Enfoques causales

Son principalmente dos los enfoques presentes en la literatura con que se aborda el análisis de causalidad de la deserción escolar (Marshall, 2003):

- el ámbito *extraescolares*: la situación socioeconómica y psicosocial de las y los estudiantes (ejemplos: condición de pobreza y marginalidad, anomia familiar, adscripción laboral temprana, embarazo adolescente) ,

- y un enfoque que refiere a las situaciones *intraescolares* que generan conflicto en la permanencia del estudiante dentro de las instituciones educativas (ejemplos: rendimiento, disciplina, convivencia)

Estos enfoques son los más reiterados y utilizados en la literatura referente a la temática de deserción escolar. Representando la convergencia mayoritaria de les autores en la materia. Es por esto, y el sentido que le hace al autor el considerar la deserción como un fenómeno multicausal y de naturaleza sistémico, que se decide considerarlos como la matriz teórica que ayude a explicar las causas y dimensiones que influencian la deserción escolar.

3.1.2.- Medición de la exclusión educativa o deserción

La forma de medir la deserción escolar depende del tipo de aproximación o características del fenómeno que se requiere analizar. Existen la tasa de incidencia, de prevalencia y de abandono, que se describen a continuación (MINEDUC, 2013):

- Tasa de incidencia

Se refiere a los estudiantes que no se matriculan en un año escolar, habiéndose matriculado el año anterior. Esta se presenta en dos subdivisiones, la tasa global de deserción, y la tasa deserción del sistema regular

  - Deserción global

Se refiere a quienes no poseen matrícula en ninguna forma del sistema educacional (regular y de adultos), no habiendo finalizado sus estudios y presentando matrícula el año anterior.

  - Deserción del sistema regular

Por sistema regular se entiende la educación escolar para niñes y jóvenes, diferenciándose de los sistemas de educación especial y de adultos. La medición de la tasa de incidencia de deserción del sistema escolar regular no considera si las y los estudiantes migraron al sistema de educación para adultos, sólo si ya no se encuentran en el sistema regular.

- Tasa de prevalencia

La tasa de prevalencia refiere a jóvenes en cierto rango etario (15 a 19 años, y 20 a 24 años) que no se han graduado de la educación media y que no asisten a establecimientos educacionales en un año respectivo de medición.

- Tasa de abandono

Se refiere a estudiantes que se han retirado formal o informalmente de un año escolar, habiendo comenzado.

3.2.- Ciencia de datos

La ciencia de datos, o ciencia intensiva en datos, en palabras de Jim Gray, ha emergido como el cuarto paradigma científico (Hey, Tansley & Tolle, 2009) luego de la ciencia experimental, teórica y computacional.

A grandes rasgos se constituye de tres vertientes principales: la programación científica, la estadística y matemática, y las áreas respectivas del conocimiento sobre el fenómeno estudiado. Siendo esta triada la responsable de que la comunidad científica, independiente de si es para la neurociencia o para la astronomía, ha cultivado en los últimos años distintos tipos de herramientas y metodologías para abordar sus problemáticas respectivas. Este ecosistema mundial se ha visto potenciado gracias al internet, lenguajes comunes de programación y la filosofía de código abierto.

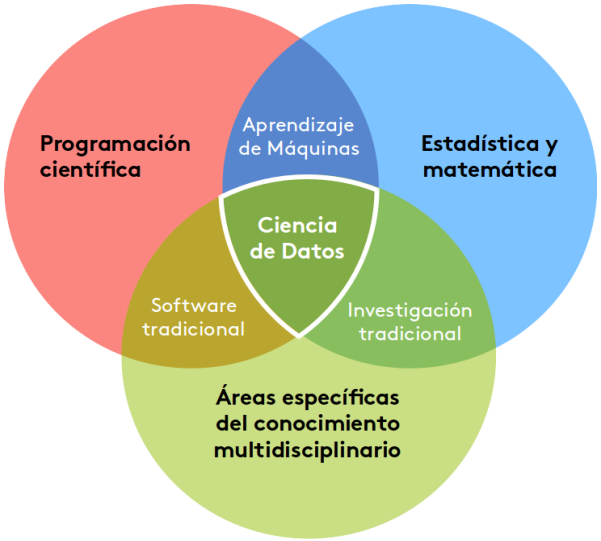

Ilustración 5: Interdisciplinareidad de la Cienca de Datos, Rayid Ghani (2017)

Es importante comprender que la ciencia de datos es eminentemente multidisciplinaria, ya que requiere de la convergencia de diversas disciplinas y técnicas. Como se intenta describir en la *Ilustración 7*, se aprecia en los círculos intersectados las distintas áreas que convergen en la ciencia de datos, como la estadística, el reconocimiento de patrones, la neurocomputación, el aprendizaje de máquinas, la inteligencia artificial, la minería de datos, las bases de datos y procesamiento de datos y la visualización. Destacando también el concepto de Descubrimiento de Conocimiento en los Datos (*Knowledge Discovery in Databases, KDD*) como el concepto precursor de la ciencia de datos en su carácter englobante, ya que comprende el proceso de inicio a fin en la extracción y descubrimiento de conocimiento nuevo a partir de los datos. En el círculo externo se pueden apreciar las distintas habilidades que se requieren para el desarrollo íntegro de la ciencia de datos desde cualquier enfoque: análisis de negocios, estrategia de negocios, dominio disciplinar, habilidades comunicativas, capacidad de presentación de información, rigurosidad, y resolución de problemas.

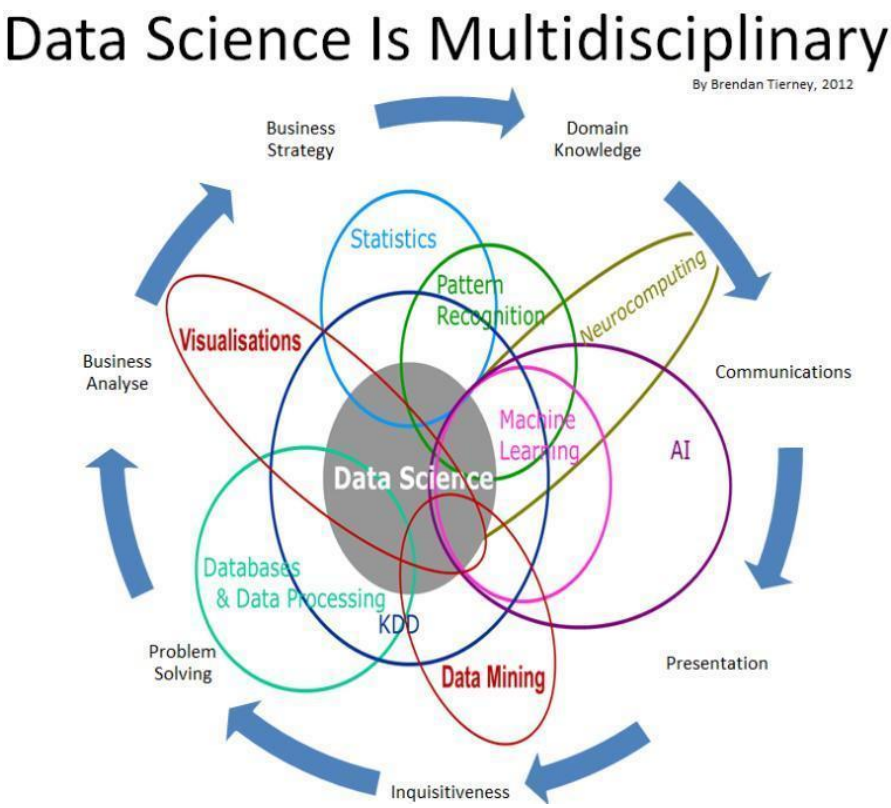

Ilustración 6: la Ciencia de Datos es Multidisciplinaria, Brendan Tierney (2012)

La ciencia de datos como paradigma para solucionar problemáticas públicas es situada por el Laboratorio de Gobierno dentro de la convergencia de la innovación pública y los procesos de modernización del Estado. Pero para que

el desarrollo de la ciencia de datos sea sostenible y efectiva en este campo, se plantea que, tanto en el ámbito de la innovación como de la modernización, se debe cumplir con ciertas condiciones y enfoques:

- Modernización del Estado: se requiere de una infraestructura habilitante, tanto en máquinas como en velocidad de conexión; a nivel institucional se requiere de la capacidad de interoperabilidad de datos entre las instituciones públicas para poder acceder a los datos necesarios para los proyectos de innovación, también la construcción de espacios multiministeriales que permitan abordar de manera multisectorial problemáticas que son de carácter multisectorial.

- Innovación pública: se requiere de poner el foco en las personas; una mirada sistémica de las problemáticas; desarrollo en cocreación y coproducción de las soluciones; generar espacios de experimentación que permitan un espacio de desarrollo innovador; y generar una alta capacidad de comunicación.

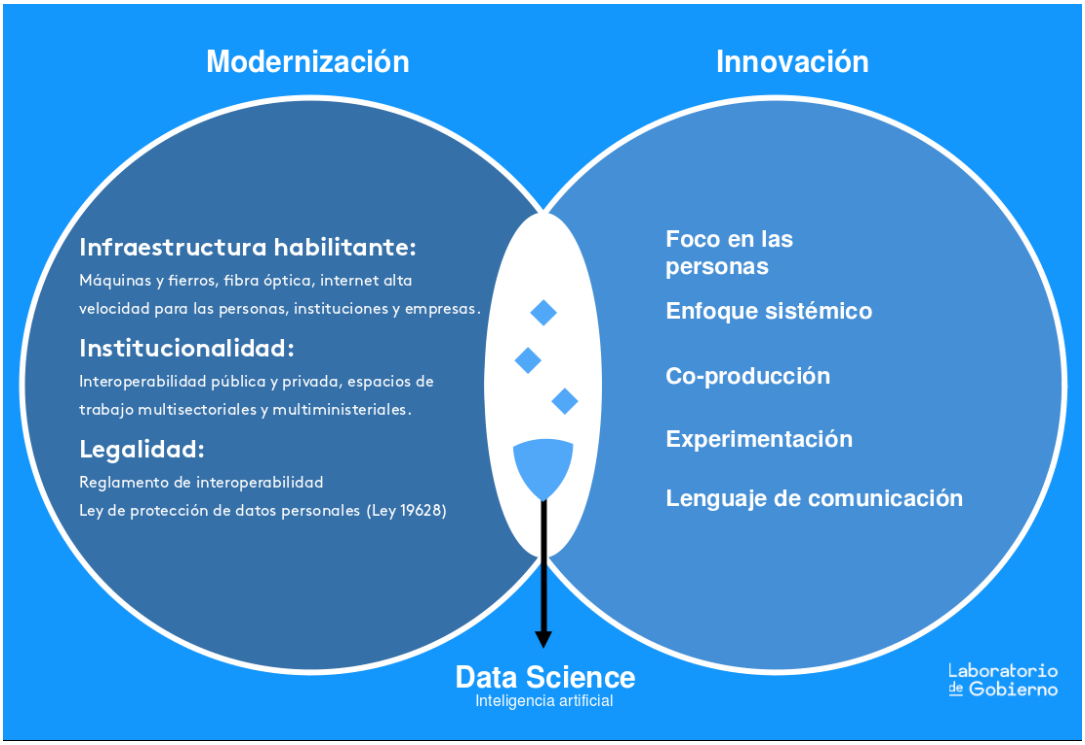

Ilustración 7: Ciencia de datos: convergencia de la Modernización del Estado e Innovación pública, Laboratorio de Gobierno y elaboración propia (2017)

3.2.1.- Aprendizaje de Máquinas (*Machine Learning*)

Ya es claro que nos encontramos en la era de los datos masivos, en los últimos años se han creado más datos que en toda la historia de la humanidad, de diversa índole y en distintos formatos. Para extraer mayor información de estas vastas cantidades de datos es que se requiere de métodos automatizados de análisis de datos: esto es el Aprendizaje de Máquinas. En particular el aprendizaje de máquinas se define como un conjunto de métodos que pueden detectar patrones de alta complejidad de manera automática, y luego utilizar este aprendizaje de los patrones para predecir con nuevos datos, o ejecutar otras tomas de decisión en entornos de incertidumbre (Murphy, 2012).

Existen diversos tipos de aprendizajes. Para efectos de este trabajo de memoria se utilizará el aprendizaje supervisado y el no supervisado:

3.2.1.1.- Aprendizaje supervisado

El aprendizaje supervisado trata de algoritmos a los cuales se les entrega un conjunto de datos muestrales en conjunto con etiquetas que categorizan dichos datos. Mediante el entrenamiento de dicho algoritmo posteriormente se puede predecir la etiqueta que corresponde a muestras de datos que no la tienen. Los tipos de algoritmos que componen el aprendizaje supervisado son de clasificación y regresión, diferenciándose en que la clasificación genera resultados de etiquetas discretas, y, la regresión, continuas (Murphy, 2012).

- Árboles de decisión (CART)

Los árboles de decisión para la clasificación y regresión (CART), en el aprendizaje de máquinas, son usados por una amplia familia de algoritmos basados en su generación automática. Son un método no-paramétrico de aprendizaje supervisado, utilizado para problemas de regresión y clasificación. Están compuestos por nodos de decisión, en que se selecciona una característica del modelo, del cual salen ramas para cada posible valor de este, hacia otro nodo, donde se selecciona otra característica a evaluar, hasta

concluir en las hojas que representan la respuesta que entrega el modelo para el dato procesado.

Para la selección de la característica a ser evaluada en un nodo, y el criterio de división en sus ramas, se utilizan métricas como la Impureza de Gini, o la Entropía, para comparar los sets de datos resultantes de todas las posibilidades de división del nodo. Así la partición que genere una mejor respuesta a las métricas mencionadas es la que se utiliza para la división del árbol.

En particular, la impureza de Gini revela la probabilidad de equivocarse al asignar una clase aleatoriamente (según la distribución en que se encuentran las clases en el set de datos) a un dato seleccionado aleatoriamente. En otras palabras, cuán impuro es un conjunto de datos, si consideramos un conjunto homogéneo, en términos de clases, como el máximo nivel de pureza (osea 0).

➔ Entonces la función de partición de Gini se define como:

$$\phi(s,t) = I_G(t) - \sum_{n}^{N} P_n * I_G(n)$$

donde S es la partición candidata, N son los nodos hijos, P_n es la proporción de muestras del nodo t que van al nodo hijo n, y $I_G(t)$ es la impureza de Gini dado por:

$$I_G(t) = 1 - \sum_{i}^{I} t_i$$

t_i: proporción de la clase i en la muestra del nodo t.

Así, se busca s^* para cada nodo tal t que:

$$\phi(s^*,t) = \max_{s \in S} \phi(s,t)$$

con S conjunto de posibles particiones.

El proceso de creación de un árbol de decisión se describe básicamente de la siguiente forma:

- ➔ Se inicializa en el nodo raíz, que contiene toda la muestra.
- ➔ Se selecciona una partición, es decir una característica y su criterio de división en ramas (usualmente la función de Gini).
- ➔ Se divide el nodo a través de las ramas, apuntando a nuevos nodos donde se repetirá el proceso hasta alcanzar el criterio de detención (hegemonía suficiente de una clase en el set de datos resultante, impureza de Gini igual a 0, u otro criterio preestablecido).
- ➔ Cuando se cumple el criterio de detención, se crea un nodo final conocido como "hojas", el cual contiene la clase hegemónica del set de datos, y que se le asignará al dato procesado por el árbol.
- ➔ Posteriormente, de manera opcional, se puede generar un proceso de poda de los nodos que no aportan mayormente al modelo, reduciendo la complejidad, y dando mayor eficiencia.

A continuación, se muestra un esquema de ejemplo de un árbol de decisión de dos ramas, construido para un set de datos con dos clases (pelotas negras y pelotas cruz), donde se muestra en A gráficamente la distribución de los datos en sus características X e Y, y las particiones definidas por el árbol de decisión. En B se muestra el esquema del árbol, donde se inicia en el nodo RAÍZ con la muestra de 20 pelotas negras y 10 pelotas cruz. Se define la partición en la característica X con criterio >3, por lo que se divide en dos nodos hijos. El nodo de la derecha resulta en una hoja, ya que toda la muestra es perteneciente a una sóla clase (etiquetado como 1); y el nodo de la izquierda, que contiene 5 muestras de pelotas negras, y 10 de pelotas cruz, define un nuevo criterio de partición para la característica Y con criterio >3. Así se resulta en dos nuevas hojas, etiquetadas con 2 y 3, en ellas se observa que no fue necesario alcanzar una homogeneidad de la clase en la muestra, sino un criterio de sensibilidad al menos menor al 20% de error (hoja número 2, sus clases están en proporción 4:1).

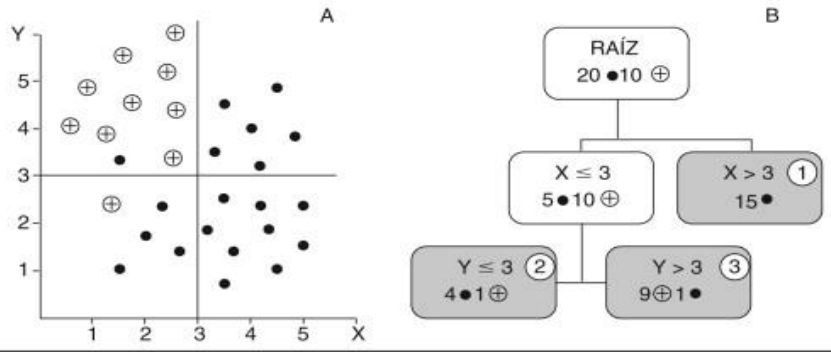

Ilustración 8: Aproximación a la metodología basada en árboles de decisión (CART), Trujillano J et al. (2008)

- Support Vector Machines

Support Vector Machines es una familia de algoritmos de aprendizaje supervisado de máquinas, usado para problemas de clasificación y de regresión.

El funcionamiento básico de las SVMs es encontrar un hiperplano óptimo que separe las clases a clasificar, al maximizar la distancia (o márgenes) entre los puntos más cercanos de cada clase con el hiperplano. En la *ilustración 11* a continuación, se muestran dos clases, la azul y la roja, donde $w^T x + b = 0$ define al hiperplano de parámetros w y b que separa las clases, y son los que se requieren encontrar.

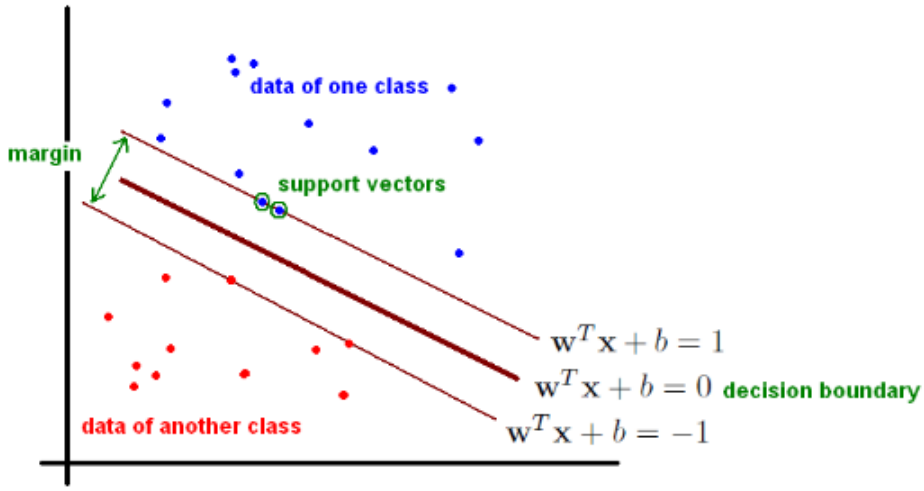

Ilustración 9: ejemplo SVM, Zoya Gavrilov

Así el problema de optimización que resuelve el ajuste de los parámetros w es la siguiente (para el caso rígido, sin regularización de relajación de clasificación):

$$\min_{w,b} \frac{1}{2} w^T w$$
$$s.t.\ y_i(w^T x_i + b) \geq 1$$
$\forall i \in \{1, 2, ..., N\}$ con N número de datos.
$$w \in R^d$$
$$b \in R$$

Los SVMs son clasificadores lineales que pueden ser utilizados como clasificadores no lineales. Esto es posible tras elevar los vectores de datos a un espacio dimensional mayor $x_i \rightarrow \phi(x_i)$ a través de una función ϕ. En este nuevo espacio dimensional elevado es posible encontrar un hiperplano que separe linealmente las clases que no eran linealmente separables en el espacio dimensional original. Lo que se traduce en transformar la restricción de la función de optimización original a:

$$s.t.\ y_i(w^T\phi(x_i) + b) \geq 1$$

A través del Lagrangiano se puede encontrar el problema dual de la recién descrita función de optimización:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i)\phi(x_j)$$

$$s.t.\ \alpha_i \geq 0\ \forall i, \sum_{i}^{N} \alpha_i y_i = 0$$

$\forall i \in \{1, 2, ..., N\}$ con N número de datos.

Y gracias al llamado "Kernel trick", que se basa en la utilización de una función Kernel para calcula el producto punto de la proyección de alta dimensionalidad de los vectores evaluados sin tener que expresarlos explícitamente en dicha alta dimensionalidad. Así las SVMs pueden utilizar una proyección de infinita dimensionalidad de los vectores de datos, ya que sólo se calcula su producto punto. En su versión de márgenes duros y no lineal (utilizando función de Kernel), el problema de optimización se puede modelar como:

$$\max_{\alpha} \sum_{i}^{N} \alpha_i - \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$s.t.\ \alpha_i \geq 0\ \forall i, \sum_{i}^{N} \alpha_i y_i = 0$$

donde N son la cantidad de la muestra de datos de entrenamiento, x_i datos de entrenamiento, y_i etiquetas de entrenamiento $\in \{-1, 1\}$, $K(x_i, x_j)$ es la función Kernel evaluada en los datos de entrenamiento x_i e x_j, y α es el vector de parámetros a optimizar.

Así la función de clasificación queda definida como:

$$f(x) = \sum_{i}^{N} \alpha_i y_i K(x_i, x) + b$$

A continuación, se muestra una imagen en que se puede apreciar cómo a la derecha, tras proyectar en un espacio dimensional superior gracias a la función de Kernel, se puede separar linealmente las clases, lo que significa en el espacio dimensional original (imagen de la izquierda) una clasificación no lineal.

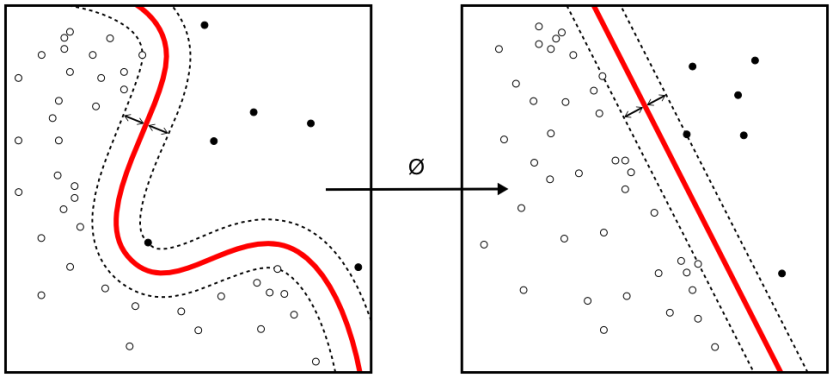

Ilustración 10: Máquinas Kernel son usadas para computar una función no separable linealmente en un espacio de alta dimensionalidad separable por una función lineal, Alisneaky (2011)

- Redes neuronales artificiales

Las redes neuronales artificiales son una familia de modelos matemáticos utilizados para resolver usualmente problemas de clasificación. Existen diversas variaciones dentro de las redes neuronales, como las redes convolucionales (utilizadas para la extracción automática de características), las redes completamente conectadas (utilizadas usualmente para la clasificación), las redes recurrentes (utilizadas en problemas como la traducción automática). En particular se describe una red completamente conectada utilizada para clasificar, ya que es el problema abordado en el presente trabajo de memoria.

El funcionamiento básico de una red neuronal artificial completamente conectada es:

- ➔ Ingresa a la red el vector que representa las características del estudiante. Cada dimensión ingresa a una neurona de entrada.
- ➔ La red neuronal computa una predicción para cada caso.
- ➔ Al contrastar la predicción con la respuesta real de la clasificación, se calcula una función de costo o pérdida.
- ➔ Con el fin de minimizar esta función de pérdida, se utiliza el método "Back-propagation" para calcular y ejecutar la modificación/actualización de los pesos de la red según su influencia en la pérdida.
- ➔ Se repite el proceso hasta alcanzar el nivel de entrenamiento deseado, prestando atención por sobre todo al sobreajuste que se pudiera producir. Esto se logra contrastando la disminución de la pérdida para el set de entrenamiento y el de test.

Otro elemento, relevante a destacar, es el Teorema de aproximación universal, que establece que una red neuronal feed-foward (como la descrita recién), con una capa oculta de finitas neuronas con funciones de activación no lineales, puede aproximarse, con precisión arbitraria, a cualquier función con número finito de discontinuidades.

La función de una red neuronal artificial feed-foward completamente conectada, ya entrenada, de una capa oculta con función de activación Relu, para dos clases (utilizando una función Sigmoidal hacia la capa/neurona de salida) está dada por la siguiente ecuación:

$$f(x) = Sigm(Relu(x * W + b) * U + c)$$

Donde x es el vector a evaluar; W es la matriz de pesos entre la capa de entrada y la capa oculta, de dimensiones $d \times N$, con d la dimensionalidad de x , y N el número de neuronas de la capa oculta y ; U es la matriz de pesos desde la capa oculta a la capa de salida, con dimensiones $N \times 1$ ya que representaremos el resultado de las dos clases en una sola neurona de salida; b y c representan el sesgo de entrenamiento; $Sigm$ (en el caso multiclase se utilizaría $Softmax$) y $Relu$ son funciones de activación definidas por:

$$Sigm(x) = \frac{1}{1 + e^{-x}}$$
$$Relu(x) = max(0, x)$$

Recalcar que son W, U, b y c las matrices y vectores de los pesos que se ajustan a través del entrenamiento de la red neuronal vía *backpropagation*.
A continuación, se presenta el esquema de una red neuronal multicapa, como la recién descrita. En este caso se muestra una capa entrada de 5 neuronas (dimensiones del vector x), una capa oculta con 6 neuronas, y una neurona de salida), una capa oculta con 6 neuronas, y una neurona de salida:

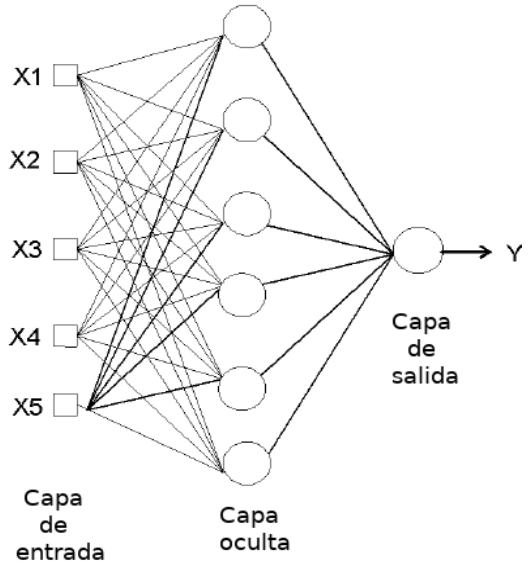

Ilustración 11: Estructura general de una arquitectura de red neuronal MLP , Braz J. (2006)

- Meta algoritmos

En el mundo de los algoritmos de aprendizaje de máquinas también existen los meta-algoritmos. Es un algoritmo que mediante la selección, variación o ensamble de modelos, construye un algoritmo final a partir de otros. A continuación, se describen meta-algoritmos tanto de Bagging, como Boosting.

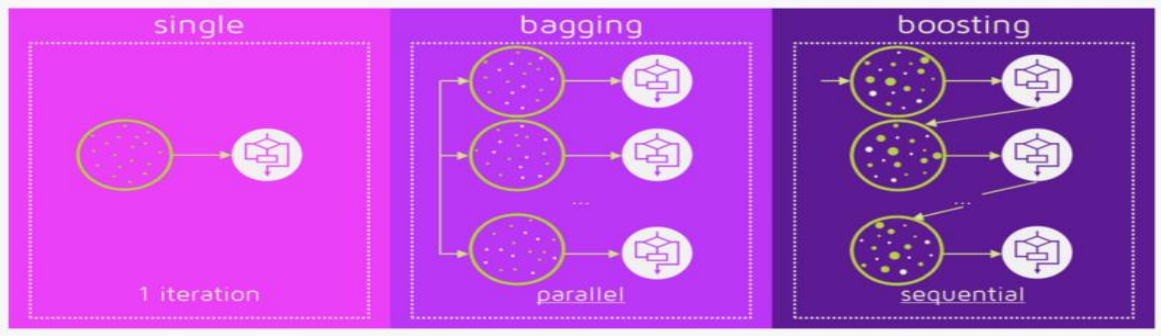

Ilustración 12: diferencia entre algoritmo único, bagging y boosting, Ana Porras (2016)

- Bagging

El “Bagging” es una forma de ensamble de algoritmos, que se basa en el entrenamiento en paralelo de clasificadores débiles, para luego combinarlos y obtener un clasificador fuerte. Siendo Random Forest su más implementación más conocida.

- Random Forest

Random Forest es un meta-algoritmo de bagging que utiliza árboles de decisión. Esto significa que árboles de decisión solitarios (usados como clasificadores “débiles”) son entrenados en paralelo para posteriormente promediar sus respuestas y construir un clasificador “fuerte”. Cada árbol es entrenado con una muestra aleatoria de los datos originales, lo que se conoce como *tree bagging* lo que reduce el *overfitting*, y a la hora de la construcción de los nodos de cada árbol se selecciona una muestra aleatoria de las características de los datos a utilizar para evaluar la partición del nodo, lo que se conoce como feature sampling y reduce la correlación entre los árboles. Es literalmente un bosque aleatorio de árboles, en que cada árbol es un predictor sin sesgo, pero de alta varianza, y que al ser promedia esta se reduce, generando una importante herramienta del aprendizaje de máquinas.

- Boosting

El “Boosting” es una técnica ensamble que se diferencia del “Bagging” en que los clasificadores débiles no son entrenados en paralelo sino de manera secuencial, a través de alguna técnica que permita construir y seleccionar dentro de un grupo de candidatos a los nuevos integrantes del ensamble que mejoran el desempeño del modelo inicial, construyendo un clasificador fuerte.

- Adaptative Boosting

El meta-algoritmo Adaptative Boosting, o Adaboost, es una implementación de Boosting. Se basa principalmente en ajustar secuencialmente la distribución de la muestra de datos de entrenamiento usado por los clasificadores débiles, acentuando los datos que fueron mal clasificados por la secuencia anterior, y así prestarles más atención en la próxima secuencia de aprendizaje; también a cada clasificador débil seleccionado para incorporar al clasificador fuerte se le pondera por un peso que modela su relevancia para la respuesta final. Como meta-algoritmo, para clasificar puede utilizar cualquier tipo de clasificador como clasificador débil, pero está comúnmente implementado con árboles de decisión, siendo utiliza de esta forma en el presente trabajo de memoria. Esta

decisión se debe a la alta capacidad de explicabilidad que brindan los árboles de decisión, aportando así a una mayor descripción del fenómeno de la exclusión educativa.

Adaboost se formula de la siguiente manera:

Dado un conjunto de datos de entrenamiento $(x_1, y_2), ., (x_m, y_m)$ donde $x_i \in \Upsilon,\ y_i \in \{-1, 1\}$.

Se inicializan, para $t = 1$, los pesos $w_1(i) = 1/m$ para $i = 1, ..., m$.
Para $t = 1, ..., T$:

- ➔ Se entrena usando la distribución en los datos D_t.
- ➔ Se entrenan diversos candidatos de clasificadores débiles $h_t(x) : \Upsilon \to \{-1, 1\}$.
- ➔ Se selecciona el $h_t(c)$ que resulte en el menor error, definido por:

$$\epsilon_t = \frac{\sum_{i=1}^{m} w_t(i) I(y_i \neq h_t(x_i))}{\sum_{i=1}^{m} w_t(i)}$$

- ➔ Se calcula $\alpha_t = L \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$, con tasa de aprendizaje $L \leq 1$.
- ➔ Se actualizan los pesos para la siguiente iteración, para todos los x_i con $i = 1, ..., m$:

$$w_{t+1}(i) = \frac{\sum_{i=1}^{m} w_t(i) \exp(-\alpha_t y_i h_t(x_i))}{\sum_{i=1}^{m} w_t(i)}$$

- ➔ Así el clasificador fuerte queda definido por:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

donde $H(x)$ es el clasificador fuerte, y recordemos que sign(x) por lo que es ideal para referenciar a las clases de etiquetadas $\{-1, 1\}$, α_t es el peso de relevancia del clasificador débil $h_t(x)$ en el clasificador fuerte $H(x)$.

- Gradient Boosting

A diferencia del meta-algoritmos Adaptative Boosting, el Gradient Boosting no modifica la distribución de los pesos del conjunto de entrenamiento según los errores de clasificación, sino que ajusta iterativamente los parámetros de modelos débiles según al *psedo residuo* del modelo fuerte.

Una formulación básica de Gradient Boosting:

- ➔ Ajusta un modelo inicial $F_0(x) = y$.
- ➔ Luego ajusta un nuevo modelo a los residuos del anterior, $h_0(x) = y - F_0(x)$.
- ➔ Crea el nuevo modelo $F_1(x) = F_0(x) + h_0(x).$
- ➔ Luego ajusta un nuevo modelo al residuo del anterior respecto al y inicial, $h_1(x) = y - F_1(x)$.
- ➔ Crea el nuevo modelo $F_2 = F_1 + h_1(x)$.
- ➔ Continúa iterativamente hasta el criterio de detención.

- ➔ Así se construye el modelo fuerte:

$$F_M(x) = F_{M-1}(x) + h_{M-1}(x)$$

con M: número de modelos agregados al inicial.

En otras palabras, Gradient Boosting es un meta-algoritmo iterativo que en cada iteración ajusta un modelo a los residuos que el modelo anterior generó, para luego ser agregado a dicho modelo, así va completando secuencialmente lo que el modelo anterior no pudo cubrir.

A diferencia del descenso de grandiente, Gradient Boosting no ajusta parámetros que interactúan con las características, sino que tuerce los modelos desde el ajuste al residuo generado en la iteración anterior.

- XGBoost

XGBoost, o eXtreme Gradient Boosting, es una meta-algoritmo basado en el framework de Gradient Boosting, implementado para árboles de decisión. Este utiliza el algoritmo Newton-Boosting, en contraste con su antecesor Gradient

Boosting Machine para árboles que utiliza Gradient Boosting Tree. Dentro de las características más relevantes de XGBoost se encuentra:

- Utiliza el hessiano empírico de la función de pérdida tanto para construir los árboles como calcular los pesos de los nodos terminales u hojas. Esto mejora la calidad de los árboles construidos (al considerar criterio de segundo orden de la función), y aumenta la eficiencia y escalabilidad del modelo ya que no debe resolver problemas de optimización para calcular el peso de las hojas.

- Agrega variabilidad en la construcción de cada árbol, aportando robustés al modelo, al permitir muestreo aleatorio tanto en las características a nivel general del modelo y por nivel (evaluación de partición de los nodos), como en la muestra de datos con que se entrena cada modelo.

- Agrega regularización tanto L1 y L2, que se traduce en tender a cero (*sparsear*) los pesos de las hojas, y tender a lo más pequeño posible estos pesos, respectivamente.

La función objetivo, para cada iteración de agregar un nuevo modelo débil, está compuesta por un componente de función de pérdida $L()$, y un componente de regularización $\Omega()$:

$$Obj^{(m)} = \sum_{i=0}^{n} L(y_i, \hat{f}^{(m-1)}(x_i) + f_m(x_i)) + \Omega(f_m) + constant$$

Donde $\hat{f}^{m-1}()$ es el modelo construido mediante iteraciones previas de agregación, y $f_m()$ es el candidato a nuevo modelo a agregar.

La función de regularización $\Omega()$ puede ser descrita de la siguiente forma:

$$\Omega(f_m) = \gamma T_m + \frac{1}{2}\lambda \sum_{j=1}^{T_m} w_j^2 + \alpha \sum_{j=1}^{T_m} |w_j|$$

donde γ penaliza la cantidad de hojas T_m, y las regularizaciones $L1$ y $L2$ tienden a cero los pesos w_j y los fuerza a ser pequeños, respectivamente.

Aplicando la expansión de Taylor de segundo orden, y removiendo las constantes, la función objetivo puede ser descritar de la siguiente manera:

$$Obj^{(m)} = \sum_{i=1}^{n} [\hat{g}_m(x_i) f_m(x_i) + \frac{1}{2} \hat{h}_m(x_i) f_m^2(x_i)] + \Omega(f_m)$$

donde $\hat{g}_m(x_i)$: gradiente empírico, y $\hat{h}_m(x_i)$: es el hessiano empírico, de la función de pérdida respecto a la función predictora pasada, evaluada en para x_i:

$$\hat{g}_m(x_i) = [\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=\hat{f}^{(m-1)}(x)}$$

$$\hat{h}_m(x_i) = [\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}]_{f(x)=\hat{f}^{(m-1)}(x)}$$

Tiene que:

$$G_j = \sum_{i \in I_j} \hat{g}_m(x_i)$$
$$H_j = \sum_{i \in I_j} \hat{h}_m(x_i)$$

con I_j : conjunto de índices del nodo j.

Y para el caso $\alpha = 0$, es decir sin regularización L1, se puede expresar la función objetivo en sólo términos de los nodos terminales o hojas T de los árboles:

$$Obj^{(m)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T$$

A continuación, se describe la construcción general del modelo, desde la construcción de cada árbol y sus pesos en las hojas, hasta la composición del modelo fuerte. Para efectos de simplicidad se continúa con el caso $\alpha = 0$:

➔ Constucción del árbol:

El aprendizaje de la estructura del árbol se basa en el método *greedy*:

- empezar con un árbol de profundidad cero
- para cada nodo, probar todas las posibles particiones y quedarse con la de mejor $Gain$ o ganancia, definida de la siguiente manera:

$$Gain_j = \frac{1}{2}[\frac{G_{jL}^2}{H_{jL}+\lambda} + \frac{G_{jR}^2}{H_{jR}+\lambda} - \frac{(G_{jL}+G_{jR})^2}{H_{jL}+H_{jR}+\lambda}] - \gamma$$

donde j denota el nodo hoja a partir, L representa el resultado de la rama izquierda de la partición evaluada en el nodo j, y R corresponde al resultado de la rama derecha en dicha partición.

Si se propusieran varios árboles, nos quedamos con la mejor score, que se desprende de la función objetivo:

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T$$

➔ Cálculo pesos de los nodos terminales u hojas:

Para el árbol resultante del proceso reción descrito, se calcula el peso de cada nodo terminal u hoja directamente con la gradiente y el hessiano empírico:

$$\hat{w}_{jm} = -\frac{G_{jm}}{H_{jm}-\lambda}$$

➔ Construcción del modelo:

Así, la función del árbol recién descrito queda definida por:

$$\hat{f}_m(x) = \eta\sum_{j=1}^{T}\hat{w}_{jm}I(x \in \hat{R}_{jm})$$

donde η representa la tasa de aprendizaje (learning rate), que amortigua o achica la influencia del árbol en la composición general, y $\hat{R}_{jm}$: es la región definida por el nodo terminal j, del modelo m.

La función, para cualquier iteración m queda definida como:

$$\hat{f}^{(m)}(x) = \hat{f}^{m-1}(x) + \hat{f}_m(x)$$

Así, iterando para $m \in \{1, \ldots, M\}$, se construye el algoritmo fuerte:

$$\hat{f}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$$

- ❖ Diferencias entre XGBoost y GBM:

La mayor diferencia entre el Newton-Boosting, utilizado en XGBoost, con el Gradient Boosting Tree, utilizado en los algoritmos de Gradient Boosting Machine (GBM) para árboles, es que el Newton-Boosting utiliza el gradiente y el hessiano empírico para construir el árbol, y para definir el peso de sus hojas (como es descrito anteriormente). Pero el Gradient Boosting Tree utiliza solo el gradiente empírico, como lo describe la siguiente función de ganancia para construir el árbol

$$Gain_j = \frac{1}{2}[\frac{G_{jL}^2}{n_{jL}} + \frac{G_{jR}^2}{n_{jR}} - \frac{(G_{jL} + G_{jR})^2}{n_{jL} + n_{jR}}]$$

para el caso sin regularización, y donde G_{jL} corresponde al gradiente empírico evaluado para el resultado de la partición izquierda evaluada para el nodo j, G_{jR} corresponde a la partición derecha.

y posteriormente resuelve el siguiente problema de optimización para calcular los pesos de los nodos terminales o hojas, a diferencia del newton boosting que lo hace de forma directa:

$$\hat{w}_{jm} = \arg\min_{w_j} \sum_{i \in I_{jm}} L(y_i, \hat{f}^{(m-1)}(x_i) + w_j)$$

esto se traduce para el Newton-Boosting, al utilizar una aproximación de segundo orden (hessiano empírico) genera mejores estructuras de árboles, pero el gradient tree boosting genera mejores pesos de hojas, pero para estructuras de árboles menos precisas. También el Newton-Boosting resulta

menos costoso computacionalmente al no tener que resolver un problema de optimización adicional para calcular el peso de las hojas, pero se restringe a funciones de pérdida doblemente derivables, ya que requiere calcular el hessiano empírico.

3.2.1.2.- Evaluación de modelos

Para dirimir qué tan bueno es un modelo de aprendizaje supervisado depende del problema en cuestión a abordar. Existen distintas métricas para evaluar el desempeño, que a continuación serán presentadas.

- *Pérdida* o loss

La *loss* o pérdida se refiere a la métrica que el modelo busca minimizar, esta puede ser tan simple como la suma de los errores, los errores medios cuadrados, o la entropía cruzada, etc. Esta métrica es la que requiere mayor atención a la hora del entrenamiento, ya que muestra el proceso de aprendizaje. Para evaluar finalmente el mejor modelo, también hay que considerar las métricas que se describirán a continuación, ya que poseen un significado concreto respecto a las clases clasificadas, en cambio la pérdida o loss es una métrica de carácter general del modelo.

- Exhaustividad o Recall

La exhaustividad o *recall* representa la capacidad del modelo de explicar o clasificar un fenómeno en particular. Se calcula como la tasa entre verdaderos positivos y verdaderos totales de la clase evaluada. En el caso particular de las políticas públicas, y en la exclusión educativa escolar, esta es una métrica muy importante, ya que describe el porcentaje de les excluides educacionales que el modelo es capaz de capturar del fenómeno total.

- *Precisión* o Precision

La precisión o *precision* representa la capacidad del modelo de clasificar certeramente una clase en específico. Se calcula como la tasa entre los verdaderos positivos y la suma entre verdaderos positivos y verdaderos negativos. En otras palabras, representa la probabilidad con que un elemento clasificado de una clase específica resulta correctamente clasificado. También es importante destacar que, para el caso de la exclusión educativa, una baja precisión puede significar la identificación de un grupo en situación de riesgo.

- *Exactitud* o Accuracy

La exactitud o *accuracy* representa la capacidad general del modelo de clasificar bien el problema abordo. Se calcula como la tasa entre la suma de los verdaderos positivos y el total de predicciones realizadas. En otras palabras es una métrica que describe la probabilidad de clasificar correctamente a nivel general, y no en particular para cada clase. En problemas donde la muestra está desbalanceada, como es el caso de la predicción de la exclusión educativa escolar (aproximadamente 97% no excluides v/s 3% excluides), es necesario completar el análisis de evaluación usando las métricas recién descritas.

3.2.1.3.- Aprendizaje no supervisado

El aprendizaje no supervisado, a diferencia del aprendizaje supervisado, no utiliza etiquetas que definan a los datos muestrales. El objetivo es descubrir estructuras de interés en los datos, en otras palabras, extraer información desde los propios datos para generar análisis y acciones futuras. Ejemplos de tipos de algoritmos de aprendizaje no supervisado son los que permiten construir agrupaciones (*clusters*) y los que permiten descubrir factores latentes (reducción de dimensionalidad) (Murphy, 2012). La utilización más significativa en el presente trabajo de memoria será a través de los modelos de agrupación, para la construcción de los perfiles de fenómenos de exclusión educativa. A continuación, se describen los principales modelos.

- Gaussian Mixture Model

Este modelo requiere de la definición del parámetro de cantidad de agrupaciones a las cuales se ajustará el modelo; asume que las agrupaciones latentes poseen naturaleza Gaussiana; y a diferencia de K-Means (modelo clásico de agrupación), Gaussian Mixture Model está implementada con la distancia de Mahalanobis, no la Euclidiana. Esto significa que para distinguir cuán cerca está un dato del centro de su agrupación, pesan más las dimensiones con mayor correlación. Cada agrupación posee su propia matriz de varianza-covarianza, o precisión (inversa de la anterior), lo que le asigna a cada agrupación una propia "personalidad", lo que es deseado para el perfilamiento de personas.

El modelo Gaussian Mixture Model es una implementación del algoritmo expectation-maximization (EM), que es un método iterativo que maximiza la verosimilitud del modelo sobre los datos de entrenamiento para la construcción y asignación de las agrupaciones.

- Distancias entre vectores

Es relevante en el proceso de aprendizaje no supervisado, para el objetivo de construcción de tipologías de fenómenos de exclusión educativa, el definir cuál es el criterio que define la distancia entre los vectores que describen a les estudiantes. La evaluación de este parámetro depende exclusivamente de la naturaleza del problema. En este apartado se busca describir el concepto teórico de las métricas de distancia atingentes al presente trabajo de memoria. Se compara la distancia Euclidiana con la distancia de Mahalanobis, describiendo el porqué la segunda es la más adecuada para el proceso de perfilamiento de fenómeno de exclusión educativa.

- *Distancia Euclidiana*

La distancia Euclidiana es igual a la raíz de la suma de los cuadrados de las diferencias entre las dimensiones de dos vectores. Asigna el mismo peso para cada diferencia, independiente de la naturaleza de la relación entre las dimensiones de los vectores.

$$D_E = \sqrt{\sum_{j}^{n} (x_j - y_j)^2}$$

D_E: distancia Euclidiana

$\vec{x}$e $\vec{y}$: vectores de dimensionalidad n

- *Distancia de Mahalanobis*

A diferencia de la distancia Euclidiana, la distancia de Mahalanobis considera la matriz de varianza-covarianza entre las dimensiones o características del conjunto de datos que describen a les estudiantes. Esto resulta bastante deseable a la hora de caracterizar perfiles de personas.

Gaussian Mixture Model está implementado utilizando esta distancia, resultando en que cada agrupación resultante posee su propia matriz de varianza-covarianza. Así la distancia entre los vectores depende de cómo covarían las dimensiones de la agrupación, dando mayor peso a las "más

estables". En el caso en que la matriz de varianza-covarianza utilizada sea sólo su diagonal, es posible observar claramente que la diferencia entre cada dimensión se pondera respecto a la inversa de la varianza de dicha dimensión.

$$D_M(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}$$

D_M : distancia de Mahalanobis
Σ : matriz varianza-covarianza

Para entender mejor lo anterior, se propone la siguiente analogía: En un galpón se encuentran reunidas cientos de personas que gustan de la música, se les entrega una lista con un mix de 100 canciones conocidas para que les asignen un puntaje de preferencia a cada una; Con estos datos procesado, se entrena un algoritmo de agrupación que identifica cuatro agrupaciones o tipologías de gustos musicales: fans de la cumbia, el reggaetón, la trova y el metal; ¿Qué tan cercano está el perfil musical de una persona, a otra dentro de la misma agrupación o tipología?, bajo la lógica de Mahalanobis, la distancia se vería influenciada por la correlación entre las características del grupo (cada grupo posee su propia matriz de varianza-covarianza de características). Es decir, a la distancia entre dos personas del grupo de la trova, la debiera influenciar más sus diferencias respecto canciones de Silvio Rodríguez (trova), que sus diferencias respecto canciones de Megadeath (metal). Bajo la lógica Euclidiana esto no sucedería, ya que las dos características importarían lo mismo a la hora de evaluar la distancia entre los dos perfiles musicales.

Como otra analogía para la comprensión de lo anterior, se propone pensar en el caso que, si tenemos una aglomeración de personas, en específico varias tribus urbanas juntas; y cada tribu representa una aglomeración obtenida por el modelo, observaríamos que en cada aglomeración hay dimensiones que se correlacionan de manera distinta. Esto significa que a la distancia entre las personas se ve afectada más por las dimensiones "estables". Es decir, para evaluar la distancia entre dos *metaleros* resultaría más relevante la diferencia entre el largo de su pelo, y la banda de música favorita que escuchan; y en el caso de unos *raperos* significarían sus diferencias respecto al sobre-ancho de sus poleras, o sus zapatillas. En el caso de la distancia Euclidiana, la influencia de cada una de las dimensiones es la misma.

3.2.2.- Visualizaciones

Un aspecto importante de la ciencia de datos, debido a su necesaria convergencia multidisciplinaria, y adecuación para la mejor toma de decisiones, es la efectiva transmisión de la información o conocimiento generado. Bajo esta máxima, es que las visualizaciones del comportamiento de los datos pueden generar mayor comprensión respecto a los fenómenos que describen. Así es, como las visualizaciones sirven para descubrir y transmitir nuevas evidencias que permitan una mejor aproximación a las problemáticas a abordadas.

3.3.- Datos abiertos

En la última década se ha instalado un nuevo paradigma sobre la administración y uso de datos recabados por parte del Estado a nivel mundial. Chile, si bien tiene muchos desafíos por delante, hoy ya ha comenzado a tomar parte a través de la agenda 2020 de modernización del Estado: los Datos Abiertos u Open Data. Este se basa en el principio de disposición sin restricción de los datos públicos o privados, obtenidos a través del ejercicio de los servicios y políticas pública. La protección de la identidad personal o la información sensible que se pudieran extraer a través de los datos respecto a la fuente que los proveyó es un resguardo fundamental de esta filosofía. Planteando necesidades de estándares ético que por ejemplo resguarden la privacidad de las personas pero que también habiliten un flujo de datos que dinamice y permita la creación de valor a partir de su uso masivo.

Los Estados, motivados tanto por potenciar mayores estándares de transparencia, fomentan así una mayor fiscalización por parte de la sociedad civil hacia las instituciones públicas, en un contexto de creciente desafección y desconfianza hacia ellas; como el permitir la incorporación de nuevos actores, ideas y métodos que potencien la innovación en la creación de soluciones y herramientas para abordar problemáticas públicas. En el caso chileno aún se encuentra en discusiones legislativas necesarias que permita el pleno uso de esta práctica.

4.- Metodología

En este capítulo se describe la metodología y herramientas de obtención y utilización de datos masivos para el cumplimiento exitoso de los objetivos del presente trabajo.

Para esto se utilizan herramientas de programación en lenguaje de programación Python, el cual cuenta con librerías de diversa índole para la exploración, tratamiento y utilización de los datos. Python cuenta con librerías pertinentes como Pandas (para el tratamiento de bases de datos), Numpy (cálculo matricial), Scikit-learn (herramientas de aprendizaje de máquinas, minería de datos y análisis de datos), TensorFlow (librería de redes neuronales), XGBoost (librería de implementación eficiente de Gradient Boosting), Matplotlib (visualización).

El desarrollo del presente trabajo, en términos de procesamiento y análisis de datos, y la construcción de modelos predictivos y su puesta en funcionamiento, se basa en la metodología CRISP-DM (por sus siglas en inglés: CRoss-Industry Standard Process for Data Mining) que describe el ciclo de vida de un proyecto de minería de datos en seis fases: Entendimiento del problema, entendimiento de los datos, preparación de los datos, modelamiento, evaluación y utilización (Chapman et al., 2000).

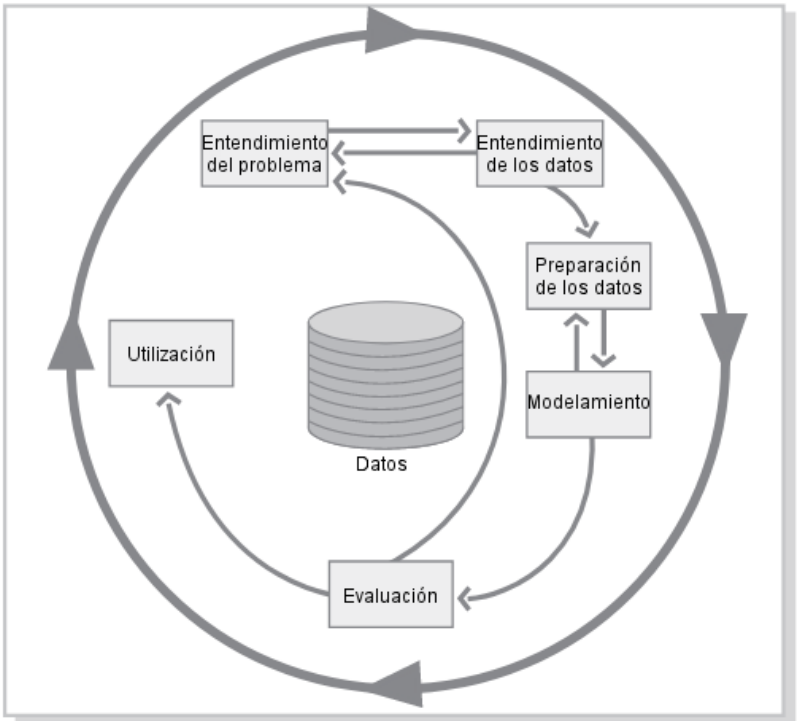

Ilustración 13: Metodología CRISP-DM, Chapman et al. (2000)

Como se describe en el esquema, a través de las flechas las fases se relacionan de una manera cíclica e iterativa, de manera interrelacionada. Evidenciando una lógica de mejoramiento constante y multi direccional.

A continuación, se detalla cada fase del ciclo:

4.1.- Entendimiento del problema

El foco de esta fase inicial está en entender el contexto, los objetivos y requerimientos del trabajo de memoria; y convertir este conocimiento en la formulación de un proyecto de minería de datos, permitiendo generar un plan preliminar para el complimiento de los objetivos.

Es por esto que en este proceso es crucial la interiorización respecto a la problemática de la exclusión educativa escolar. Tanto a través de la revisión de la literatura existente, entrevistas con las instituciones públicas incumbentes, entrevistas con organizaciones no gubernamentales que trabajen el tema, y entrevistas con quienes han vivido el fenómeno en carne propia, o revisión de testimonios recopilados.

4.2.- Entendimiento de los datos

La fase de entendimiento de los datos está compuesta por la obtención de los datos, y las actividades de exploración preliminar que permitan generar una mayor comprensión de estos: identificar problemas de calidad de los datos, obtener conocimiento no evidente de los datos (“insights”) y/o identificar subconjuntos de datos para formular hipótesis sobre información oculta. A continuación, se describen en mayor profundidad.

4.2.1.- Obtención de datos

Los datos son la materia prima para la concepción y desarrollo de modelos y herramientas que permiten la obtención de información y creación de conocimiento a partir de estos, para abordar y proponer soluciones a las problemáticas planteadas. Esto hace que la obtención de datos sea un paso inicial crucial para el presente trabajo de memoria de título.

En este contexto es que la utilización de los diversos portales públicos de obtención de Datos Abiertos como http://datosabiertos.mineduc.cl del Centro de Estudios del Ministerio de Educación; los mecanismos de transparencia del Estado de Chile que provee la Ley 20.285 sobre Transparencia y Acceso a la Información Pública; y el contacto directo con las Instituciones pertinentes y sus áreas de estudios: son los métodos pertinentes para la obtención de información pública.

- Datos ideales

En concordancia con la teoría, que plantea que los factores que inciden en la exclusión educativa se pueden categorizar como extraescolares e intraescolares, es que el conjunto de datos ideal para abordar el modelo de exclusión educativa debe comprender dichas dimensiones.

Así es como se puede dividir los factores extraescolar en información del estudiante en específico, del interior de su hogar, y del ambiente donde este está inserto, entre otros; y los factores intraescolar en el relacionamiento del estudiante con su entorno, el rendimiento o sintonía con una trayectoria educativa exitosa, y condiciones de la institución educativa, entre otros.

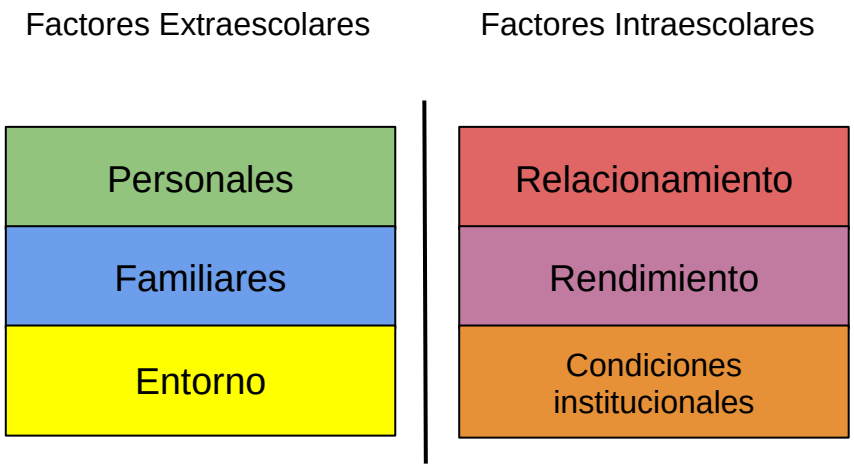

Ilustración 14: Factores incidentes en la deserción escolar, elaboración propia (2018)

Los datos ideales para el modelo debieran describir al menos estas dos categorías, y sus seis subcategorías. Por lo que los datos del MINEDUC, la Agencia de Calidad de la Educación, y la JUNAEB son atingentes a la hora de describir el aspecto intraescolares, y en menor medida lo personal y familiar; Para la descripción del ámbito extrainstitucionales los datos que pudiera

proveer el Registro Social de Hogares RSH, instrumento que sustituye la Ficha de Protección Social, son considerados claves para una mejor comprensión de la realidad de los hogares de les estudiantes. Este instrumento es administrado por el Ministerio de Desarrollo Social.

4.2.2.- Exploración preliminar

Al tratarse de cantidades masivas de datos (cientos de miles o millones de muestras, con decenas o cientos dimensiones, en cada conjunto de dato a utilizar) el análisis de calidad y la obtención de visiones internas a evaluar, la utilización de métodos de programación es fundamental ya que no se justifica la verificación manual que puede demorar demasiado tiempo. Ejemplo de esto es la evaluación de la cantidad de datos nulos en las bases de datos.

4.2.3.- Visualizaciones de datos para comprender mejor el problema

Para el mejor entendimiento del problema, a través del entendimiento de los datos, es que se deben explorar visualizaciones que permitan sentar nueva evidencia para la innovación, diseño e implementación de políticas públicas integrales para abordar la exclusión educativa escolar. El autor considera que no basta con predecir, sino que también se debe aportar a mejorar la comprensión del fenómeno de la exclusión educativa. Es por esto que esta temática se comprende como un objetivo específico. Es fundamental en este paso el formular preguntas que desafíen los límites de la literatura actual sobre exclusión educativa escolar, con el objetivo de ampliar y fortalecer la base en que se yerguen políticas y servicios públicos para su prevención.

4.3.- Preparación de los datos

La fase de preparación de datos trata de todas las actividades que se requieran para construir, a partir de los datos en bruto, el conjunto final de datos que serán recibido por los modelos a construir. Tanto la identificación de los tipos de datos (nominal, ordinal, intervalo o razón (SPSS FREE, n.d.)), la normalización de los datos, y la construcción de indicadores a través del análisis previo.

4.4.- Modelamiento

En esta fase se selecciona, entrena, y evalúa variados modelos predictivos, por lo que se requiere la calibración de sus respectivos parámetros para el desempeño óptimo. El proceso de modelamiento conversa constantemente con la preparación de los conjuntos de datos a utilizar, tanto en distintas variantes requeridas de su preprocesamiento (la imputación de valores que reemplacen los miss-values), o la selección de características (features) para mejorar el desempeño.

4.4.1.- Selección y construcción de modelos

Para esta fase se abordan dos tipos de modelos distintos, a considerar para el cumplimiento de los primeros dos objetivos específicos del presente trabajo: modelos de clasificación para realizar predicciones probabilísticas del potencial de exclusión educativa para cada estudiante; y modelos agrupamiento, para calcular y construir las tipologías de fenómeno de exclusión educativa, que serán asignadas para caracterizar a cada estudiante. A continuación se describe cada tipo de modelo.

- Selección de características

A través de modelamiento preliminares y evaluaciones previas, se identifican las características, o variables a utilizar, dentro de cada modelo. Esto es relevante dado que permite alcanzar una mayor identificación de los factores que inciden en el fenómeno estudiado, y lograr mayor eficiencia computacional.

Para esto es que se utiliza una técnica de selección de características basadas en árboles de decisión. Esto ya que en el proceso de la construcción de los árboles de decisión, se identifican las características que mayor ganancia aportan a la predicción.

- Modelos de clasificación

Los modelos de clasificación se utilizan para asociar patrones complejos de datos a una clase específica. Esto se logra mediante un proceso de aprendizaje supervisado. Con estas herramientas se busca que las características que describen a cada estudiante sean asociadas a un potencial de exclusión educativa del sistema escolar.

Para esto es que se configuran y entrenan diversos tipos de clasificadores o regresores (XGBoost, Adaboost, Random Forest, Support Vector Machines SVMs, Redes Neuronales), de los cuales se seleccionarán los con mejor desempeño tras sus respectivos ajustes.

- Modelos de agrupación

Los modelos de agrupación se utilizan para construir aglomeraciones dentro de los conjuntos de datos bajo distintos criterios, que pueden ser de distancia o relación. En particular este tipo de modelos se utilizará en el trabajo de memoria para la construcción de perfiles o tipologías de los fenómenos de exclusión educativa de les estudiantes. Al agrupar e identificar los vectores de datos que describen a les estudiantes, se construye la materia prima para que, tras un trabajo de interpretación, los perfiles puedan ser utilizados para la prevención de la exclusión educativa escolar. Esta es una de los puentes interdisciplinarios de la ciencia de datos, entre el aprendizaje de máquinas no supervisado y las ciencias sociales, que resulta crucial para el efecto que una política pública de prevención pudiera tener. Es por esto que, la construcción y asignación de tipologías de fenómenos de exclusión educativa, se conforma como objetivo específico del presente trabajo de memoria. Los modelos a evaluar son Gaussian Mixture y K-Means, utilizando el primero debido a su mayor idoneidad al estar implementado con la distancia de Mahalanobis.

4.4.2.- Construcción de perfiles de exclusión educativa

Luego de obtener las agrupaciones de los perfiles de les estudiantes, es necesario construir una muestra de datos por cada agrupación que represente las características más relevantes que describen a dicho conjunto. Para esto se proponen los siguientes pasos:

1. Se debe identificar, gracias a la diagonal de la matriz de varianza-covarianza de cada agrupación, las dimensiones de menor varianza. Osea las más estables dentro del cluster.

2. Obtener una muestra de los datos más cercanos al centro de la agrupación. Así se obtendrán los magnitudes que permitirán la posterior interpretación del perfil de exclusión educativa, como la mediana, máximo, mínimo, y desviación estándar, de las características más relevantes encontradas en el paso anterior.

3. Analizar las magnitudes y relación existente entre las características de cada agrupación, para la construcción de una descripción que caracteriza el perfil o tipología de la agrupación. Este paso requiere de habilidades profesionales ligadas a las ciencias sociales, como sicología, sociología y/o antropología. El autor considera que este paso representa un vínculo interdisciplinario relevante para la ciencia de datos, convergiendo la matemática y procesamiento computacional, con las ciencias sociales.

4.5.- Evaluación

La fase de evaluación trata de la comparación de las métricas de resultado de los modelos ya ajustados (loss, recall, precision, accuracy) e identificar el balance óptimo de estas para el problema en cuestión. Esto se representa en una tabla comparativa. También se utiliza para evaluar modificaciones a los modelos.

En el caso de la exclusión educativa escolar, y las políticas públicas que apunten la cobertura de programas o servicios críticos para personas, es muy relevante el recall, ya que una menor exhaustividad o *recall* significa estudiantes con potencial de exclusión educativa no cubiertos por un programa de prevención. Por otro lado a mayor precisión o *precision* es menor el gasto innecesario incurrido en por la política pública de prevención.

5.- Desarrollo del proyecto

En el presente capítulo se describe el proceso de desarrollo de la solución propuesta. Cabe recordar que la metodología propuesta no es de carácter lineal, sino que iterativo y de retroalimentación entre los distintos niveles. Esto debido a que en cada nivel pueden surgir nuevas luces que nutran el proceso en general.

5.1.- Entendimiento del problema

Para generar el entendimiento del problema es que, aparte de investigar la literatura respectiva, se agendaron y realizaron entrevistas con el Centro de Estudios del MINEDUC, y la Agencia de la Calidad de la Educación. Identificando así los aspectos teóricos de la exclusión educativa, como su realidad descrita previamente, y las mejores aproximaciones al problema propuesto.

Así también se calculó la exclusión educativa escolar del sistema regular, comprendida entre los períodos 2010-2018. Es relevante destacar que la exclusión educativa o deserción global representa consistentemente cerca del 60% de la regular.

Período	Exclusión educativa regular (%)	Exclusión educativa regular (N)	Exclusión educativa global (%)	Exclusión educativa global (N)	% Exclusión educativa global/regular	Matrícula regular (N)
2010-2011	3.48%	105,160	2.18%	65,765	62.54%	3,022,074
2011-2012	4.30%	127,554	2.64%	78,366	61.44%	2,969,142
2012-2013	3.48%	100,910	2.14%	61,986	61.43%	2,901,190
2013-2014	3.42%	98,048	2.07%	59,267	60.45%	2,869,544
2014-2015	3.38%	96,193	2.00%	56,975	59.23%	2,849,600
2015-2016	3.23%	91,711	1.93%	54,963	59.93%	2,842,672
2016-2017	3.00%	85,374	1.82%	51,885	60.77%	2,844,187
2017-2018	2.87%	82,092	1.81%	51,882	63.20%	2,859,064

Tabla 1: Exclusión educativa regular y global del sistema escolar regular, períodos 2010-2018, elaboración propia (2018) – Fuenta datos: MINEDUC

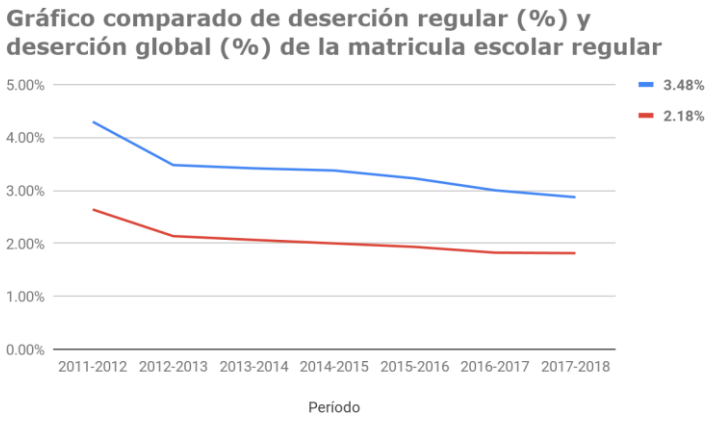

Ilustración 15: evolución de la deserción regular y global del sistema regular en los últimos años, elaboración propia (2018) – Fuente datos: MINEDUC

Lo que significas una trayectoria ligeramente descendente a lo largo de los últimos años, mostrando una baja de la barrera de los 3% en el último año respecto a la deserción regular.

También se evidencia que en el período 2017-2018 cerca de la mitad de les estudiantes que desertaron lo hicieron habiendo aprobado el año escolar (46,2%), un cuarto fue reprobado (27,1%) y el cuarto restante fue retirado (26,7%) lo que significa que por razones formales o informales no estaba en condiciones de ser evaluado.

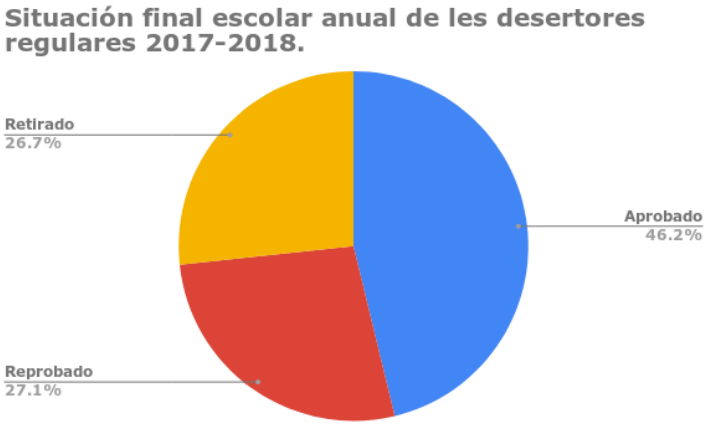

Ilustración 16: situación final escolar anual de les desertores regulares 2017-2018, elaboración propia (2018) – Fuente datos: MINEDUC

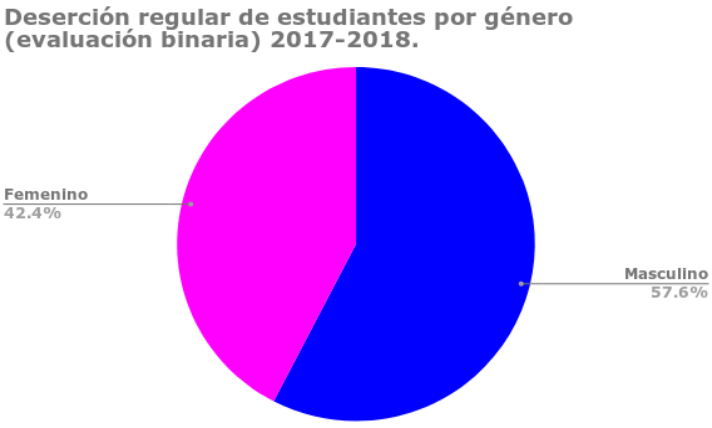

Ilustración 17: Deserción regular de estudiantes por género (evaluación binaria) 2017-2018, elaboración propia – Fuente datos: MINEDUC

En el análisis de género, limitado a una evaluación binaria debido a no existir un levantamiento que considere las distintas identidades de género, se evidencia que el género masculino sobrepasa al femenino en la deserción regular del período 2017-2018.

5.2.- Entendimiento de los datos

El entendimiento de los datos es un aspecto fundamental, ya que son la materia prima para la solución propuesta. A continuación, se describen los distintos niveles de este proceso.

En esta etapa se realizaron distintas diligencias con el fin de la obtención de la más amplia gama de datos, que se ajustaran a los requerimientos que la teoría de exclusión educativa plantea como relevantes, y que fueran disponibles transversalmente para toda la matrícula escolar. A continuación, se describen las fuentes y resultados de la obtención de los datos, como los datos faltantes y su efecto en la robustés del modelo.

5.2.1.- Datos obtenidos

- La fuente principal de datos proviene de los Datos Abiertos del Ministerio de Educación. Se recolectó información respecto a la descripción de les estudiantes según su: Matrícula (edad, género, etnia, etc), Rendimiento (asistencia anual, promedio, resultado), Asistencia mensual, SEP (Subvención escolar preferencial); de los Establecimientos educacionales: Establecimiento, Dotación docente, Dotación asistentes, Jornada Escolar Completa JECD, Matrícula por curso, Matrícula por unidad educativa, Matrícula por establecimiento, Rendimiento por unidad educativa, SEP (Subvención escolar preferencial), SNED (Sistema nacional de evaluación de desempeño), Subvenciones.
- De la Agencia de Calidad de la Educación se obtuvieron los datos por Establecimiento referentes a: Indicadores de desarrollo personal y social, Simce.
- De la JUNAEB se obtuvo los datos por Establecimiento: Índice de vulnerabilidad de establecimientos.
- Del CIAE se obtuvo los datos del Índice de Desarrollo Educacional IDE, el cual mide la trayectoria del desarrollo educacional de los establecimientos a lo largo de los años.

En total significan una muestra depurada de aproximada de 2.8 millones de estudiantes por año (2014-2017), donde consistentemente les estudiantes

excluidos del sistema educacional representan aproximadamente el 3% los casos de exclusión educativa anual (del orden de 80 mil estudiantes cada año).

En total convergen 20 fuentes de datos, de 4 instituciones distintas. Cada estudiante se caracteriza en términos iniciales (es decir, tras la primera selección, y sin considerar la codificación de las variables categóricas en "dummys") por 948 dimensiones, las cuales 187 son directamente datos del estudiante, y 761 caracterizan a nivel del establecimiento educacional. A continuación, se presenta la tabla de las fuentes de datos, y en ANEXO A se encuentra la tabla con todas las dimensiones a considerar. Los dato provenientes del MINEDUC pueden ser encontrados en http://datosabiertos.mineduc.cl/; de la JUNAEB en https://www.junaeb.cl/ive; la Agencia de Calidad de la Educación a la fecha revocó la disponibilidad de libre descarga; y los datos provenientes del CIAE tienen carácter confidencial, y fueron obtenidos mediante la firma de un acuerdo de confidencialidad.

N	Datos	Escala	Institución
1	Matrícula	Estudiante	MINEDUC
2	Rendimiento	Estudiante	MINEDUC
3	Asistencia mensual	Estudiante	MINEDUC
4	Subvención escolar preferencial SEP	Estudiante	MINEDUC
5	Establecimiento	Establecimiento	MINEDUC
6	Dotación docente	Establecimiento	MINEDUC
7	Dotación asistentes	Establecimiento	MINEDUC
8	JECD	Establecimiento	MINEDUC
9	Matrícula por curso	Estudiante	MINEDUC
10	Matrícula por unidad educativa	Establecimiento	MINEDUC
11	Matrícula por establecimiento	Establecimiento	MINEDUC
12	Rendimiento por unidad educativa	Establecimiento	MINEDUC
13	Subvención escolar preferencial SEP	Establecimiento	MINEDUC

14	Sistema nacional de evaluación de desempeño SNED	Establecimiento	MINEDUC
15	Subvenciones	Establecimiento	MINEDUC
16	Indicadores de desarrollo personal y social (ACE)	Establecimiento	Agencia de Calidad de la Educación
17	SIMCE	Establecimiento	Agencia de Calidad de la Educación
18	Índice de desarrollo personal y social (IDPS)	Establecimiento	Agencia de Calidad de la Educación
19	Índice de vulnerabilidad de establecimientos (IVE)	Establecimiento	JUNAEB
20	Índice de Desempeño Educacional (IDE)	Establecimiento	CIAE

Tabla 2: Fuentes de datos, elaboración propia (2018)

5.2.2.- Datos no incorporados

En este apartado se describen los datos que fueron considerados relevantes, pero por diversos motivos no fueron incorporados. Junto con describir el impacto de esto en el proyecto y la robustés del modelo.

Existen dos criterios que priman a la hora de no disponibilizar un dato para el modelo: no se encuentra disponible para la generalidad de la matrícula; o no fue autorizada su entrega por parte de la custodia correspondiente.

- Registro Social de Hogares

Este instrumento significa un gran potencial para una mejor caracterización de la realidad familiar, del hogar, del estudiante. Siendo una pieza clave para capturar parte importante de los factores de exclusión educativa pertenecientes al ámbito extraescolares.

Tras solicitar vía Ley de Transparencia al Ministerio de Desarrollo Social, se niega la obtención del Registro Social de Hogares por razones de privacidad

de los datos; y al no estar mandatados por la Ley al preprocesamiento de datos previa su entrega. Se adjunta en ANEXO B la respuesta completa.

Sin embargo, la disponibilidad general de datos alcanzada para el presente trabajo de memoria, descrita anteriormente, resulta suficientemente abundante para la construcción de un modelo que permita dar luces del fenómeno de exclusión educativa. Queda como recomendación y trabajo futuro la incorporación del Registro Social de Hogares al modelo predictivo.

- Información territorial CIAE

El Centro de Investigación Avanzada en Educación posee también un registro descriptivo de carácter territorial para les estudiantes, donde se detalla la georreferenciación habitacional de les estudiantes, su distancia al colegio, y caracterización sociodemográfica de su entorno. Estos datos aún no alcanzan la cobertura necesaria, a criterio del autor, para la incorporación al modelo, ya que una máxima en la construcción de éste es abarcar al sistema escolar regular en su conjunto, y no una muestra reducida para la que estén disponibles los datos.

- Estudiantes de establecimientos Particulares Pagados

En el proceso de exploración de los datos de asistencia para cada estudiante, se identifica que sólo se encuentra disponible para menos del 10% de los estudiantes de este tipo de dependencia del establecimiento educacional. Es por lo que, al considerarse una variable crítica, y ser corroborado posteriormente, se decide acotar el presente trabajo de memoria de título al Sistema Escolar Regular Público, Municipal y Particular Subvencionado.

Esto si bien significa una disminución en la cobertura del presente modelo, no atenta contra la robustés en cuanto a identificar el fenómeno de exclusión educativa.

5.3.- Desarrollo de la propuesta de solución por objetivos

A modo de exponer el desarrollo de las soluciones propuestas para el presente trabajo de memoria, es que a continuación se describe los procesos transversales para las soluciones, como la construcción de la base de datos y

la selección de características, para posteriormente plantear el desarrollo específico para cada objetivo planteado: la predicción de la potencial exclusión educativa escolar, la construcción y asignación de los perfiles de exclusión educativa, y el aporte al entendimiento del fenómeno de exclusión educativa a través de visualizaciones.

5.3.1.- Construcción de las bases de datos

Utilizando las bases de datos recién mencionadas, se construye una base consolidada para les estudiantes, y para cada establecimiento. Estas por separado debido a la alta dimensionalidad de ambas, y la alta cantidade datos en la de estudiantes por año. Generar una base única genera problemas que su solución escapan del alcance del presente trabajo de memoria, ya que involucra la puesta en marcha y utilización de sistemas de computación distribuida que permitan mayor acceso a la masividad de datos.

Para construir un modelo que considere la disponibilidad de los datos a la hora de predecir, es que se define que cada base de datos que describa un año en particular estará compuesta sólo con la información que a lo largo del tiempo reciente evidenciara disponibilidad con la antelación suficiente para ser usadas por el modelo. Lo que se traduce que para describir a les estudiantes se utilizarán sólo bases de datos desde al año anterior al que se describe, exceptuando la matrícula de cada año, la asistencia mes a mes hasta julio, y la información de la matrícula del curso de les estudiantes. Para el caso de los establecimientos, se constituyen de igual manera sólo con bases disponibilizadas durante el año anterior, o meses previos al inicio de clases, salvo el directorio de establecimientos para dicho año, que juega el mismo rol que la matrícula para les estudiantes.

A modo de construcción de nuevas características, agregadas al cruce de datos bruto obtenido, es que se agregan las siguientes características para cada estudiante:

1. Rendimiento anual general de los últimos 5 años, que comprende promedio general, asistencia general, situación final.
2. Trayectoria mes a mes de asistencia para los últimos 4 años, junto con la contabilización de establecimientos a los que asiste de forma efectiva le estudiante mes a mes.

3. Ranking relativo entre estudiantes por curso, colegio, provincia y región, respecto al promedio general y asistencia generla, para los últimos 4 años.
4. Sobre edad de cada estudiante, respecto al curso al que están asistiendo.
5. Promedio entre resultados SIMCE por nivel por establecimiento.

Para la caracterización de los modelos como de *alerta temprana*, se define que la predicción será situada temporalmente un semestre antes del cierre del año escolar. Esto porque se identifica la importancia de la asistencia mes a mes dentro del año para evaluar el decaimiento efectivo en la trayectoria de asistencia le estudiante. Se considera la asistencia hasta julio del mismo año dentro del conjunto de dimensiones a utilizar. Así la predicción podrá servir para gatillar intervenciones preventivas durante el año escolar.

Dentro de las dimensiones se identifican las naturalezas categóricas y numéricas de cada dimensión, las cuales son especificadas en el ANEXO A, donde también se identifican las dimensiones a utilizar previo a la selección de características. En total, luego de la codificación en dimensiones "dummys" de las dimensiones categóricas, se aumenta de 724 a 1188 dimensiones. Las cuales serán sometidas a un proceso de selección descrito a continuación.

5.3.2. - Selección de características

Para discriminar entre las dimensiones que aportan efectivamente al proceso de predicción de la exclusión educativa escolar, se acude a la técnica de selección de características basada en árboles de decisión. Esto ya que naturalmente los árboles de decisión encuentran, a través de la selección de particiones, las dimensiones que aportan a la decisión del modelo. En este contexto es que se selecciona al meta-algoritmo XGBoost en base de árboles de decisión, en el se define un umbrel de aporte de ganancia al modelo de 0.0001, lo cual reduce de 1188 dimensiones a 490, las cuales se presentan en el ANEXO C en su orden de relevancia para el model. Destaca la sobre edad, y la trayectoria de asistencia y promedio en los últimos años como dimensiones relevantes y que describen la trayectoria educative de le estudiante.

5.3.3.- Algoritmos de predicción de la exclusión educativa escolar

Para abordar la predicción de la exclusión educativa del sistema regular chileno, acotado al caso de matrícula pública, municipal y particular subvencionada; y definido el escenario temporal en que se llevará a cabo la predicción a mitad de año (asistencia hasta julio); se plantean los siguientes modelos:

1) Support Vector Classificator (Kernel polynomial)
2) Red Neuronal Artificial
3) RandomForest
4) AdaBoost (Árboles de decisión)
5) XGBoost (Árboles de decisión)

Para los Support Vector Classificator se utiliza una muestra aleatoria de entreamiento de 10mil muestras balanceadas al 50%-50% entre cada clase. Esto debido a que las SVMs escalan en complejidad según la cantidad de la muestra con que se les entrene. La muestra de datos fue imputada en sus miss-values con 0, debido a la naturaleza numérica de las dimensiones que lo requerían. Se utiliza la implementación SVC de la librería Scikit-Learn.

Para la Red Neuronal Artificial implementada corresponde a un Multi Layer Perceptrón (MLP) de arquitectura completamente conectada, dos capas ocultas con 8mil neuronas cada una, ambas con función de activación Relu, y drop-out de 0.2 en la primera capa oculta, con Softmax en la capa de salida, y función de pérdida de Entropía cruzada. Se utiliza una muestra aleatoria balanceada al 50% entre las dos clases de 150mil datos. La muestra de datos fue imputada en sus miss-values con 0, debido a la naturaleza numérica de las dimensiones que lo requerían. Se utiliza la librería TensorFlow para su implementación.

Para el modelo RandomForest se utiliza una muestra aleatoria balanceada al 50% entre las dos clases en 150mil muestras. La muestra de datos fue imputada en sus miss-values con 0, debido a la naturaleza numérica de las dimensiones que lo requerían. Se utiliza la implementación RandomForestClassifier de la librería Scikit-Learn.

El modelo AdaBoost es utilizado con árboles de decisión como estimadores débiles, se utiliza una muestra aleatoria balanceada al 50% entre las dos

clases en 150mil muestras. La muestra de datos fue imputada en sus miss-values con 0, debido a la naturaleza numérica de las dimensiones que lo requerían. Se utiliza la implementación AdaboostClassifier de la librería Scikit-Learn.

El Modelo XGBoost, o eXtreme Gradient Boosting, es utilizado con árboles de decisión como estimadores débiles, y se entrena con una muestra aleatoria balanceada al 50% entre las dos clases en 150mil muestras. Se utiliza la librería creada colaborativamente XGBoost.

A continuación, se presentan la tabla de desempeños para cada uno de los algoritmos, evaluados en el conjunto de validación de 100mil estudiantes con balance natural del fenómeno de exclusión educativa escolar (97% no excluides – 3% excluides).

Modelo	Exhaustividad	Precisión	Exactitud
XGBoost	82.167%	19.534%	89.311%
AdaBoost	76.800%	17.030%	88.291%
RandomForest	81.930%	14.550%	85.027%
SVC POLY	68.767%	12.596%	84.748%
Red Neuronal MLP	83.100%	8.500%	72.670%

Tabla 3: resultados validación modelos predictivos, elaboración propia (2018)

A continuación, se presentan la tabla de desempeño para el algoritmo XGBoost que presenta mejor desempeño en la validación, evaluado en el conjunto de test, disjunto con el de validación, de 100mil estudiantes con balance natural del fenómeno de exclusión educativa escolar (97% no excluides – 3% excluides).

Modelo	Exhaustividad	Precisión	Exactitud
XGBoost	81.500%	19.290%	89.215%

Tabla 4: resultado test modelo predictivo XGBoost, elaboración propia (2018)

5.3.4.- Perfiles de fenómenos de exclusión educativa

Sumado a predecir el potencial de exclusión educativa escolar para cada estudiante, se construyen perfiles de fenómenos de exclusión educativa con el objetivo de describir y caracterizar de mejor manera a cada potencial excluido. Esto con el objetivo de que, al aportar mayor información, las políticas y programas de prevención de la exclusión educativa puedan tener una aproximación más adecuada y contextualizada al estudiante, y así ser más afectivas.

Esto se desarrolla en su primer paso con un proceso de aprendizaje de máquinas no supervisado, en el cual se utiliza el modelo Gaussian Mixture Model, con matriz diagonal de varianza-covarianza para efectos de complejidad computacional, siendo también un estándar, para construir aglomeraciones que distribuyen cada una como una Normal Multivariada. Cada aglomeración posee un vector de medias y una matriz de varianza-covarianza, lo que los hacen característicos de una "personalidad" propia, idónea para la construcción de perfiles de personas.

Como la mayoría de los modelos de aglomeración, Gaussian Mixture Model requiere como parámetro el número de agrupaciones a construir. Para identificar la idoneidad de esta cantidad, es que se realiza el test de scores tanto calculando el AIC (Criterio de información de *Akaike)*, el BIC (Criterio de información bayesiano) y el Score (Logverosimilitud promedio de las muestras).

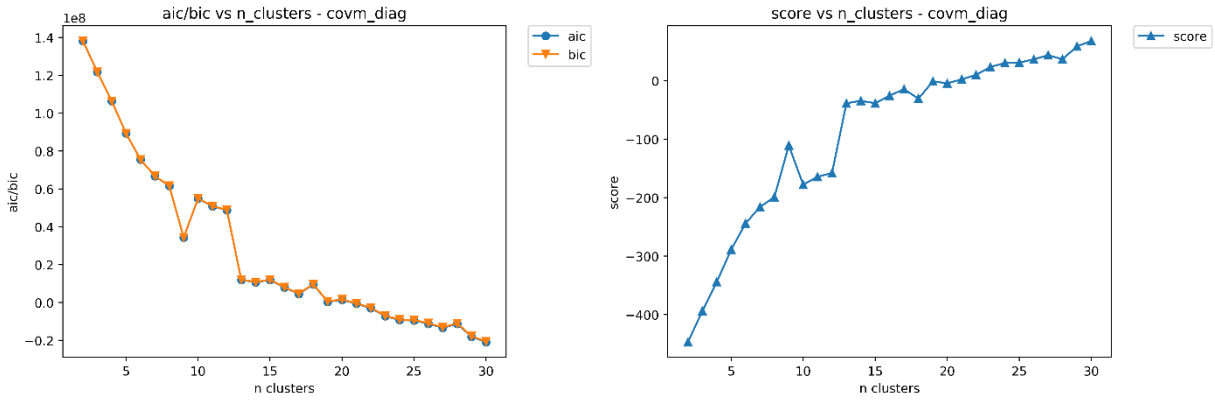

Ilustración 18: test de scores, elaboración propia (2018)

Estos tests evidencian en 9 agrupaciones un máximo local en la Logverosimilitud (mínimo local en AIC/BIC), lo que bajo un criterio de *trade-*

off entre idoneidad y complejidad interpretativa, es suficiente para seleccionarlas como una cantidad adecuada de agrupaciones.

Es relevante destacar que la interpretación de una cantidad *"correcta"* de agrupaciones tiende a ser carácter arbitrario. Para datos de naturaleza y calidad de presentación silmilares, los algoritmos de agrupación que fuerzan la construcción de aglomeraciones bajo un criterio específico, cambiando las estructuras interpretables. Es por lo que más que encontrar un número de agrupaciones correcto, es también relevante considerar la utilidad de la interpretación que se pueda obtener de estos.

Tras obtener las 9 agrupaciones, se analiza para cada una de ellas su matriz diagonal de varianza-covarianza para encontrar las dimensiones más estables para cada agrupación. Estas son las que sirven de base para la construcción de los perfiles de exclusión educativa.

Para esto es que se evalúa $Log(1/var(x))$ para $x \in \{\text{dimensiones}\}$. Donde $1/var(x)$ busca separar las dimensiones de varianzas bajas (las que menores a 1, y las más cercanas a cero, generan valores altos) y las con alta varianza generan valores bajos, como muestra la ilustración 22 que hace referencia a una agrupación. Se le aplica logaritmo para suavizar esta dispersión y lograr sentar el escenario en que se utilice un umbral arbitrario de selección. En este caso se selecciona el valor $log(1/var(x)) \geq 2$ como umbral de selección de variable crítica para la construcción del perfil de fenómenos de exclusión educativa, ya que se observa en el histograma de los datos recién descritos (ilustración 22) que desde ese punto comienza una cola alargada de dimensiones estables más allá del comportamiento normal correspondiente a las dimensiones más ruidosas. Este se presenta como patrón para todos los casos.

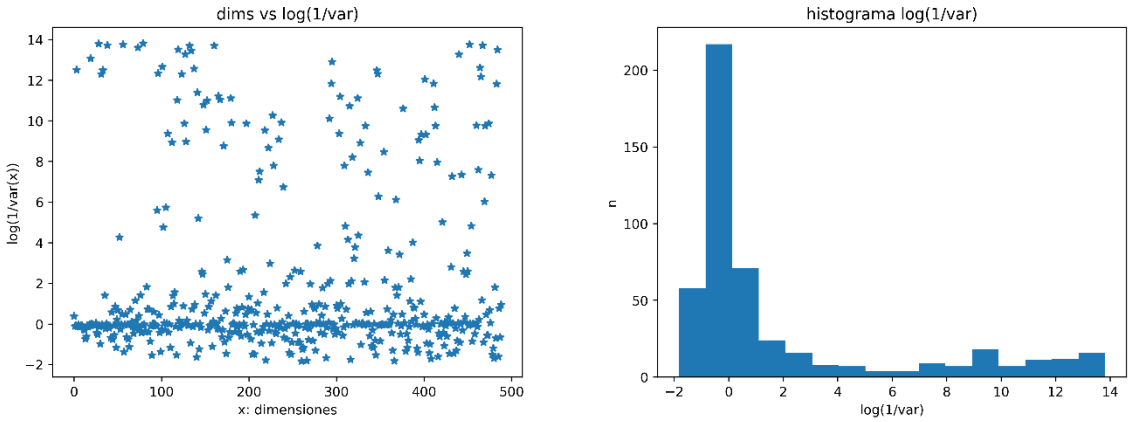

Ilustración 19: Análisis de representatividad dimensional de la agrupación H, elaboración propia (2018).

Posterior a la selección de las dimensiones críticas, se calcula la media y mediana de cada característica para cada agrupación, y se cruza con la importancia relativa de las características en el model (ANEXO C). Así, al considerar la media y mediana de cada dimensión crítica en cada agrupación, jerarquizada por la importancia en el modelo predictivo, analizando su relación con los otros componentes críticos de la agrupación, y su relación general con el fenómeno de la exclusión educativa, es que se construyen descripciones cualitativas de cada agrupación en base a la recién descrita información cuantitativa. Se utiliza la vinculación interdisciplinaria con las ciencias sociales a modo de metodología de interpretación, en específico se trabaja junto a Beatriz Hasbún, cientista social del Laboratorio de Gobierno para la validación metodológica para la descripción cualitativa de los perfiles de fenómenos de exclusión educativa.

A continuación, se presentan las descripciones de los 9 perfiles de fenómenos de exclusión educativa construidos:

Agrupación	Representación (%)	Descripción
A	13,03 %	Educacion media rural.
B	10,64 %	Primero o segundo básico, o sin trayectoria anterior hasta 5 años atrás. Establecimientos ni autónomo ni emergente, según la evaluación SEP.
C	18,93 %	Establecimientos con básicas pequeñas o inexistentes, simces básicas en la media y mediana general.
D	8,08 %	Establecimientos con internado en zona urbana, simces básica superiores al promedio (20 puntos).
E	17,41 %	Sobre edad de 1 año, establecimientos sin básica.
F	7,32 %	Establecimientos sin retiros en 1° grado.
G	3,23 %	Enseñanza media.
H	16,87 %	Promedio ficom más bajo, retiro de hombres y mujeres en 1° tendiente a cero y más bajo de lo normal.
I	4,49 %	Enseñanza media.

Tabla 5: Descripción fenómenos de exclusión educativa, elaboración propia (2018)

5.3.5.- Visualizaciones para portar a la comprensión del fenómeno

Como enfoque para intentar aportar a ampliar la comprensión del fenómeno de exclusión educativa es que se formulan preguntas que, aunque simples, aún no se han evidenciado en la literatura actual de la exclusión educativa escolar para el caso chileno. Esto también forma parte del proceso de comprensión general, tanto de los datos como del problema, siendo expresión de la naturaleza iterativa y bidireccional de la metodología utilizada.

En este contexto, aparte de las visualizaciones aportadas en las secciones previas del presente trabajo, es que los siguientes hallazgos destacan tanto por su pregunta antes no realizada al menos explícitamente, y la evidenciación de patrones importantes en sus respuestas. Estas son:

A. ¿Cuál es el último mes al que asisten les estudiantes que no se matriculan al año siguiente debiendo hacerlo?
B. ¿Cómo se comporta mes a mes la asistencia de les estudiantes que desertan en un mes en específico?

A continuación, se responden estas preguntas.

5.3.5.1.- Último mes en que asisten les excluides

Para responder la pregunta *"¿Cuál es el último mes al que asisten les estudiantes excluides?", s*e propone la visualización del comportamiento mes a mes de la exclusión educativa, identificando el último mes en que asisten les estudiantes excluides a clases a través de su registro de asistencia. Para esto es que genera la visualización para los años 2011 al 2016, donde cada ilustración contiene 3 filas y 8 columnas, siendo la primera fila correspondiente a la enseñanza básica, la segunda a la enseñanza media científico humanista, y la tercera la enseñanza técnico profesional y artística (cod_ense2= 2,5,7). Es por esto que sólo la primera fila utiliza las 8 columnas, ya que están ordenadas de 1ro básico a 8vo básico. Y la segunda y tercera están ordenadas de 1ro medio a 4to medio, ocupando sólo las primeras 4 columnas. Como resultado se obtiene el hallazgo de un patrón de comportamiento de la distribución de la exclusión educativa a lo largo del año escolar. Consistentemente cerca del 50% de les excluides educacionalmente asisten hasta diciembre a clases, cosa que pareciera contraintuitiva, y su análisis será profundizado con la siguiente pregunta.

Año 2011:

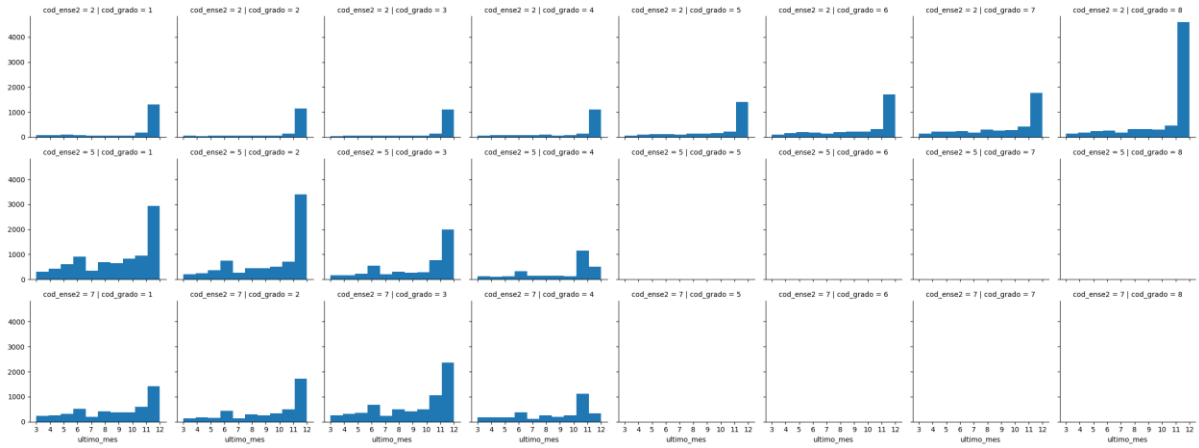

Ilustración 20: último mes de asistencia excluides educacionales regulares por grado y nivel de enseñanza 2011, elaboración propia (2018)

Año 2012:

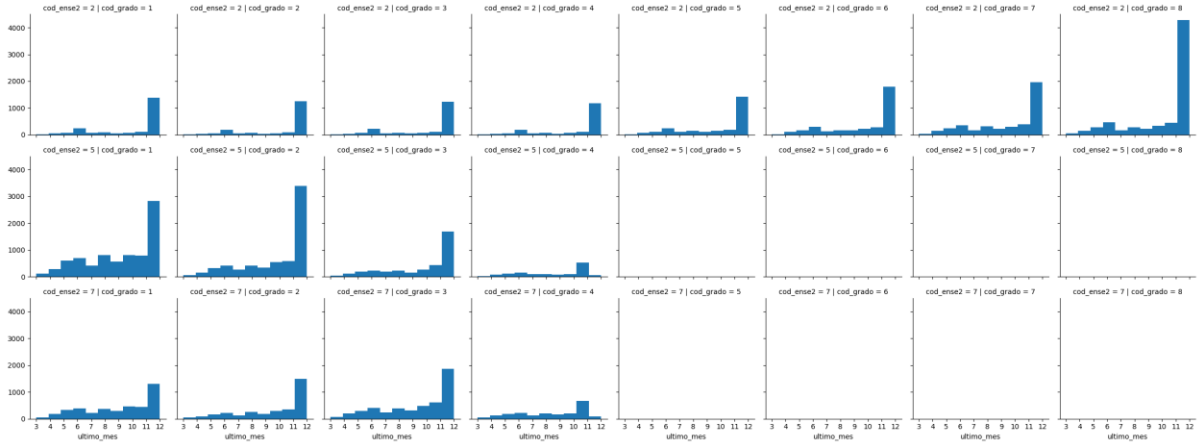

Ilustración 21: último mes de asistencia excluides educacionales regulares por grado y nivel de enseñanza 2012, elaboración propia (2018)

Año 2013:

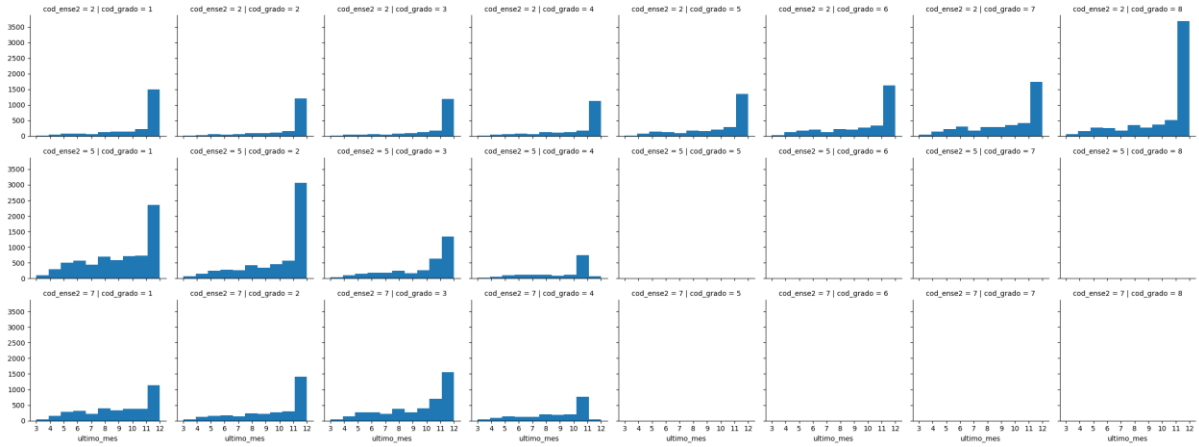

Ilustración 22: último mes de asistencia excluides educacionales regulares por grado y nivel de enseñanza 2013, elaboración propia (2018)

Año 2014:

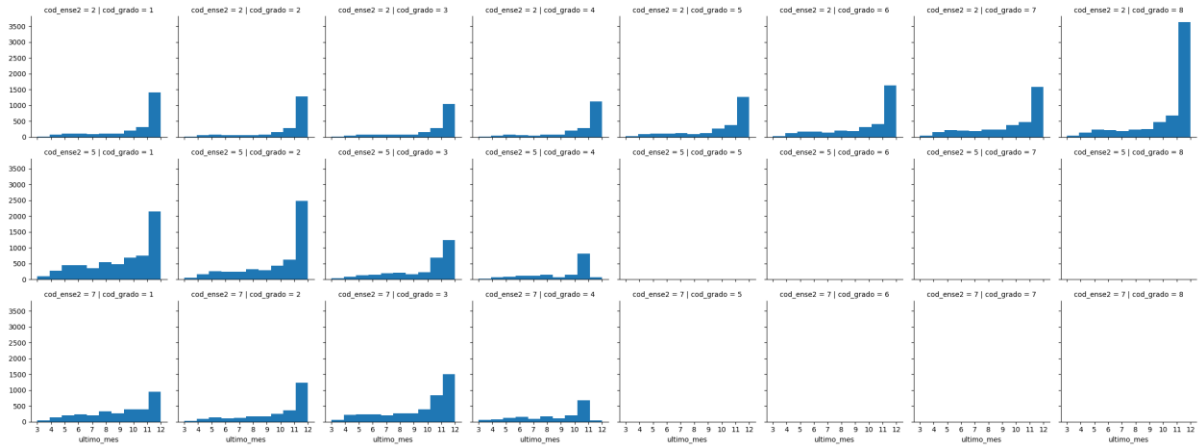

Ilustración 23: último mes de asistencia excluides educacionales regulares por grado y nivel de enseñanza 2014, elaboración propia (2018)

Año 2015:

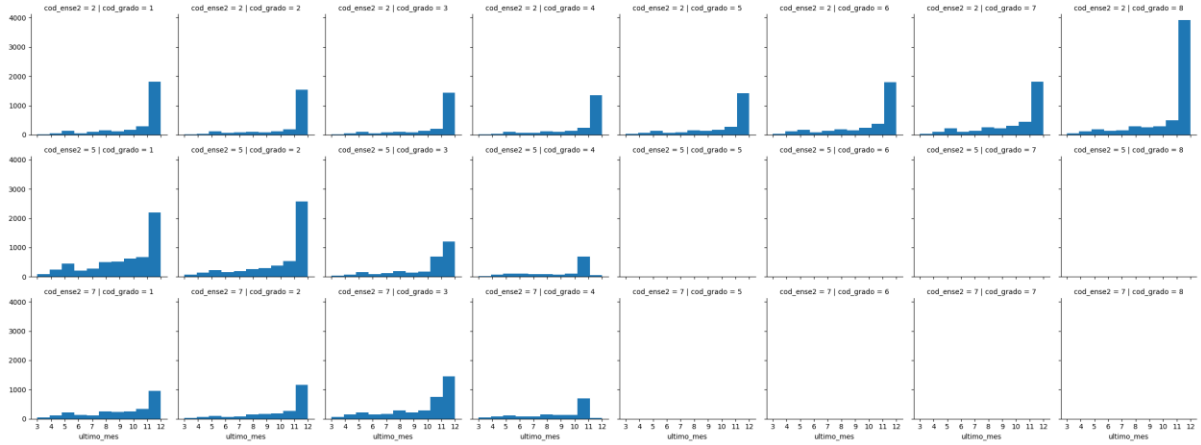

Ilustración 24: último mes de asistencia excluides educacionales regulares por grado y nivel de enseñanza 2015, elaboración propia (2018)

Año 2016:

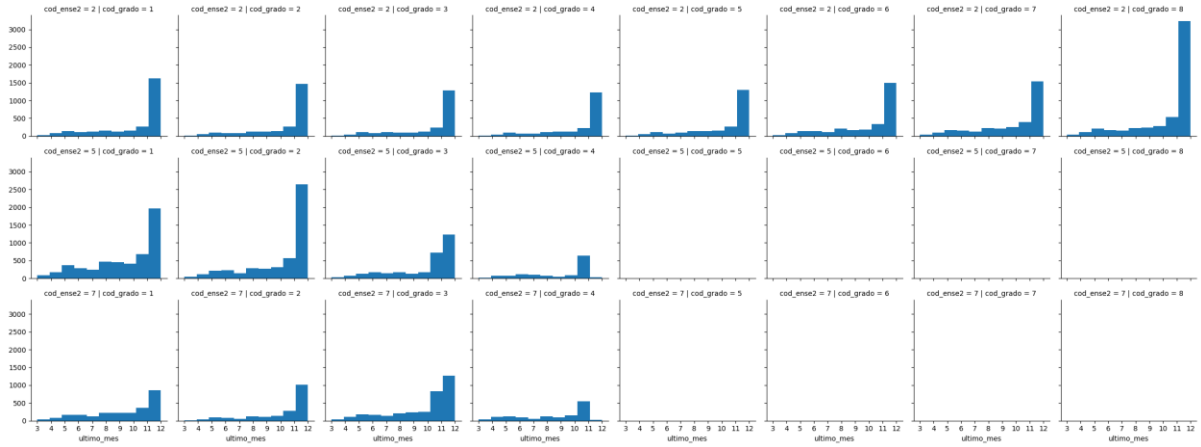

Ilustración 25: último mes de asistencia excluides educacionales regulares por grado y nivel de enseñanza 2016, elaboración propia (2018)

5.3.5.2.- Promedio asistencia mes a mes

Para analizar el comportamiento de la exclusión educativa a nivel mensual, es que se grafica el comportamiento de la media y desviación estándar de la asistencia mes a mes de les estudiantes excluides que asisten a clases hasta cierto mes en específico. En otras palabras, se aisló a cada grupo de excluides por mes en que asisten por última vez a clases, y se graficó su asistencia durante el año. Como hallazgo se obtiene que el comportamiento de les estudiantes excluides que asisten hasta diciembre a clases (del orden del 50% del total) lo hacen de manera regular, en los márgenes de un estudiante que aprueba por asistencia de forma normal: sobre el 80% de asistencia en promedio mes a mes en la educación básica, y sobre 80% general en la media.

2011:

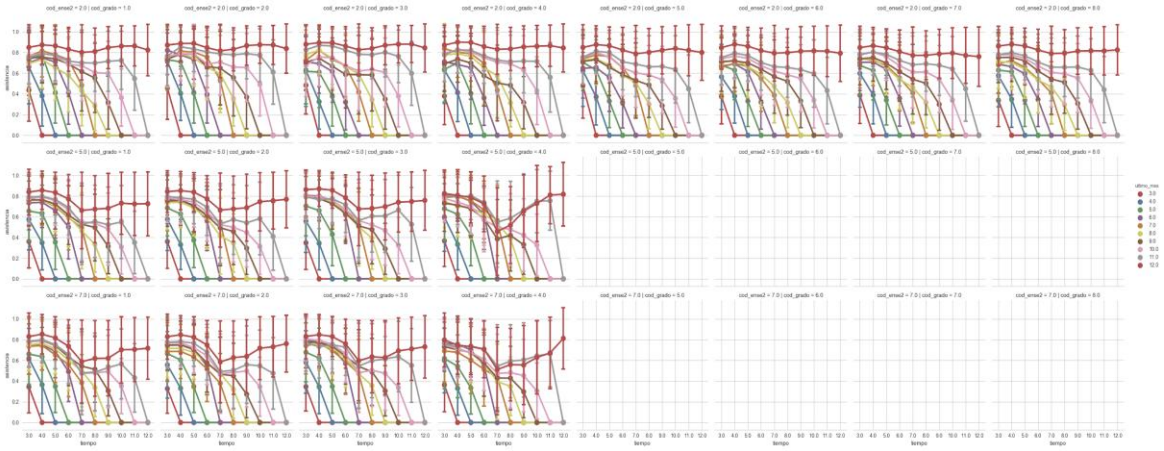

Ilustración 26: asistencia mes a mes excluides educacionales regulares por grado y nivel de enseñanza 2011, elaboración propia (2018)

2012:

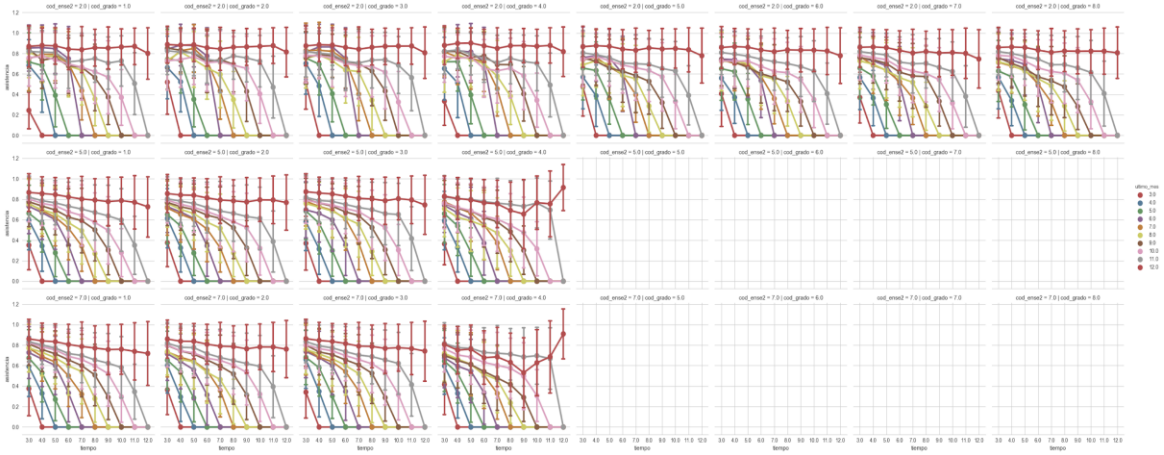

Ilustración 27: asistencia mes a mes excluides educacionales regulares por grado y nivel de enseñanza 2012, elaboración propia (2018)

2013:

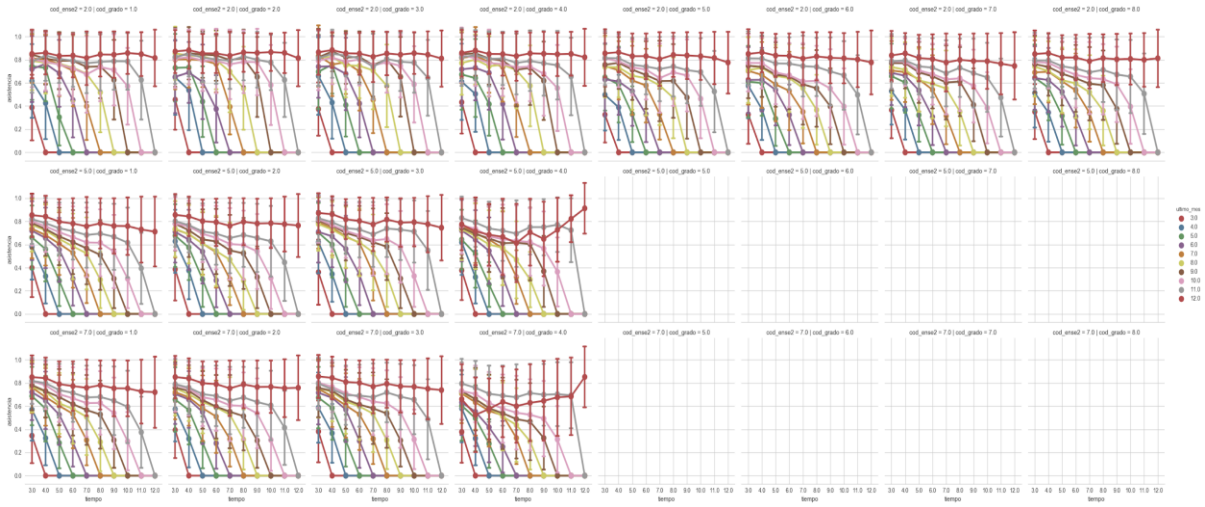

Ilustración 28: asistencia mes a mes excluides educacionales regulares por grado y nivel de enseñanza 2013, elaboración propia (2018)

2014:

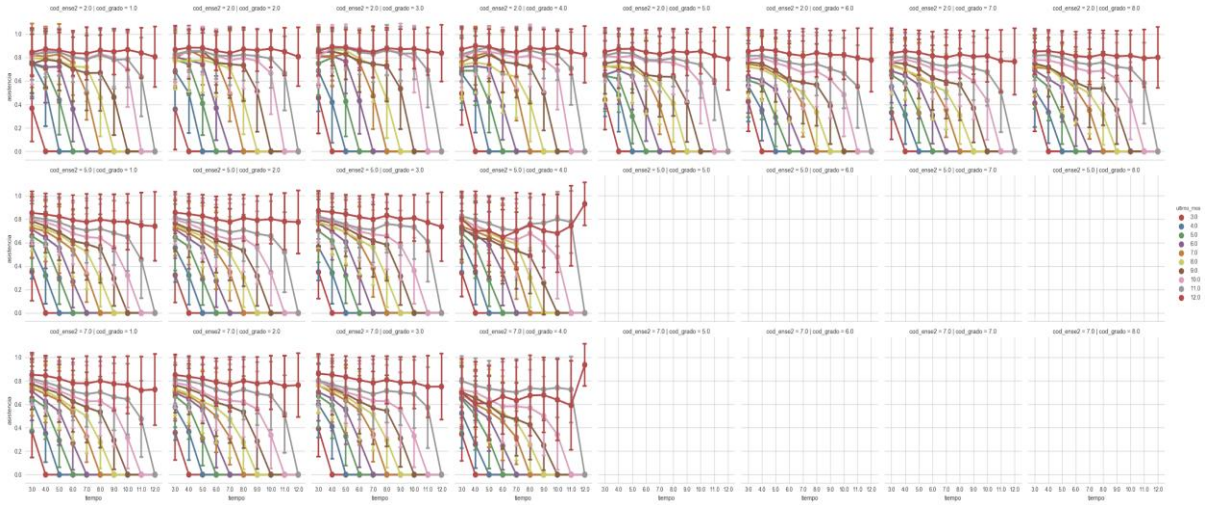

Ilustración 29: asistencia mes a mes excluides educacionales regulares por grado y nivel de enseñanza 2014, elaboración propia (2018)

2015:

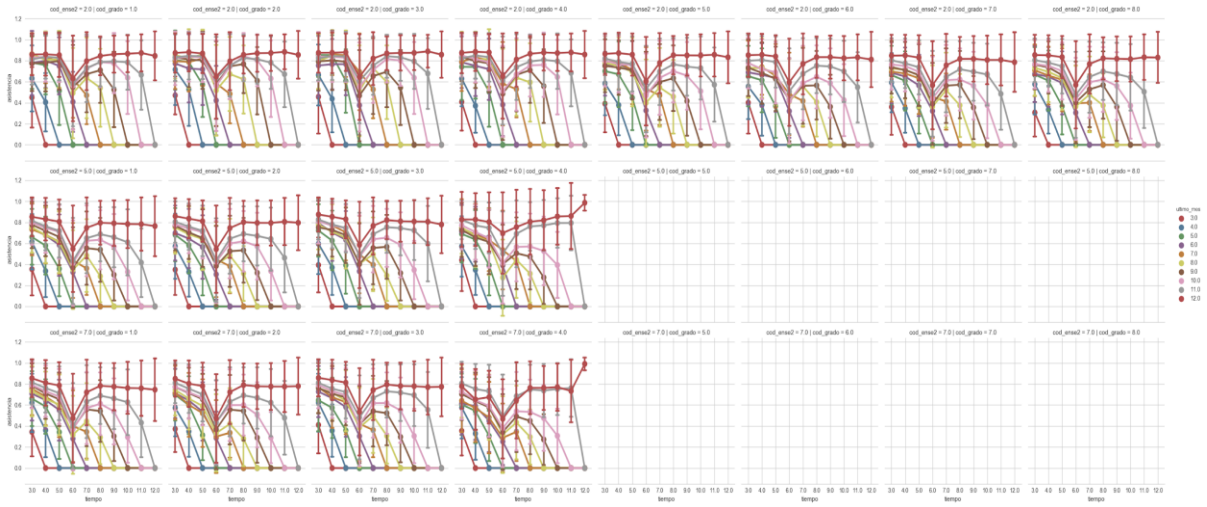

Ilustración 30: asistencia mes a mes excluides educacionales regulares por grado y nivel de enseñanza 2015, elaboración propia (2018)

2016:

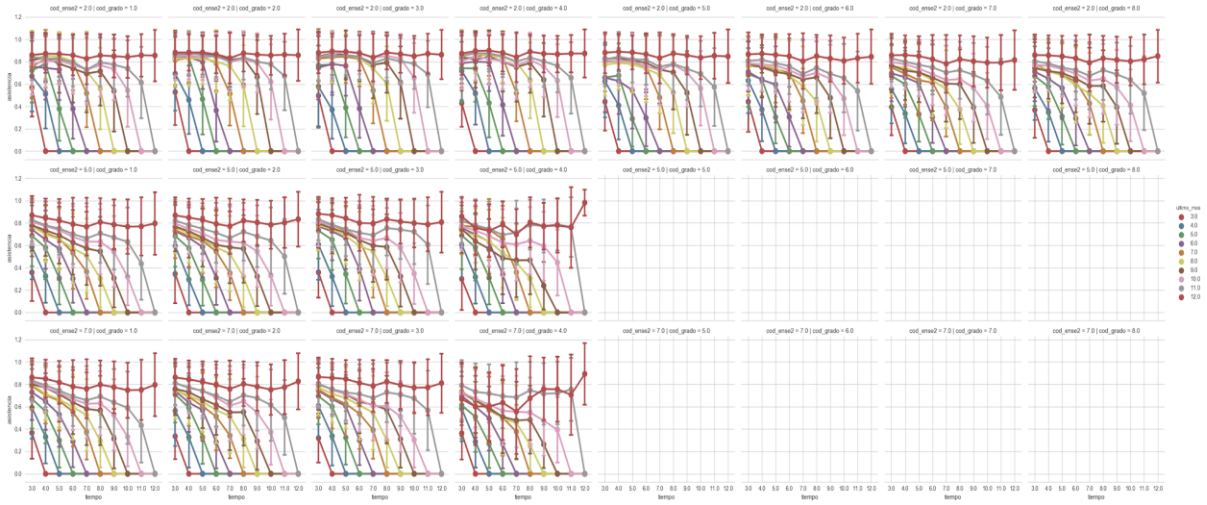

Ilustración 31: asistencia mes a mes excluides educacionales regulares por grado y nivel de enseñanza 2016, elaboración propia (2018)

6.- Análisis de resultado

En el presente capítulo se analiza los resultados obtenidos en el trabajo de memoria en sus tres dimensiones: predicción de exclusión educativa, perfiles de fenómenos de exclusión educativa, y la visualización de nuevas evidencias que ayuden a aumentar el entendimiento respecto a la exclusión educativa escolar para el caso chileno.

6.1.- Predicción de la exclusión educativa escolar

El resultado que mejor equilibra la exhaustividad (cobertura del fenómeno de exclusión educativa) y precisión, ya que logra reducir el problema de detectar a excluides educacionales de 1 en 33 (3%) a uno de 1 en 5 (19.290%), cubriendo el 81.500% del fenómeno (exhaustividad o recall), es el del modelo XGBoost.

Si bien se logra generar poder de predicción desde el entrenamiento en años escolares pasados hacia futuros, y desde una perspectiva temprana, aún existe un margen de aumento de la precisión. Mayor cantidad de información del estudiante, como el Registro Social de Hogares, permitirían al modelo crecer en efectividad a la hora de diferenciar entre estudiantes que se aproximan a trayectorias de exclusión educativa escolar con los que no. Esto basado en la importancia que la literatura referente a la exclusión educativa escolar le atañe a factores extraescolares, que son los más débiles en el modelo.

Al observar las características más relevantes que aportan al modelo XGBoost, se destaca la importancia de la sobre edad como fuerte factor de influencia para la decisión del modelo, la cual también podría ser asociada a la migración al sistema para adultos o el ingreso prematuro a la vida laboral por parte de le estudiante que deserta; así también destaca la relevancia de la trayectoria educativa de los últimos años, descrita por dimensiones como la asistencia mes a mes, el cambio de establecimiento dentro de los meses del año, el promedio general, la asistencia general, la situación final de le estudiante, su ranking relativo a diferente escala, entre otras, que se presentan para sucesivos años anteriores.

6.2.- Perfiles de fenómenos de exclusión educativa escolar

La construcción matemática de las aglomeraciones de estudiantes excluidos educacionalmente consta de dos componentes críticos que requieren análisis de resultados para el avance de la construcción de perfiles de fenómenos de exclusión educativa: la selección de la cantidad de aglomeraciones del modelo propuesto, y la selección de las dimensiones críticas que caracterizan a la aglomeración.

6.2.1.- Selección de la cantidad de aglomeraciones para el modelo propuesto

En este proceso de selección se utiliza el test de scores (*Ilustración 22*), para determinar un número adecuado de aglomeraciones a construir por el modelo. Mientras mayor score (Logversimilitud) o menor AIC (Criterio de información de Akaike) o menor BIC (Criterio de información Bayesiana) mayor es la idoneidad del modelo. Es relevante destacar que tanto el AIC y BIC utilizan en su cálculo la Logverosimilitud, por lo que es consistente la asociación que se puede hacer entre uno y otro. En particular AIC y BIC penalizan por la cantidad de parámetros que el modelo requiere, como medidas que permitan combatir el sobreajuste.

Es en este escenario que destaca el número 9 de agrupaciones, ya que representa un máximo local para la Logverosimilitud, y un máximo local para AIC/BIC. A esto es importante agregar que no es correcto pensar que siempre existe solo una forma o cantidad adecuada de generar agrupaciones, ya que lo que varía en cada uno de los casos (número de agrupaciones) es la estructura misma de las aglomeraciones, osea también su interpretación. Por lo que no es inadecuado plantear que 9 es un buen número de aglomeraciones independientemente de que existen mejores resultados de score, ya que debe existir un trade-off entre idoneidad matemática y complejidad de caracterización cualitativa. Ya que, en otras palabras, es más práctico describir 9 aglomeraciones y entender sus relaciones, que describir 200 aglomeraciones y perder las nociones diferenciadoras sustantivas que motivan la misma generación de estos perfiles: mejorar el escenario de intervención preventiva concreta.

6.2.2.- Selección de dimensiones críticas para la construcción de perfiles

Para la selección de las dimensiones críticas se propone un análisis de estabilidad de las dimensiones que componen la aglomeración. Para esto es que se genera una transformación que permita identificar de mejor manera los criterios de discriminación. Como se muestra en la *Ilustración 23,* la transformación elegida es el logaritmo de la inversa de la varianza, así mientras mayor sea la varianza, su imagen será menor, y mayor la varianza se alejará en dirección positiva. Al evaluar que efectivamente se genera una dispersión que muestra dimensiones con alta y poca varianza, es que también se genera un histograma para la imagen construida de la varianza de cada dimensión, mostrando que existen dimensiones que se comportan de manera "ruidosa" en forma normal, y dimensiones que se escapan en su estabilidad. Así se define arbitrariamente el umbral de $log(1/var(x)) \geq 2$, definiendo a las dimensiones que lo satisfagan como las variables críticas que definen la aglomeración debido a su carácter de estable dentro de la misma.

6.2.3.- Descripción cualitativa de los perfiles de fenómeno de exclusión educativa

Tras analizar la jerarquía de características críticas relevantes para el problema, en cada una de las 9 agrupaciones construidas, es que se logra construir perfiles cualitativos para cada subfenómeno de exclusión educativa. Estos resultan ser significativamente distinguibles, y permiten en la mayoría de los casos una interpretación que aporte mayor contexto al fenómeno de exclusión educativa. Estos sí no son suficientes para entregar una contextualización mayor, ya que al no contar con una descripción acabada por estudiante de su contexto familiar y de entorno, es que las aglomeraciones son construidas en mayor medida por descripciones transversales o generales (a nivel establecimiento educacional, comunal o regional) más que por las características propias del sujeto. Esto refuerza la necesidad de incorporar el Registro Social de Hogares como insumo de información contextual familiar.

6.3.- Visualizaciones: la asistencia de les estudiantes excluides educacionales durante el año escolar

A continuación, se presentan las nuevas evidencias encontradas gracias a la visualización de datos, acerca del fenómeno de exclusión educativa escolar chilena; junto con preguntas y reflexiones que estas pudieran despertar.

Al analizar los resultados obtenidos con las visualizaciones del último mes en que asisten les estudiantes, y la del comportamiento de asistencia promedio mes a mes de los grupos de excluides educacionales según su último mes de asistencia, se evidencia un patrón común año a año para el comportamiento de exclusión educativa durante el año escolar. ***Sobre el 50% de les estudiantes excluides educacionales asisten hasta diciembre a clases incurriendo en un régimen regular de asistencia, o suficiente para aprobar: 80% promedio mes a mes en la enseñanza básica y sobre el 80% general en la media (destacando una mayor irregularidad en 4to Medio)*.** Lo que plantea una realidad no evidenciada en la literatura, y que se propone como insumo a la base factual para procesos de innovación y diseño de políticas públicas para la prevención de la exclusión educativa.

A juicio del autor, gracias a lo recién descrito, esto hace nacer preguntas como: ¿Hace sentido que el proceso de cobertura institucional sea discontinuado durante el verano para les estudiantes en riesgo de exclusión educativa?, ¿Pueden ser prevenidos los factores críticos que hacen propender a los meses de enero y febrero como los de mayor escape del sistema escolar?

También esta evidencia se presenta como posible explicación de la limitación que condiciona a los sistemas de alerta temprana arbitrarios. Estos analizan sólo el comportamiento aislado de la intermitencia de asistencia en perspectiva presente, no pudiendo explotar los patrones de alta complejidad que los algoritmos de aprendizaje de máquinas sí, no alcanzando la capacidad de identificar a más del 50% de les estudiantes excluides que no presentan comportamiento intermitente en la asistencia durante el año.

7.- Conclusiones y recomendaciones

Los resultados del presente trabajo de memoria logran permitir concluir que la predicción de la exclusión educativa escolar puede ser realizada, con cierta precisión y exhaustividad, a través del entrenamiento de algoritmos de aprendizaje de máquinas usando datos que en su mayoría describen ciclos educativos previos al evaluado, y disponibles en el momento de hacer la predicción.

También se concluye que es posible generar un sistema de alerta temprana, basado en aprendizaje de máquinas, que con al menos un semestre de margen de espacio a la acción institucional y comunitaria respectiva para la prevención efectiva de la exclusión educativa escolar de les estudiante.

Respecto a los perfiles de fenómenos de exclusión educativa, se concluye que matemáticamente es posible construir perfiles con interpretabilidad cualitativa, donde las dimensiones con menor variabilidad evaluadas en sus medias y medianas constituyen un elemento crítico para la caracterización. También se concluye que los perfiles de exclusión educativa alcanzan características diferenciadoras, pero debido a la falta de características descriptivas del entorno personal de les estudiante, es que sólo se alcanza un aporte superficial de información.

A modo de recomendación, se hace evidente la necesaria incorporación del Registro Social de Hogares para aumentar el nivel de caracterización personal, familiar y del entorno de les estudiantes. Esto en base a la amplia referencia dentro de la literatura respecto a la importancia de los factores extraescolares en la causalidad de la exclusión educativa. El autor considera que la incorporación del RSH aportaría mayor precisión en el modelo, aportando mayores antecedentes para que se pueda discriminar de mejor manera entre estudiante en trayectorias de exclusión educativa y los que no. También aportaría a la exhaustividad, ya que involucra factores que son nula o débilmente expresados en el actual modelo.

8.- Trabajos futuros

A continuación, se describen las principales propuestas de trabajos futuros, tanto en el tema particular del desarrollo algorítmico, como de la apertura general al paradigma de la ciencia de datos en temáticas de educación o política pública.

- Ensamble de modelos

Como evolución del modelo propuesto, se propone la construcción de un ensamble de algoritmos que integre una nueva capa de clasificación a la cual se ensamblen algoritmos individuales diferentes que predigan la exclusión educativa. Esto con el objetivo de que el nuevo clasificador, al ser entrenado con los resultados de los clasificadores individuales y las etiquetas originales de entrenamiento, se ajuste a la función que modela la discrepancia entre los algoritmos, y así mejorar el poder de predicción. Esto se puede entender como un clasificador que sepa cuándo creerle a cada algoritmo dependiendo de lo que dicen los demás. La *ilustración 32* representa el esquema propuesto.

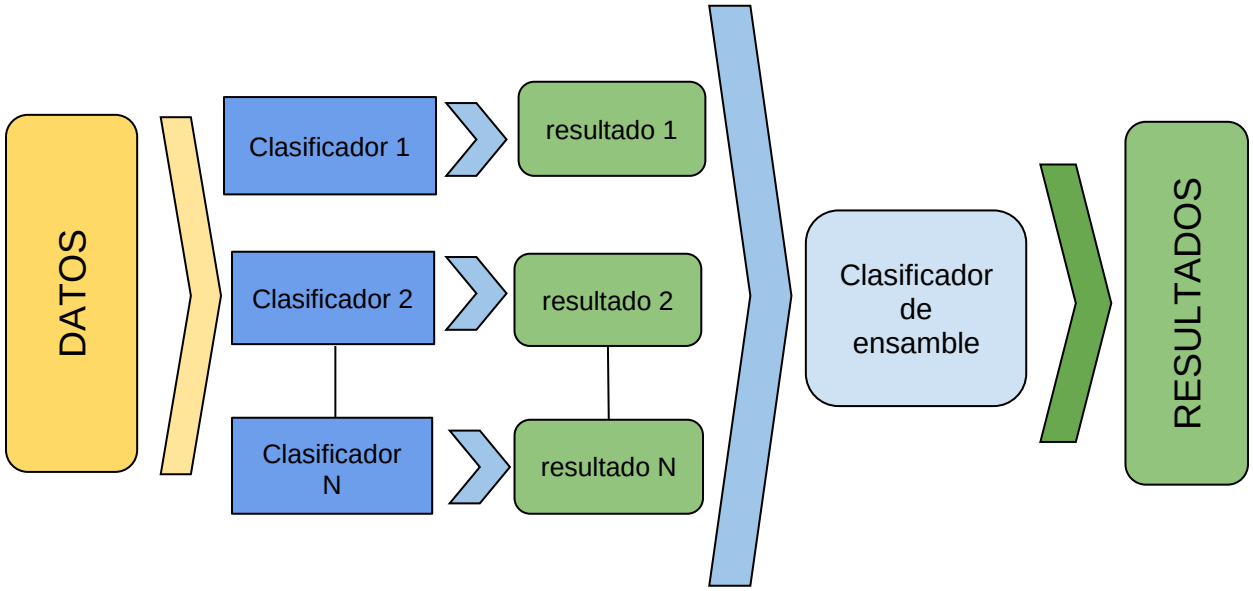

Ilustración 32: Propuesta de trabajo futuro (ensamble), elaboración propia (2018)

- Múltiples evaluaciones a lo largo del año escolar:

Como otra arista de trabajo futuros, se propone la cobertura del año escolar por distintos modelos, lo que permitan ir evaluando la trayectoria en que se

desarrolla el potencial de exclusión educativa, permitiendo eventualmente detectar de manera temprana, intervenir con enfoque de prevención, y evaluar los efectos a lo largo del año según la variación que pueda tener el potencial de exclusión educativa en le estudiante. Se recomienda generar modelos para inicio de año (sin asistencia, o sólo meses iniciales), y finalizado el año escolar, a modo de evaluación y ratificación de intervenciones para con les estudiantes.

- Nuevas herramientas para una nueva aula de clases:

La incorporación la ciencia de datos al proceso educativo significa un potencial de transformación paradigmático para el concepto de sala de clases clásico: Instituciones y profesores con mayores herramientas para el diagnóstico, implementación, y evaluación de prácticas educativas; detección temprana de factores de riesgo o fenómenos negativos; incorporación de plataformas que permitan una educación contextualizada a las capacidades y habilidades de les estudiantes; recaracterización del proceso educativo homogéneo o estático; son sólo uno de los pocos ejemplos que abre esta nueva área.

9.- Bibliografía

Baeza, J., y Fuentes, R. (2004). Antecedentes y fundamentos de las políticas de gestión y administración en el sistema educativo chileno 1980-2003. Santiago: Mineduc

Barro, R. (1991) "Economic Growth in a Cross Section of Countries". The Quarterly Journal of Economics, Vol.106, No. 2 (May, 1991), pp. 407-443.

Capacidades para Innovar. (s. f.). Recuperado 11 de abril de 2017, a partir de http://www.lab.gob.cl/capacidades-para-innovar/

CETIUC. (2017). Caracterización de la Interoperabilidad en el Estado de Chile, Informe N° 3. CETIUC.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS.

CIAE - Universidad de Chile. (2018). Recuperado 6 de julio del 2018, a partir de http://www.ciae.uchile.cl/

CIAE. (2018). Proyecto del CIAE estimará el riesgo de deserción del sistema educacional y sistematizará experiencias para prevenirlo. Recuperado 13 de abril de 2018, a partir de http://www.ciae.uchile.cl/index.php?page=view_noticias&langSite=es&id=1330

Corporación de Fomento de la Producción. (2014). EJECUTA ACUERDO DE CONSEJO N° 2.826, DE 2014, CREA "COMITÉ DE INNOVACIÓN EN EL SECTOR PÚBLICO" Y APRUEBA TEXTO DEL REGLAMENTO POR EL QUE DEBERÁ REGIRSE. [PDF file]. Santiago, CL: Secretaria General. Recuperado de http://www.agendaproductividad.cl/wp-content/uploads/sites/22/2014/10/50-RA-CORFO-2014-toma-razon_3-6-19-2014.pdf

El Lab. (s. f.). Recuperado 10 de abril de 2017, a partir de http://www.lab.gob.cl/el-lab/

Fundación San Carlos de Maipo. (2017). Uno de cada dos reos pasó su infancia o adolescencia en un centro del Sename | Emol.com. Emol. Retrieved 27 November 2017, from http://www.emol.com/noticias/Nacional/2017/03/20/850222/Uno-de-cada-dos-reos-paso-su-infancia-o-adolescencia-en-un-centro-del-Sename.html

Gestión de Ecosistemas e Inversiones. (s. f.). Recuperado 11 de abril de 2017, a partir de http://www.lab.gob.cl/gestion-de-ecosistemas-e-inversiones/

González, R. (2017). Las consecuencias de (no) completar la educación media para la población adulta en Chile Hallazgos a partir de la Evaluación Internacional de Competencias en Población Adulta PIAAC-OECD - DOCUMENTODE TRABAJO N°7. Centro de estudios MINEDUC.

Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm. Microsoft Research.

JUNAEB. (2018b). Beca de Apoyo a la Retención Escolar – Postulación Educación Media. Recuperado 13 de abril de 2018, a partir de https://www.junaeb.cl/becas-educacion-media/beca-de-apoyo-a-la-retencion-escolar-postulacion

JUNAEB. (2018b). PROGRAMA DE APOYO A LA RETENCIÓN ESCOLAR. Recuperado 13 de abril de 2018, a partir de https://www.junaeb.cl/programa-de-apoyo-a-la-retencion-escolar
Ley N° 18525. Diario Oficial de la República de Chile, Santiago, Chile, 29 de mayo de 2003.

LEY-19628 28-AGO-1999 MINISTERIO SECRETARÍA GENERAL DE LA PRESIDENCIA - Ley Chile - Biblioteca del Congreso Nacional. (2012). Ley Chile - Biblioteca del Congreso Nacional. Retrieved 27 November 2017, from https://www.leychile.cl/Navegar?idNorma=141599

LEY-20285 20-AGO-2008 MINISTERIO SECRETARÍA GENERAL DE LA PRESIDENCIA - Ley Chile - Biblioteca del Congreso Nacional. (2016). Ley Chile - Biblioteca del Congreso Nacional. Retrieved 27 November 2017, from https://www.leychile.cl/Navegar?idNorma=276363

Marshall, T. (2003). Algunos factores que explican la deserción temprana. En Seminario. Internacional "Abriendo Calles". Santiago, CONACE-SENAME.
MINEDUC. (2013). Serie Evidencias: Medición de la deserción escolar en Chile.

MINEDUC. (2018). Resumen Estadístico de la Educación 2017. Recuperado 12 de abril de 2018, a partir de https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2018/04/Resumen-Estad%C3%ADstico-de-la-Educaci%C3%B3n.-A%C3%B1o-2017.pdf

Ministerio de Economía y Turismo. (2017). BALANCE DE GESTIÓN INTEGRAL AÑO 2016. Economia.gob.cl. Retrieved 27 November 2017, from http://www.economia.gob.cl/wp-content/uploads/2017/04/3-BGI-2016-CORFO.pdf

Ministerio Secretaría General de la Presidencia. (2017). Consulta Pública: Norma de Interoperabilidad. Modernización y Gobierno Digital. Retrieved 27 November 2017, from http://www.modernizacion.gob.cl/es/noticias/consulta-publica-norma-de-interoperabilidad/

Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. London: The MIT Press.

SENAME. (2013). Estadística de los niños(as) y adolecentes vigentes en la red SENAME para enfoque de género.

SPSS FREE. Escalas De Medida: Nominal Ordinal Intervalo y Razón - Curso de Spss Gratis. Spssfree.com. Retrieved 27 November 2017, from http://www.spssfree.com/curso-de-spss/analisis-descriptivo/escalas-de-medida.html

Tele13. (2017). Los ejes del proyecto de ley que busca proteger los datos personales. https://www.facebook.com/teletrece. Retrieved 27 November 2017, from http://www.t13.cl/noticia/nacional/Presidenta-Bachelet-firma-proyecto-de-ley-que-regula-tratamiento-de-datos-personales

Vega, & Grau. (2016). Tesis: Impacto del Programa "Aquí, Presente" en la deserción escolar. Santiago: Universidad de Chile.

Anexos

ANEXO A

A continuación, se declaran todas las dimensiones en un principio evaluadas por el modelo. Donde tipo ‘n’: numéricas, y ‘c’: categóricas.

Dimensión	**tipo**
curso_n_alu_si	n
curso_n_alu_hom	n
curso_n_alu_muj	n
curso_n_alu	n
ranking_prom_gral_curso_1	n
ranking_asis_gral_curso_1	n
ranking_prom_gral_curso_2	n
ranking_asis_gral_curso_2	n
ranking_prom_gral_curso_3	n
ranking_asis_gral_curso_3	n
ranking_prom_gral_curso_4	n
ranking_asis_gral_curso_4	n
ranking_depe2_prom_gral_grado_rbd_1	n
ranking_depe2_asis_gral_grado_rbd_1	n
ranking_depe2_prom_gral_grado_rbd_2	n
ranking_depe2_asis_gral_grado_rbd_2	n
ranking_depe2_prom_gral_grado_rbd_3	n
ranking_depe2_asis_gral_grado_rbd_3	n
ranking_depe2_prom_gral_grado_rbd_4	n
ranking_depe2_asis_gral_grado_rbd_4	n
ranking_depe2_prom_gral_grado_com_1	n
ranking_depe2_asis_gral_grado_com_1	n
ranking_depe2_prom_gral_grado_com_2	n
ranking_depe2_asis_gral_grado_com_2	n
ranking_depe2_prom_gral_grado_com_3	n
ranking_depe2_asis_gral_grado_com_3	n
ranking_depe2_prom_gral_grado_com_4	n
ranking_depe2_asis_gral_grado_com_4	n
ranking_depe2_prom_gral_grado_pro_1	n
ranking_depe2_asis_gral_grado_pro_1	n
ranking_depe2_prom_gral_grado_pro_2	n
ranking_depe2_asis_gral_grado_pro_2	n
ranking_depe2_prom_gral_grado_pro_3	n
ranking_depe2_asis_gral_grado_pro_3	n

ranking_depe2_prom_gral_grado_pro_4	n
ranking_depe2_asis_gral_grado_pro_4	n
ranking_depe2_prom_gral_grado_reg_1	n
ranking_depe2_asis_gral_grado_reg_1	n
ranking_depe2_prom_gral_grado_reg_2	n
ranking_depe2_asis_gral_grado_reg_2	n
ranking_depe2_prom_gral_grado_reg_3	n
ranking_depe2_asis_gral_grado_reg_3	n
ranking_depe2_prom_gral_grado_reg_4	n
ranking_depe2_asis_gral_grado_reg_4	n
cod_depe	c
cod_depe2	c
estado_estab	c
cod_ense	c
cod_ense2	c
cod_ense3	c
cod_grado	c
cod_grado2	c
cod_jor	c
cod_tip_cur	c
gen_alu	c
cod_etnia_alu	c
int_alu	c
cod_int_alu	c
cod_nac_alu	c
pais_origen_alu	c
cod_sec	c
cod_espe	c
cod_rama	c
ens	c
marzo_asistencia_1	n
marzo_n_inst_efectivo_1	n
abril_n_inst_efectivo_1	n
mayo_asistencia_1	n
mayo_n_inst_efectivo_1	n
junio_asistencia_1	n
junio_n_inst_efectivo_1	n
julio_asistencia_1	n
julio_n_inst_efectivo_1	n
agosto_asistencia_1	n
agosto_n_inst_efectivo_1	n
septiembre_asistencia_1	n
septiembre_n_inst_efectivo_1	n

octubre_asistencia_1	n
octubre_n_inst_efectivo_1	n
noviembre_asistencia_1	n
noviembre_n_inst_efectivo_1	n
diciembre_asistencia_1	n
diciembre_n_inst_efectivo_1	n
marzo_asistencia_2	n
marzo_n_inst_efectivo_2	n
abril_asistencia_2	n
abril_n_inst_efectivo_2	n
mayo_asistencia_2	n
mayo_n_inst_efectivo_2	n
junio_asistencia_2	n
junio_n_inst_efectivo_2	n
julio_asistencia_2	n
julio_n_inst_efectivo_2	n
agosto_asistencia_2	n
agosto_n_inst_efectivo_2	n
septiembre_asistencia_2	n
septiembre_n_inst_efectivo_2	n
octubre_asistencia_2	n
octubre_n_inst_efectivo_2	n
noviembre_asistencia_2	n
noviembre_n_inst_efectivo_2	n
diciembre_asistencia_2	n
diciembre_n_inst_efectivo_2	n
marzo_asistencia_3	n
marzo_n_inst_efectivo_3	n
abril_asistencia_3	n
abril_n_inst_efectivo_3	n
mayo_asistencia_3	n
mayo_n_inst_efectivo_3	n
junio_asistencia_3	n
junio_n_inst_efectivo_3	n
julio_asistencia_3	n
julio_n_inst_efectivo_3	n
agosto_asistencia_3	n
agosto_n_inst_efectivo_3	n
septiembre_asistencia_3	n
septiembre_n_inst_efectivo_3	n
octubre_asistencia_3	n
octubre_n_inst_efectivo_3	n
noviembre_asistencia_3	n

noviembre_n_inst_efectivo_3	n
diciembre_asistencia_3	n
diciembre_n_inst_efectivo_3	n
asistencia_1	n
asistencia_2	n
asistencia_3	n
asistencia_4	n
asistencia_5	n
edad	n
prom_gral_1	n
prom_gral_2	n
prom_gral_3	n
prom_gral_4	n
prom_gral_5	n
sit_fin_r_1	c
sit_fin_r_2	c
sit_fin_r_3	c
sit_fin_r_4	c
sit_fin_r_5	c
sobre_edad	n
ben_sep	c
clasificacion_sep	c
convenio_sep	c
criterio_sep	c
grado_sep	c
marzo_asistencia_0	n
marzo_n_inst_efectivo_0	n
abril_asistencia_0	n
abril_n_inst_efectivo_0	n
mayo_asistencia_0	n
mayo_n_inst_efectivo_0	n
junio_asistencia_0	n
junio_n_inst_efectivo_0	n
julio_asistencia_0	n
julio_n_inst_efectivo_0	n
estab_cod_depe	c
estab_cod_depe2	c
estab_convenio_pie	c
estab_ens_01	c
estab_ens_02	c
estab_ens_03	c
estab_ens_04	c
estab_ens_05	c

estab_ens_06	c
estab_ens_07	c
estab_ens_08	c
estab_ens_09	c
estab_matricula	c
estab_estado_estab	c
estab_ori_religiosa	c
estab_pago_matricula	c
estab_pago_mensual	c
docentes_dc_a	n
docentes_dc_dir	n
docentes_dc_dir_sost	n
docentes_dc_educ_trad	n
docentes_dc_ig	n
docentes_dc_jutp	n
docentes_dc_oes	n
docentes_dc_of	n
docentes_dc_or	n
docentes_dc_pdir	n
docentes_dc_prof_enc	n
docentes_dc_subdir	n
docentes_dc_sup_sost	n
docentes_dc_tot	n
docentes_dc_tp_sost	n
docentes_dc_utp	n
docentes_hh_a	n
docentes_hh_dir	n
docentes_hh_dir_sost	n
docentes_hh_educ_trad	n
docentes_hh_ig	n
docentes_hh_jutp	n
docentes_hh_oes	n
docentes_hh_of	n
docentes_hh_or	n
docentes_hh_pdir	n
docentes_hh_prof_enc	n
docentes_hh_subdir	n
docentes_hh_sup_sost	n
docentes_hh_tot	n
docentes_hh_tp_sost	n
docentes_hh_utp	n
asistentes_jorn_aux	n
asistentes_jorn_para	n

asistentes_jorn_prof	n
asistentes_jorn_sin_dato	n
asistentes_n_asis	n
asistentes_n_aux	n
asistentes_n_hombres	n
asistentes_n_mujeres	n
asistentes_n_para	n
asistentes_n_prof	n
asistentes_n_sin_dato	n
asistentes_tot_jorn	n
mat_rbd_cur_comb_01	n
mat_rbd_cur_comb_02	n
mat_rbd_cur_comb_tot	n
mat_rbd_cur_sim_01	n
mat_rbd_cur_sim_02	n
mat_rbd_cur_sim_03	n
mat_rbd_cur_sim_04	n
mat_rbd_cur_sim_05	n
mat_rbd_cur_sim_06	n
mat_rbd_cur_sim_07	n
mat_rbd_cur_sim_08	n
mat_rbd_cur_sim_tot	n
mat_rbd_mat_ens_1	n
mat_rbd_mat_ens_2	n
mat_rbd_mat_ens_3	n
mat_rbd_mat_ens_4	n
mat_rbd_mat_ens_5	n
mat_rbd_mat_ens_6	n
mat_rbd_mat_ens_7	n
mat_rbd_mat_ens_8	n
mat_rbd_mat_gd_total	n
mat_rbd_mat_hom_1	n
mat_rbd_mat_hom_2	n
mat_rbd_mat_hom_3	n
mat_rbd_mat_hom_4	n
mat_rbd_mat_hom_5	n
mat_rbd_mat_hom_6	n
mat_rbd_mat_hom_7	n
mat_rbd_mat_hom_8	n
mat_rbd_mat_hom_gd	n
mat_rbd_mat_hom_tot	n
mat_rbd_mat_muj_1	n
mat_rbd_mat_muj_2	n

mat_rbd_mat_muj_3	n
mat_rbd_mat_muj_4	n
mat_rbd_mat_muj_5	n
mat_rbd_mat_muj_6	n
mat_rbd_mat_muj_7	n
mat_rbd_mat_muj_8	n
mat_rbd_mat_muj_gd	n
mat_rbd_mat_muj_tot	n
mat_rbd_mat_si_2	n
mat_rbd_mat_si_tot	n
mat_rbd_mat_total	n
sep_clasificacion_sep	c
sep_n_ben	n
sep_n_prio	n
sned_efectivr	n
sned_igualdr	n
sned_indicer	n
sned_iniciar	n
sned_integrar	n
sned_mejorar	n
sned_sel	c
sned_sel_25	c
sned_sel_35	c
sned_superar	n
sned_tipo_est	c
subv_%_zona	n
subv_adeco	n
subv_aep	n
subv_aporte_estado_ficom	n
subv_area	c
subv_avdi	n
subv_brp	n
subv_dependencia	c
subv_descuento_escolaridad	n
subv_descuento_ficom	n
subv_desempeÃ±o_dificil	n
subv_discrepancia	n
subv_escolaridad	n
subv_escolaridad_pie	n
subv_internado	n
subv_liquido_pago	n
subv_mantenimiento	n
subv_meses	n

subv_multas	n
subv_pag_pend	n
subv_piso_rural	n
subv_profesor_encargado	n
subv_prom_matricula	n
subv_proretencion	n
subv_reforzamiento	n
subv_reintegros	n
subv_reliquidacion	n
subv_retenciones	n
subv_ruralidad	n
subv_sep	n
subv_sned	n
subv_subv_adicional_especial	n
subv_subv_asistentes_educacion	n
subv_subv_normal	n
subv_zona	n
rend_apr_hom_01	n
rend_apr_hom_02	n
rend_apr_hom_03	n
rend_apr_hom_04	n
rend_apr_hom_05	n
rend_apr_hom_06	n
rend_apr_hom_07	n
rend_apr_hom_08	n
rend_apr_hom_to	n
rend_apr_ind_02	n
rend_apr_muj_01	n
rend_apr_muj_02	n
rend_apr_muj_03	n
rend_apr_muj_04	n
rend_apr_muj_05	n
rend_apr_muj_06	n
rend_apr_muj_07	n
rend_apr_muj_08	n
rend_apr_muj_to	n
rend_prom_asis	n
rend_prom_asis_apr	n
rend_prom_asis_apr_hom	n
rend_prom_asis_apr_ind	n
rend_prom_asis_apr_muj	n
rend_prom_asis_rep	n
rend_prom_asis_rep_hom	n

rend_prom_asis_rep_muj	n
rend_rep_hom_01	n
rend_rep_hom_02	n
rend_rep_hom_03	n
rend_rep_hom_04	n
rend_rep_hom_05	n
rend_rep_hom_06	n
rend_rep_hom_07	n
rend_rep_hom_08	n
rend_rep_hom_to	n
rend_rep_muj_01	n
rend_rep_muj_02	n
rend_rep_muj_03	n
rend_rep_muj_04	n
rend_rep_muj_05	n
rend_rep_muj_06	n
rend_rep_muj_07	n
rend_rep_muj_08	n
rend_rep_muj_to	n
rend_ret_hom_01	n
rend_ret_hom_02	n
rend_ret_hom_03	n
rend_ret_hom_04	n
rend_ret_hom_05	n
rend_ret_hom_06	n
rend_ret_hom_07	n
rend_ret_hom_08	n
rend_ret_hom_to	n
rend_ret_muj_01	n
rend_ret_muj_02	n
rend_ret_muj_03	n
rend_ret_muj_04	n
rend_ret_muj_05	n
rend_ret_muj_06	n
rend_ret_muj_07	n
rend_ret_muj_08	n
rend_ret_muj_to	n
rend_si_hom_01	n
rend_si_hom_02	n
rend_si_hom_03	n
rend_si_hom_04	n
rend_si_hom_05	n
rend_si_hom_06	n

rend_si_hom_07	n
rend_si_hom_to	n
rend_si_muj_01	n
rend_si_muj_02	n
rend_si_muj_03	n
rend_si_muj_07	n
rend_si_muj_to	n
rend_tra_hom_01	n
rend_tra_hom_02	n
rend_tra_hom_03	n
rend_tra_hom_04	n
rend_tra_hom_05	n
rend_tra_hom_06	n
rend_tra_hom_07	n
rend_tra_hom_08	n
rend_tra_hom_to	n
rend_tra_ind_02	n
rend_tra_muj_01	n
rend_tra_muj_02	n
rend_tra_muj_03	n
rend_tra_muj_04	n
rend_tra_muj_05	n
rend_tra_muj_06	n
rend_tra_muj_07	n
rend_tra_muj_08	n
rend_tra_muj_to	n
mat_ue_com_jor_ma	n
mat_ue_com_jor_mt	n
mat_ue_com_jor_ta	n
mat_ue_com_jor_ve	n
mat_ue_cur_comb	n
mat_ue_cur_sim_01	n
mat_ue_cur_sim_02	n
mat_ue_cur_sim_03	n
mat_ue_cur_sim_04	n
mat_ue_cur_sim_05	n
mat_ue_cur_sim_06	n
mat_ue_cur_sim_07	n
mat_ue_cur_sim_08	n
mat_ue_cur_sim_tot	n
mat_ue_cursos_dual_3	n
mat_ue_cursos_dual_4	n
mat_ue_cursos_dual_tot	n

mat_ue_mat_cur_dual_3	n
mat_ue_mat_cur_dual_4	n
mat_ue_mat_cur_dual_hom	n
mat_ue_mat_cur_dual_hom_3	n
mat_ue_mat_cur_dual_hom_4	n
mat_ue_mat_cur_dual_muj	n
mat_ue_mat_cur_dual_muj_3	n
mat_ue_mat_cur_dual_muj_4	n
mat_ue_mat_cur_dual_tot	n
mat_ue_mat_gd_total	n
mat_ue_mat_gra_1	n
mat_ue_mat_gra_2	n
mat_ue_mat_gra_3	n
mat_ue_mat_gra_4	n
mat_ue_mat_gra_5	n
mat_ue_mat_gra_6	n
mat_ue_mat_gra_7	n
mat_ue_mat_gra_8	n
mat_ue_mat_hom_1	n
mat_ue_mat_hom_2	n
mat_ue_mat_hom_3	n
mat_ue_mat_hom_4	n
mat_ue_mat_hom_5	n
mat_ue_mat_hom_6	n
mat_ue_mat_hom_7	n
mat_ue_mat_hom_8	n
mat_ue_mat_hom_gd	n
mat_ue_mat_hom_tot	n
mat_ue_mat_jor_ma	n
mat_ue_mat_jor_mt	n
mat_ue_mat_jor_ne	n
mat_ue_mat_jor_ta	n
mat_ue_mat_jor_tot	n
mat_ue_mat_jor_ve	n
mat_ue_mat_muj_1	n
mat_ue_mat_muj_2	n
mat_ue_mat_muj_3	n
mat_ue_mat_muj_4	n
mat_ue_mat_muj_5	n
mat_ue_mat_muj_6	n
mat_ue_mat_muj_7	n
mat_ue_mat_muj_8	n
mat_ue_mat_muj_gd	n

mat_ue_mat_muj_tot	n
mat_ue_mat_si_2	n
mat_ue_mat_si_tot	n
mat_ue_mat_total	n
mat_ue_sim_jor_ma	n
mat_ue_sim_jor_mt	n
mat_ue_sim_jor_ne	n
mat_ue_sim_jor_ta	n
mat_ue_sim_jor_ve	n
jecd_id_bas_jecd	c
jecd_id_esp_jecd	c
jecd_id_hc_jecd	c
jecd_id_hos_jecd	c
jecd_id_parv_jecd	c
jecd_id_pie_jecd	c
jecd_id_tp_jecd	c
jecd_mat_310_1	n
jecd_mat_310_2	n
jecd_mat_310_3	n
jecd_mat_310_4	n
jecd_mat_410_1	n
jecd_mat_410_2	n
jecd_mat_410_3	n
jecd_mat_410_4	n
jecd_mat_510_1	n
jecd_mat_510_2	n
jecd_mat_510_3	n
jecd_mat_510_4	n
jecd_mat_610_1	n
jecd_mat_610_2	n
jecd_mat_610_3	n
jecd_mat_610_4	n
jecd_mat_710_1	n
jecd_mat_710_2	n
jecd_mat_710_3	n
jecd_mat_710_4	n
jecd_mat_810_1	n
jecd_mat_810_2	n
jecd_mat_810_3	n
jecd_mat_810_4	n
jecd_mat_b_1	n
jecd_mat_b_2	n
jecd_mat_b_3	n

jecd_mat_b_4	n
jecd_mat_b_5	n
jecd_mat_b_6	n
jecd_mat_b_7	n
jecd_mat_b_8	n
jecd_mat_cj	n
jecd_mat_e	n
jecd_mat_hos	n
jecd_mat_k	n
jecd_mat_pie	n
jecd_mat_pk	n
ficom_asi_prom	n
ficom_cobropromedio	n
ficom_en_use	n
ficom_ing_efectivo	n
ficom_prom_mensual	n
ficom_tramo	c
ficom_valor_use	n
ive_basica_primera prioridad	n
ive_basica_segunda prioridad	n
ive_basica_tercera prioridad	n
ive_basica_no vulnerables	n
ive_basica_sin informacion	n
ive_basica_total matricula basica 2	n
ive_basica_ive-sinae basica 1	n
ive_media_primera prioridad	n
ive_media_segunda prioridad	n
ive_media_tercera prioridad	n
ive_media_no vulnerables	n
ive_media_sin informacion	n
ive_media_total matricula media 2	n
ive_media_ive-sinae media 1	n
ive_comuna_primera prioridad	n
ive_comuna_segunda prioridad	n
ive_comuna_tercera prioridad	n
ive_comuna_no vulnerables	n
ive_comuna_sin informacion	n
ive_comuna_total matricula basica-media 2	n
ive_comuna_ive-sinae comunal 1	n
idps_4b_ind_am	n
idps_4b_ind_cc	n
idps_4b_ind_pf	n

idps_4b_ind_hv	n
idps_6b_ind_am	n
idps_6b_ind_cc	n
idps_6b_ind_pf	n
idps_6b_ind_hv	n
idps_2m_ind_am	n
idps_2m_ind_cc	n
idps_2m_ind_pf	n
idps_2m_ind_hv	n
simce_cod_depe1	c
simce_cod_depe2	c
simce_cod_grupo	c
simce_cod_rural_rbd	c
simce_dif_lect4b_com	n
simce_dif_lect4b_deprov	n
simce_dif_lect4b_rbd	n
simce_dif_lect4b_reg	n
simce_dif_mate4b_com	n
simce_dif_mate4b_deprov	n
simce_dif_mate4b_rbd	n
simce_dif_mate4b_reg	n
simce_difgru_lect4b_rbd	n
simce_difgru_mate4b_rbd	n
simce_marca_lect4b_rbd	c
simce_marca_mate4b_rbd	c
simce_marcadif_lect4b_rbd	c
simce_marcadif_mate4b_rbd	c
simce_nalu_lect4b_com	n
simce_nalu_lect4b_deprov	n
simce_nalu_lect4b_rbd	n
simce_nalu_lect4b_reg	n
simce_nalu_mate4b_com	n
simce_nalu_mate4b_deprov	n
simce_nalu_mate4b_rbd	n
simce_nalu_mate4b_reg	n
simce_noaplica	c
simce_palu_eda_ade_lect4b_rbd	n
simce_palu_eda_ade_lect4b_reg	n
simce_palu_eda_ade_mate4b_rbd	n
simce_palu_eda_ade_mate4b_reg	n
simce_palu_eda_ele_lect4b_rbd	n
simce_palu_eda_ele_lect4b_reg	n
simce_palu_eda_ele_mate4b_rbd	n

simce_palu_eda_ele_mate4b_reg	n
simce_palu_eda_ins_lect4b_rbd	n
simce_palu_eda_ins_lect4b_reg	n
simce_palu_eda_ins_mate4b_rbd	n
simce_palu_eda_ins_mate4b_reg	n
simce_prom_lect4b_com	n
simce_prom_lect4b_deprov	n
simce_prom_lect4b_rbd	n
simce_prom_lect4b_reg	n
simce_prom_mate4b_com	n
simce_prom_mate4b_deprov	n
simce_prom_mate4b_rbd	n
simce_prom_mate4b_reg	n
simce_sigdif_lect4b_com	c
simce_sigdif_lect4b_deprov	c
simce_sigdif_lect4b_rbd	c
simce_sigdif_lect4b_reg	c
simce_sigdif_mate4b_com	c
simce_sigdif_mate4b_deprov	c
simce_sigdif_mate4b_rbd	c
simce_sigdif_mate4b_reg	c
simce_siggru_lect4b_rbd	c
simce_siggru_mate4b_rbd	c
simce_dif_lect8b_com	n
simce_dif_lect8b_deprov	n
simce_dif_lect8b_rbd	n
simce_dif_lect8b_reg	n
simce_dif_mate8b_com	n
simce_dif_mate8b_deprov	n
simce_dif_mate8b_rbd	n
simce_dif_mate8b_reg	n
simce_dif_nat8b_com	n
simce_dif_nat8b_deprov	n
simce_dif_nat8b_rbd	n
simce_dif_nat8b_reg	n
simce_difgru_lect8b_rbd	n
simce_difgru_mate8b_rbd	n
simce_difgru_nat8b_rbd	n
simce_marca_lect8b_rbd	c
simce_marca_mate8b_rbd	c
simce_marca_nat8b_rbd	c
simce_marcadif_lect8b_rbd	c
simce_marcadif_mate8b_rbd	c

simce_marcadif_nat8b_rbd	c
simce_nalu_lect8b_com	n
simce_nalu_lect8b_deprov	n
simce_nalu_lect8b_rbd	n
simce_nalu_lect8b_reg	n
simce_nalu_mate8b_com	n
simce_nalu_mate8b_deprov	n
simce_nalu_mate8b_rbd	n
simce_nalu_mate8b_reg	n
simce_nalu_nat8b_com	n
simce_nalu_nat8b_deprov	n
simce_nalu_nat8b_rbd	n
simce_nalu_nat8b_reg	n
simce_palu_eda_ade_lect8b_rbd	n
simce_palu_eda_ade_lect8b_reg	n
simce_palu_eda_ade_mate8b_rbd	n
simce_palu_eda_ade_mate8b_reg	n
simce_palu_eda_ade_nat8b_rbd	n
simce_palu_eda_ade_nat8b_reg	n
simce_palu_eda_ele_lect8b_rbd	n
simce_palu_eda_ele_lect8b_reg	n
simce_palu_eda_ele_mate8b_rbd	n
simce_palu_eda_ele_mate8b_reg	n
simce_palu_eda_ele_nat8b_rbd	n
simce_palu_eda_ele_nat8b_reg	n
simce_palu_eda_ins_lect8b_rbd	n
simce_palu_eda_ins_lect8b_reg	n
simce_palu_eda_ins_mate8b_rbd	n
simce_palu_eda_ins_mate8b_reg	n
simce_palu_eda_ins_nat8b_rbd	n
simce_palu_eda_ins_nat8b_reg	n
simce_prom_lect8b_com	n
simce_prom_lect8b_deprov	n
simce_prom_lect8b_rbd	n
simce_prom_lect8b_reg	n
simce_prom_mate8b_com	n
simce_prom_mate8b_deprov	n
simce_prom_mate8b_rbd	n
simce_prom_mate8b_reg	n
simce_prom_nat8b_com	n
simce_prom_nat8b_deprov	n
simce_prom_nat8b_rbd	n
simce_prom_nat8b_reg	n

simce_sigdif_lect8b_com	c
simce_sigdif_lect8b_deprov	c
simce_sigdif_lect8b_rbd	c
simce_sigdif_lect8b_reg	c
simce_sigdif_mate8b_com	c
simce_sigdif_mate8b_deprov	c
simce_sigdif_mate8b_rbd	c
simce_sigdif_mate8b_reg	c
simce_sigdif_nat8b_com	c
simce_sigdif_nat8b_deprov	c
simce_sigdif_nat8b_rbd	c
simce_sigdif_nat8b_reg	c
simce_siggru_lect8b_rbd	n
simce_siggru_mate8b_rbd	n
simce_siggru_nat8b_rbd	n
simce_dif_lect2m_com	n
simce_dif_lect2m_deprov	n
simce_dif_lect2m_rbd	n
simce_dif_lect2m_reg	n
simce_dif_mate2m_com	n
simce_dif_mate2m_deprov	n
simce_dif_mate2m_rbd	n
simce_dif_mate2m_reg	n
simce_dif_soc2m_rbd	n
simce_dif_soc2m_reg	n
simce_difgru_lect2m_rbd	n
simce_difgru_mate2m_rbd	n
simce_difgru_soc2m_rbd	n
simce_marca_lect2m_rbd	c
simce_marca_mate2m_rbd	c
simce_marca_soc2m_rbd	c
simce_marcadif_lect2m_rbd	c
simce_marcadif_mate2m_rbd	c
simce_nalu_lect2m_com	n
simce_nalu_lect2m_deprov	n
simce_nalu_lect2m_rbd	n
simce_nalu_lect2m_reg	n
simce_nalu_mate2m_com	n
simce_nalu_mate2m_deprov	n
simce_nalu_mate2m_rbd	n
simce_nalu_mate2m_reg	n
simce_nalu_soc2m_com	n
simce_nalu_soc2m_deprov	n

simce_nalu_soc2m_rbd	n
simce_nalu_soc2m_reg	n
simce_palu_eda_ade_lect2m_rbd	n
simce_palu_eda_ade_lect2m_reg	n
simce_palu_eda_ade_mate2m_rbd	n
simce_palu_eda_ade_mate2m_reg	n
simce_palu_eda_ele_lect2m_rbd	n
simce_palu_eda_ele_lect2m_reg	n
simce_palu_eda_ele_mate2m_rbd	n
simce_palu_eda_ele_mate2m_reg	n
simce_palu_eda_ins_lect2m_rbd	n
simce_palu_eda_ins_lect2m_reg	n
simce_palu_eda_ins_mate2m_rbd	n
simce_palu_eda_ins_mate2m_reg	n
simce_prom_lect2m_com	n
simce_prom_lect2m_deprov	n
simce_prom_lect2m_rbd	n
simce_prom_lect2m_reg	n
simce_prom_mate2m_com	n
simce_prom_mate2m_deprov	n
simce_prom_mate2m_rbd	n
simce_prom_mate2m_reg	n
simce_prom_soc2m_com	n
simce_prom_soc2m_deprov	n
simce_prom_soc2m_rbd	n
simce_prom_soc2m_reg	n
simce_sigdif_lect2m_com	c
simce_sigdif_lect2m_deprov	c
simce_sigdif_lect2m_rbd	c
simce_sigdif_lect2m_reg	c
simce_sigdif_mate2m_com	c
simce_sigdif_mate2m_deprov	c
simce_sigdif_mate2m_rbd	c
simce_sigdif_mate2m_reg	c
simce_sigdif_soc2m_rbd	c
simce_sigdif_soc2m_reg	c
simce_siggru_lect2m_rbd	n
simce_siggru_mate2m_rbd	n
simce_siggru_soc2m_rbd	n
simce_dif_lect2b_com	n
simce_dif_lect2b_deprov	n
simce_dif_lect2b_rbd	n
simce_dif_lect2b_reg	n

simce_difgru_lect2b_rbd	n
simce_marca_lect2b_rbd	n
simce_marcadif_lect2b_rbd	n
simce_nalu_lect2b_com	n
simce_nalu_lect2b_deprov	n
simce_nalu_lect2b_rbd	n
simce_nalu_lect2b_reg	n
simce_palu_eda_ade_lect2b_rbd	n
simce_palu_eda_ade_lect2b_reg	n
simce_palu_eda_ele_lect2b_rbd	n
simce_palu_eda_ele_lect2b_reg	n
simce_palu_eda_ins_lect2b_rbd	n
simce_palu_eda_ins_lect2b_reg	n
simce_prom_lect2b_com	n
simce_prom_lect2b_deprov	n
simce_prom_lect2b_rbd	n
simce_prom_lect2b_reg	n
simce_sigdif_lect2b_com	c
simce_sigdif_lect2b_deprov	c
simce_sigdif_lect2b_rbd	c
simce_sigdif_lect2b_reg	c
simce_siggru_lect2b_rbd	c
simce_dif_lect6b_com	n
simce_dif_lect6b_deprov	n
simce_dif_lect6b_rbd	n
simce_dif_lect6b_reg	n
simce_dif_mate6b_com	n
simce_dif_mate6b_deprov	n
simce_dif_mate6b_rbd	n
simce_dif_mate6b_reg	n
simce_dif_soc6b_com	n
simce_dif_soc6b_deprov	n
simce_dif_soc6b_rbd	n
simce_dif_soc6b_reg	n
simce_difgru_lect6b_rbd	n
simce_difgru_mate6b_rbd	n
simce_difgru_soc6b_rbd	n
simce_marca_lect6b_rbd	c
simce_marca_mate6b_rbd	c
simce_marca_soc6b_rbd	c
simce_marcadif_lect6b_rbd	c
simce_marcadif_mate6b_rbd	c
simce_marcadif_soc6b_rbd	c

simce_nalu_lect6b_com	n
simce_nalu_lect6b_deprov	n
simce_nalu_lect6b_rbd	n
simce_nalu_lect6b_reg	n
simce_nalu_mate6b_com	n
simce_nalu_mate6b_deprov	n
simce_nalu_mate6b_rbd	n
simce_nalu_mate6b_reg	n
simce_nalu_soc6b_com	n
simce_nalu_soc6b_deprov	n
simce_nalu_soc6b_rbd	n
simce_nalu_soc6b_reg	n
simce_prom_lect6b_com	n
simce_prom_lect6b_deprov	n
simce_prom_lect6b_rbd	n
simce_prom_lect6b_reg	n
simce_prom_mate6b_com	n
simce_prom_mate6b_deprov	n
simce_prom_mate6b_rbd	n
simce_prom_mate6b_reg	n
simce_prom_soc6b_com	n
simce_prom_soc6b_deprov	n
simce_prom_soc6b_rbd	n
simce_prom_soc6b_reg	n
simce_sigdif_lect6b_com	c
simce_sigdif_lect6b_deprov	c
simce_sigdif_lect6b_rbd	c
simce_sigdif_lect6b_reg	c
simce_sigdif_mate6b_com	c
simce_sigdif_mate6b_deprov	c
simce_sigdif_mate6b_rbd	c
simce_sigdif_mate6b_reg	c
simce_sigdif_soc6b_com	c
simce_sigdif_soc6b_deprov	c
simce_sigdif_soc6b_rbd	c
simce_sigdif_soc6b_reg	c
simce_siggru_lect6b_rbd	n
simce_siggru_mate6b_rbd	n
simce_siggru_soc6b_rbd	n
simce_prom_lect_general_rbd	n
simce_prom_lect_general_com	n
simce_prom_lect_general_deprov	n
simce_prom_lect_general_reg	n

simce_prom_mat_general_rbd	n
simce_prom_mat_general_com	n
simce_prom_mat_general_deprov	n
simce_prom_mat_general_reg	n
simce_prom_soc_general_rbd	n
simce_prom_soc_general_com	n
simce_prom_soc_general_deprov	n
simce_prom_soc_general_reg	n
ide_4b_score13	n
ide_4b_score10	n
ide_4b_score9	n
ide_4b_score8	n
ide_4b_score7	n
ide_4b_score6	n
ide_4b_score5	n
ide_4b_score4	n
ide_4b_score3	n
ide_4b_score2	n
ide_4b_score1	n
ide_4b_score0	n
ide_4b_score13_10	n
ide_4b_score10_9	n
ide_4b_score9_8	n
ide_4b_score8_7	n
ide_4b_score7_6	n
ide_4b_score6_5	n
ide_4b_score5_4	n
ide_4b_score4_3	n
ide_4b_score3_2	n
ide_4b_score2_1	n
ide_4b_score1_0	n
ide_8b_score11	n
ide_8b_score8	n
ide_8b_score6	n
ide_8b_score4	n
ide_8b_score2	n
ide_8b_score1	n
ide_8b_score0	n
ide_8b_score11_8	n
ide_8b_score8_6	n
ide_8b_score6_4	n
ide_8b_score4_2	n
ide_8b_score2_1	n

ide_8b_score1_0	n
ide_2m_score14	n
ide_2m_score12	n
ide_2m_score9	n
ide_2m_score7	n
ide_2m_score5	n
ide_2m_score3	n
ide_2m_score2	n
ide_2m_score1	n
ide_2m_score0	n
ide_2m_score14_12	n
ide_2m_score12_9	n
ide_2m_score9_7	n
ide_2m_score7_5	n
ide_2m_score5_3	n
ide_2m_score3_2	n
ide_2m_score2_1	n
ide_2m_score1_0	n

ANEXO B

Ministerio de Desarrollo Social
Gobierno de Chile

Carta N° 20 / 4448 /

Santiago, 18 OCT 2017

Señor
Diego Ibáñez
diegoii91@gmail.com
PRESENTE

De mi consideración:

En relación a su solicitud de acceso a la información pública ingresada el día 3 de octubre de 2017 con el folio N° AI001T0000943, derivado desde la Subsecretaría de Servicios Sociales, donde indica: *"Solicito la base de datos completa, y documentación para la interpretación, del registro social de hogares, y sus modificaciones a lo largo del tiempo a nivel mensual, desde sus inicios (o en su defecto enero 2016) hasta la fecha.*
Específicamente requiero los registros a nivel de individuo enmascarados bajo "MRUN" (campo utilizado por el ministerio de educación para ocultar los RUN de los estudiantes). No requiero poder identificar las características familiares y su variación en el tiempo de cada estudiante enmascarado bajo MRUN.
Soy memorista del Laboratorio de Gobierno, estoy construyendo un algoritmo que prediga la deserción escolar utilizando factores contextuales e inteligencia artificial. Por ello requiero el contar con la caracterización familiar de las y los estudiantes, bajo la máscara "MRUN" que me permite hacer cruce con las bases de datos del Ministerio de Educación.
Desde el Centro de Estudios del MINEDUC están dispuestos a compartir internamente con el MDS el diccionario MRUNRUN para efectos de este trabajo, en caso de que ustedes no cuenten con ello.", puedo señalar a Ud. lo siguiente:

En primer lugar, es preciso señalar que el Registro Social de Hogares (RSH) es el nuevo Sistema de Apoyo a la Selección de Usuarios de Prestaciones Sociales, regulado en el Decreto N° 22, de 2015, del Ministerio de Desarrollo Social, y tiene por objetivo apoyar los distintos procesos de selección de usuarios de beneficios, prestaciones y programas sociales, a través de la provisión de un conjunto amplio de información, principalmente proveniente de registros administrativos del Estado, entre la cual se incluye la construcción de una Calificación Socioeconómica de los hogares. El reglamento del RSH está disponible en la siguiente dirección:

http://www.registrosocial.gob.cl/wp-content/uploads/2017/02/Decreto-N°22-2015_MDS.pdf

En seguida, en atención al deber de someter todo tratamiento de datos personales y sensibles a lo dispuesto en la Ley N° 19.628, sobre Protección de la Vida Privada, no es posible hacer entrega de información con identificación de las personas, eso es, bases de datos nominadas, por cuanto dicho acto permitiría, en último término, asociar a las personas y sus familias con los datos

personales entregados por ellos con otra finalidad. En esta misma línea, respecto de la propuesta de uso de la máscara MRUN del Ministerio de Educación, debo indicar que este ministerio no trabaja con dicho dato. Además, su uso implicaría la nominación de información bajo la tutela de este ministerio, para quienes manejen el diccionario MRUNRUN mencionado en su solicitud.

Por otra parte, conforme al Artículo 10º de la Ley N° 20.285, Sobre Acceso a la Información Pública, el acceso a la información comprende el derecho de acceder a las informaciones contenidas en actos, resoluciones, actas, expedientes, contratos y acuerdos, así como a toda información elaborada con presupuesto público, cualquiera sea el formato o soporte en que se contenga, salvo las excepciones legales. Esto es, se refiere a la información que obra en poder del órgano de la Administración del Estado al que se dirige, no a la obtención del procesamiento de datos o la elaboración de antecedentes nuevos en base a información existente. En esta línea, no es posible dar cumplimiento a su solicitud, toda vez que para ello, en atención al deber de protección sobre los datos señalado en el punto anterior, su requerimiento involucra el procesamiento de datos y la preparación de información de al menos 22 meses, ya que ésta no se encuentra disponible en los términos solicitados.

Finalmente, por lo expuesto, de conformidad con lo dispuesto en el artículo 14 de la Ley Nº 20.285, se da por evacuado requerimiento de transparencia ya individualizado.

Saluda atentamente a Ud.,

MINISTERIO DE DESARROLLO SOCIAL
SUBSECRETARIA DE EVALUACIÓN SOCIAL

JUAN BERNER HERRERA
SUBSECRETARIA DE EVALUACIÓN SOCIAL

MRM/JSP/CCC/VAA/

Distribución:
Gabinete Subsecretaría Evaluación Social
División Información Social
Oficina de Partes (2)
Expediente: 49807

ANEXO C

A continuación, se presentan las 490 dimensiones seleccionadas en su orden relativo de importancia.

dimensiones	importancia
sobre_edad	1686,686049
noviembre_asistencia_1	728,0540841
julio_n_inst_efectivo_3	449,8941775
mat_ue_mat_gra_5	319,0829532
mayo_asistencia_0	283,2121907
julio_asistencia_0	264,808854
edad	249,4859802
junio_asistencia_0	247,831358
marzo_asistencia_2	244,8559189
prom_gral_1	208,5684663
marzo_asistencia_1	190,490884
junio_n_inst_efectivo_1	164,8054769
diciembre_n_inst_efectivo_1	161,4003655
rend_ret_hom_to	157,3165285
rend_ret_muj_01	156,6546873
marzo_asistencia_0	155,5870939
septiembre_asistencia_1	154,5565191
junio_n_inst_efectivo_2	153,4509861
julio_asistencia_3	138,5742536
mat_ue_mat_si_tot	134,876923
prom_gral_2	133,9514982
diciembre_asistencia_1	131,3375977
abril_asistencia_0	128,7537141
octubre_asistencia_2	113,3612136
noviembre_asistencia_2	110,5643507
sned_efectivr	102,4650981
julio_n_inst_efectivo_1	100,702698
rend_ret_muj_to	91,9137045
subv_profesor_encargado	81,32355753
junio_n_inst_efectivo_3	79,24355407
mat_ue_mat_total	78,81600075
mat_ue_mat_gra_6	75,40269818
sit_fin_r_1_P	67,45339765
mat_ue_mat_gra_7	66,79137611
diciembre_n_inst_efectivo_2	64,5342465
ive_basica_tercera prioridad	63,69901106
marzo_n_inst_efectivo_1	61,32396362

sit_fin_r_1_R	50,82256258
mat_ue_mat_muj_5	48,25027493
junio_asistencia_3	42,90181948
junio_n_inst_efectivo_0	42,37877027
diciembre_n_inst_efectivo_3	42,29581553
mat_rbd_mat_muj_tot	41,52842917
simce_nalu_mate2m_rbd	41,15426755
rend_ret_hom_01	40,30242706
mayo_n_inst_efectivo_0	38,18044317
subv_prom_matricula	37,26776265
ive_basica_sin informacion	35,68964034
rend_apr_muj_to	35,45762483
sep_n_ben	33,63830667
simce_prom_soc_general_rbd	33,47312174
prom_gral_3	32,30196197
mat_rbd_mat_muj_1	32,09862002
abril_n_inst_efectivo_1	31,44323783
diciembre_asistencia_3	31,30693784
subv_sep	31,17804937
rend_apr_hom_05	31,06411243
mat_ue_mat_gra_4	30,30690663
agosto_asistencia_3	29,24386052
simce_prom_lect_general_rbd	29,08970078
subv_subv_adicional_especial	27,96080702
ranking_depe2_asis_gral_grado_rbd_1	27,79436804
rend_ret_hom_04	27,76445908
estab_pago_matricula_$10,001 A $25,000	27,6677246
simce_prom_mate6b_com	27,07164523
sit_fin_r_2_P	27,00466395
simce_prom_mat_general_rbd	26,49083299
octubre_asistencia_1	26,33910529
mat_rbd_mat_total	26,27864777
abril_n_inst_efectivo_0	25,55575006
clasificacion_sep_AUTONOMO	25,02641906
ive_comuna_sin informacion	24,7836384
agosto_n_inst_efectivo_3	24,73863392
mat_ue_mat_hom_6	23,50976622
ficom_prom_mensual	23,34227141
julio_n_inst_efectivo_0	23,24895775
simce_dif_mate6b_reg	23,11665251
subv_sned	22,48109445
septiembre_asistencia_2	22,3543399

mat_ue_mat_muj_6	21,8899913
clasificacion_sep_EMERGENTE	21,77946627
noviembre_n_inst_efectivo_1	21,5768471
octubre_n_inst_efectivo_1	20,42358315
simce_dif_lect6b_rbd	20,41318138
simce_prom_soc6b_com	20,40068019
subv_subv_normal	20,02410357
simce_prom_soc_general_reg	19,572948
rend_prom_asis_apr_hom	19,55332348
julio_n_inst_efectivo_2	19,42985408
sned_cluster	19,39950014
rend_apr_hom_to	19,25436315
ranking_depe2_prom_gral_grado_pro_4	19,09381009
rend_prom_asis_apr	18,91858072
ranking_depe2_prom_gral_grado_com_4	18,89540338
marzo_asistencia_3	18,73474029
octubre_n_inst_efectivo_2	18,7055473
simce_prom_mate2m_reg	18,53373623
rend_rep_hom_08	18,4489994
ficom_asi_prom	18,43216507
ide_2m_score3_2	18,40766949
mat_ue_mat_gra_3	18,16421636
mayo_n_inst_efectivo_1	17,99419548
rend_ret_muj_05	17,86691223
simce_prom_lect_general_reg	17,85565004
ranking_depe2_prom_gral_grado_reg_2	17,77997021
marzo_n_inst_efectivo_0	17,62997462
jecd_mat_b_1	17,40831878
ranking_depe2_prom_gral_grado_reg_4	17,36444098
rend_apr_hom_07	16,75826078
asistentes_n_aux	16,74871287
asistencia_2	16,7479654
rend_prom_asis_apr_muj	16,51800992
simce_dif_lect6b_reg	16,41129622
rend_apr_hom_08	16,11496111
subv_descuento_ficom	15,92601655
ide_4b_score8_7	15,45417154
simce_dif_mate6b_com	15,08212173
marzo_n_inst_efectivo_2	14,98886108
mat_ue_mat_hom_tot	14,72683293
subv_ruralidad	14,61749392
rend_apr_muj_05	14,56606
ive_comuna_tercera prioridad	14,51818451

estab_pago_matricula_$1,000 A $10,000	14,4878693
sep_n_prio	14,34494097
mat_ue_mat_gra_8	14,26714244
prom_gral_5	14,26416665
simce_prom_soc6b_rbd	13,9103796
simce_prom_mate4b_reg	13,8949413
mat_ue_mat_hom_5	13,89406754
mat_ue_mat_hom_8	13,84382026
curso_n_alu	13,76298295
abril_asistencia_2	13,72311668
marzo_n_inst_efectivo_3	13,62783207
rend_apr_muj_06	13,51922292
mat_ue_cur_sim_tot	13,40282644
rend_ret_muj_06	13,38426467
ive_basica_segunda prioridad	13,15611245
octubre_n_inst_efectivo_3	13,14061165
mat_ue_mat_muj_8	13,1268015
simce_difgru_lect6b_rbd	12,90277422
sit_fin_r_3_P	12,83328405
rend_rep_hom_02	12,634244
mat_rbd_cur_comb_tot	12,4472227
ide_4b_score2_1	12,39479693
rend_apr_muj_02	12,3702803
ranking_depe2_prom_gral_grado_pro_3	12,31780239
ive_basica_ive-sinae basica 1	12,30091388
ide_4b_score5	12,30073759
simce_prom_mat_general_reg	12,22208
rend_ret_hom_05	12,1889486
asistentes_n_prof	12,16417285
simce_siggru_lect6b_rbd	12,1470137
rend_apr_muj_07	11,78943989
simce_palu_eda_ele_lect4b_rbd	11,78641586
agosto_n_inst_efectivo_1	11,6719342
simce_prom_mate6b_rbd	11,52790396
mayo_n_inst_efectivo_3	11,28847342
ranking_depe2_asis_gral_grado_reg_1	11,15803187
sned_igualdr	11,14522072
rend_rep_hom_01	11,02357736
ive_comuna_total matricula basica-media 2	10,99928713
rend_apr_hom_06	10,96482098
mat_ue_mat_jor_mt	10,90623017

ide_4b_score7_6	10,87289298
ranking_asis_gral_curso_1	10,73022164
ide_2m_score1_0	10,72696738
simce_palu_eda_ins_mate4b_rbd	10,69130821
mat_rbd_cur_sim_06	10,6693773
rend_rep_muj_05	10,608202
ive_comuna_ive-sinae comunal 1	10,54604685
rend_rep_muj_04	10,51899742
ficom_en_use	10,45659534
jecd_mat_b_3	10,35179069
rend_rep_muj_01	10,33802965
subv_escolaridad_pie	10,28985457
rend_apr_muj_03	10,27828132
mat_rbd_mat_hom_1	10,25151988
simce_difgru_mate6b_rbd	10,12161631
ive_media_total matricula media 2	10,03699602
asistentes_jorn_aux	10,01762846
subv_area_Rural	9,98536015
simce_prom_lect4b_com	9,819439175
ranking_depe2_asis_gral_grado_com_1	9,692393158
ranking_depe2_prom_gral_grado_pro_2	9,625803693
mat_rbd_mat_muj_6	9,59259068
subv_liquido_pago	9,591560067
simce_palu_eda_ele_mate2m_rbd	9,545014378
simce_prom_lect6b_reg	9,5279665
simce_difgru_mate2m_rbd	9,521520053
docentes_hh_oes	9,4825881
ranking_depe2_asis_gral_grado_pro_1	9,447302674
ide_4b_score4_3	9,41323574
mat_rbd_mat_hom_6	9,364845515
idps_6b_ind_pf	9,325381296
ide_4b_score4	9,300591082
ranking_depe2_prom_gral_grado_reg_3	9,269374267
subv_mantenimiento	9,228981411
subv_%_zona	9,166463646
simce_prom_mate4b_com	9,155514732
jecd_mat_310_2	9,09617615
subv_aporte_estado_ficom	8,996646201
ide_4b_score8	8,981078103
simce_prom_soc_general_com	8,90527995
rend_ret_hom_02	8,812998775
subv_internado	8,78770351
mat_ue_mat_muj_4	8,687168118

noviembre_asistencia_3	8,681136362
mat_ue_mat_muj_3	8,655572747
ide_4b_score2	8,63675886
sned_indicer	8,623901015
septiembre_n_inst_efectivo_1	8,61903472
ide_4b_score9	8,617185692
rend_tra_hom_06	8,444285175
rend_tra_muj_05	8,430934205
ive_media_sin informacion	8,372538562
mat_ue_mat_gra_1	8,369688223
subv_escolaridad	8,344146796
julio_asistencia_1	8,339060051
simce_difgru_soc6b_rbd	8,278116505
mat_ue_mat_hom_7	8,268289022
rend_apr_hom_03	8,266450422
ranking_depe2_prom_gral_grado_com_3	8,235888927
simce_palu_eda_ele_lect2m_rbd	8,22440152
jecd_mat_b_4	8,21887112
subv_adeco	8,198037907
mat_rbd_mat_hom_4	8,16576195
ranking_prom_gral_curso_3	8,128996013
subv_brp	8,034733831
simce_difgru_mate4b_rbd	7,99360502
docentes_hh_educ_trad	7,957911788
diciembre_asistencia_2	7,938188121
mat_rbd_mat_ens_6	7,919507992
curso_n_alu_muj	7,814891403
ive_basica_no vulnerables	7,735134305
noviembre_n_inst_efectivo_3	7,719122423
rend_tra_hom_08	7,69225073
ranking_depe2_prom_gral_grado_rbd_3	7,690626513
ive_media_primera prioridad	7,679271256
idps_4b_ind_hv	7,64491916
sit_fin_r_3_R	7,577341284
ide_4b_score1_0	7,473182045
curso_n_alu_hom	7,462745538
junio_asistencia_1	7,394005574
ide_4b_score6	7,335660669
ive_media_tercera prioridad	7,324825077
asistentes_tot_jorn	7,320225003
sned_superar	7,314697788
rend_rep_hom_to	7,302855166
idps_2m_ind_pf	7,241357782

mat_ue_mat_hom_2	7,199132733
ive_media_segunda prioridad	7,178800904
rend_prom_asis_rep_muj	7,135949972
sit_fin_r_4_P	7,10253191
jecd_mat_b_6	7,06723286
subv_subv_asistentes_educacion	7,01453658
ranking_prom_gral_curso_2	6,98151658
subv_zona	6,978022174
rend_rep_muj_to	6,967138065
octubre_asistencia_3	6,940726732
rend_ret_hom_03	6,93565321
simce_nalu_mate6b_rbd	6,879275238
rend_apr_hom_04	6,837058325
mat_rbd_mat_hom_tot	6,823203153
mat_rbd_cur_sim_05	6,818054748
simce_palu_eda_ins_mate2m_rbd	6,669333837
sned_iniciar	6,669163676
ranking_depe2_asis_gral_grado_com_4	6,654640836
mat_ue_mat_hom_4	6,605673166
prom_gral_4	6,594110896
rend_rep_muj_02	6,578265507
simce_dif_lect2m_rbd	6,540123291
ive_media_ive-sinae media 1	6,50421544
julio_asistencia_2	6,447565907
mat_rbd_mat_ens_2	6,44720101
rend_rep_hom_04	6,37914681
simce_prom_lect6b_rbd	6,334759445
subv_aep	6,326930143
mat_rbd_mat_ens_5	6,310243918
asistencia_5	6,2899017
mat_rbd_mat_ens_1	6,284251708
ranking_depe2_asis_gral_grado_reg_2	6,280834941
docentes_hh_dir	6,27653503
ranking_depe2_prom_gral_grado_rbd_1	6,269123903
ide_2m_score2	6,244578414
rend_prom_asis_rep	6,237345999
ive_basica_primera prioridad	6,23220478
ive_comuna_segunda prioridad	6,17310254
jecd_mat_310_1	6,13119936
idps_6b_ind_hv	6,072679291
mat_ue_sim_jor_ta	6,064255595
simce_palu_eda_ade_mate2m_rbd	6,04557824
ide_8b_score2	5,977988825

junio_asistencia_2	5,965440798
docentes_hh_prof_enc	5,90740728
simce_dif_mate6b_rbd	5,895822695
simce_dif_lect4b_rbd	5,847766537
simce_palu_eda_ins_lect4b_rbd	5,807120418
simce_difgru_lect2m_rbd	5,742983935
ranking_prom_gral_curso_1	5,667216957
docentes_hh_tot	5,598545841
rend_prom_asis_rep_hom	5,564632736
ide_4b_score9_8	5,542363783
idps_6b_ind_cc	5,528908835
docentes_hh_a	5,446091028
rend_rep_hom_05	5,402820765
rend_rep_muj_07	5,376452625
ive_basica_total matricula basica 2	5,334306877
simce_dif_lect2m_com	5,308417708
mat_rbd_cur_sim_07	5,209143313
mayo_asistencia_1	5,190702842
simce_prom_lect2m_reg	5,183743
simce_dif_lect4b_reg	5,163794995
ide_4b_score10	5,132647425
septiembre_asistencia_3	5,113695485
simce_dif_mate4b_reg	5,109921695
simce_prom_lect_general_com	5,028841234
idps_4b_ind_cc	5,003182239
mat_ue_mat_hom_3	4,99647641
ide_8b_score2_1	4,98796053
jecd_mat_b_2	4,98087764
abril_asistencia_3	4,979496038
mat_rbd_cur_sim_01	4,96754408
ide_8b_score1	4,965714864
jecd_mat_cj	4,96559182
ranking_depe2_prom_gral_grado_com_1	4,895627535
jecd_mat_pk	4,874195704
simce_dif_mate4b_com	4,838985724
ranking_depe2_asis_gral_grado_com_2	4,836487764
simce_palu_eda_ele_mate4b_rbd	4,835415981
mat_ue_mat_gra_2	4,818024196
sit_fin_r_5_P	4,80614376
ranking_depe2_prom_gral_grado_rbd_2	4,797368634
agosto_asistencia_2	4,794396874
mat_rbd_cur_sim_04	4,79093226
idps_4b_ind_am	4,777413412

simce_dif_mate2m_com	4,774210574
ranking_depe2_asis_gral_grado_rbd_4	4,773726694
simce_dif_mate4b_rbd	4,707216104
ranking_depe2_asis_gral_grado_pro_2	4,696449042
docentes_hh_or	4,690936805
docentes_hh_jutp	4,68599963
rend_tra_hom_02	4,68088265
ide_2m_score2_1	4,67796891
mat_ue_mat_muj_tot	4,583655901
ranking_depe2_prom_gral_grado_rbd_4	4,557645763
rend_tra_muj_01	4,492526173
rend_rep_muj_03	4,477795573
simce_dif_lect4b_com	4,469611586
simce_palu_eda_ade_mate4b_rbd	4,466923999
ide_4b_score1	4,446690367
asistencia_1	4,438104951
docentes_hh_of	4,415997034
mayo_asistencia_3	4,269918162
ranking_depe2_asis_gral_grado_rbd_2	4,268527303
ranking_depe2_asis_gral_grado_reg_3	4,171046576
rend_tra_muj_06	4,15371084
ide_4b_score6_5	4,12241427
mat_ue_sim_jor_mt	4,092881291
mayo_asistencia_2	4,012241144
simce_prom_lect6b_com	3,982530974
ide_4b_score3	3,982360108
ranking_depe2_asis_gral_grado_rbd_3	3,974477995
ranking_depe2_prom_gral_grado_com_2	3,897359754
ive_comuna_primera prioridad	3,880680576
docentes_hh_ig	3,865807409
rend_apr_muj_01	3,865609981
sned_mejorar	3,852715254
mat_rbd_mat_hom_5	3,84525657
idps_2m_ind_cc	3,836133686
simce_nalu_lect4b_rbd	3,82278073
jecd_mat_810_4	3,7753582
asistencia_4	3,711350631
jecd_mat_410_3	3,700688211
ide_4b_score3_2	3,679257985
ide_2m_score3	3,672207616
abril_n_inst_efectivo_3	3,60026303
asistentes_jorn_prof	3,555368564
simce_difgru_lect4b_rbd	3,544560393

jecd_mat_b_8	3,487844269
docentes_hh_utp	3,478111992
mat_ue_mat_cur_dual_muj_4	3,460436345
docentes_dc_a	3,45238256
mat_ue_mat_muj_1	3,445587313
jecd_mat_310_3	3,444172775
sned_integrar	3,409963421
simce_dif_lect6b_com	3,388936489
simce_prom_mate2m_rbd	3,385846349
asistentes_jorn_para	3,38559085
simce_prom_mate2m_com	3,38347258
ide_4b_score10_9	3,340185877
idps_6b_ind_am	3,330505163
asistentes_n_para	3,300048768
ranking_depe2_asis_gral_grado_com_3	3,224785451
mat_ue_mat_muj_2	3,22469756
mat_ue_mat_jor_ma	3,191154589
rend_ret_hom_07	3,18916416
jecd_mat_k	3,156791252
simce_prom_mat_general_com	3,105433746
idps_4b_ind_pf	3,070951863
agosto_asistencia_1	3,062013628
mayo_n_inst_efectivo_2	3,054213585
jecd_mat_pie	3,051002495
rend_ret_hom_08	3,039780798
simce_palu_eda_ade_lect4b_rbd	3,018035811
rend_tra_hom_to	2,975640051
ranking_prom_gral_curso_4	2,939720319
ide_2m_score0	2,91590521
ive_media_no vulnerables	2,90433292
rend_tra_hom_01	2,88676548
mat_ue_mat_jor_ta	2,876852193
asistentes_n_asis	2,840854463
ranking_asis_gral_curso_3	2,837264944
jecd_mat_410_1	2,831480874
subv_reliquidacion	2,8169108
simce_prom_lect4b_rbd	2,790742042
jecd_mat_b_7	2,770599191
rend_tra_muj_07	2,755439959
rend_rep_hom_03	2,747630179
mat_rbd_cur_sim_tot	2,746123797
rend_tra_muj_to	2,744691725
ranking_depe2_prom_gral_grado_reg_1	2,674429071

ranking_depe2_prom_gral_grado_pro_1	2,672974513
simce_prom_lect2m_com	2,672666204
docentes_dc_ig	2,6439929
rend_tra_hom_04	2,589001547
jecd_mat_610_3	2,578525146
subv_desempeÃ±o_dificil	2,451277856
rend_tra_muj_03	2,407891764
ive_comuna_no vulnerables	2,353745034
docentes_dc_pdir	2,26896524
mat_rbd_mat_muj_5	2,263144842
ranking_asis_gral_curso_2	2,198784087
simce_prom_lect2m_rbd	2,19168152
ide_4b_score0	2,176517156
ranking_depe2_asis_gral_grado_reg_4	2,132817618
mat_ue_mat_cur_dual_tot	2,1230371
simce_dif_mate2m_reg	2,041676211
subv_avdi	2,007981884
mat_ue_sim_jor_ma	1,97458136
rend_ret_muj_04	1,92753657
simce_dif_lect2m_reg	1,920183822
rend_rep_hom_07	1,91596758
jecd_mat_510_3	1,871769926
ranking_asis_gral_curso_4	1,869431214
rend_tra_hom_05	1,833710416
ide_2m_score1	1,807660381
simce_nalu_lect6b_rbd	1,800463358
mat_rbd_mat_hom_2	1,77826923
asistentes_n_mujeres	1,73423576
ide_4b_score5_4	1,696902491
idps_2m_ind_hv	1,678149283
simce_nalu_soc6b_rbd	1,62878859
simce_dif_mate2m_rbd	1,616399806
rend_si_hom_to	1,61080253
subv_dependencia_Municipal	1,572122702
idps_2m_ind_am	1,537313224
rend_tra_muj_02	1,530644414
ranking_depe2_asis_gral_grado_pro_4	1,523335577
asistencia_3	1,508759391
rend_ret_muj_02	1,449534642
ranking_depe2_asis_gral_grado_pro_3	1,43077951
mat_ue_mat_muj_7	1,401383875
simce_palu_eda_ade_lect2m_rbd	1,400708696
rend_tra_muj_08	1,3713367

jecd_mat_610_4	1,37118471
subv_piso_rural	1,35705161
rend_ret_hom_06	1,35578418
ide_4b_score7	1,299226026
jecd_mat_710_3	1,2328237
mat_ue_cur_sim_01	1,03815773
simce_nalu_mate4b_rbd	1,008111955
simce_prom_mate4b_rbd	0,959836754
rend_prom_asis	0,895208418
docentes_dc_or	0,851032376
asistentes_n_hombres	0,802272201
simce_prom_lect4b_reg	0,748884021
simce_nalu_lect2m_rbd	0,716701176
sep_clasificacion_sep_EMERGENTE	0,686448991
rend_tra_hom_07	0,63557677
rend_apr_hom_02	0,621123982
rend_tra_hom_03	0,384689472
mat_ue_mat_cur_dual_hom_3	0,356906235
jecd_mat_410_4	0,312042933
docentes_dc_oes	0,240463227
rend_rep_hom_06	0,187542404
jecd_mat_610_2	0,158783302
rend_apr_hom_01	0,112402664
simce_prom_soc6b_reg	0,104737841
rend_tra_muj_04	0,08332926
simce_palu_eda_ins_lect2m_rbd	0,028421156

CÓDIGO RELEVANTE

A continuación, se adjunta el código de la implementación de los predictores.

```
import xgboost as xgb
from ..procesador.herramientasutiles import *
#from algoritmo.aprendizaje.evaluacion import *
from sklearn.metrics import mean_squared_error as mse
from .preparacion import cargar_datos_y_etiquetas
import time, os, pickle
from sklearn.metrics import confusion_matrix
from sklearn.ensemble import AdaBoostClassifier, RandomForestClassifier
import tensorflow as tf
from .evaluacion import *
from datetime import datetime
from sklearn import svm
from sklearn.metrics import accuracy_score



from sklearn.metrics import accuracy_score
from random import shuffle
import matplotlib.pyplot as plt


def corregir_etiquetas(y):
    y.replace(to_replace=0, value=-1, inplace=True)
    return y

def SVC(ruta_train, ruta_test, C=0.5, kernel='poly', degree=3, bias=50,n=2000,
gma=1.2,etiqueta='desercion_regular', ruta_modelo_fs = None):
    def polynomial_kernel(x, y, degree, coef):
        k = (np.dot(x, y) + coef) ** degree
        return k

    def gram_matrix_poly(X, Y):
        G = np.zeros((X.shape[0], Y.shape[0]))
        for i, x in enumerate(X):
            for j, y in enumerate(Y):
                G[i, j] = polynomial_kernel(x, y, degree, bias)
        return G


    def g_kernel(u, v, gma):
        kk = np.exp(-gma * (np.linalg.norm(u - v)) ** 2)
        return kk

    def gram_matrix_g(X, Y):
        GG = np.zeros((X.shape[0], Y.shape[0]))
```

```
        for i, u in enumerate(X):
            for j, v in enumerate(Y):
                GG[i, j] = g_kernel(u, v, gma)
        return GG

    if kernel=='poly':
        gram_matrix = gram_matrix_poly
    if kernel=='gauss':
        gram_matrix = gram_matrix_g

    etiquetas = ['desercion_regular', 'desercion_global', 'ultimo_mes']
    x_te, y_te = cargar_datos_y_etiquetas(ruta_test, etiquetas, etiqueta, fillna=True)
    x_tr, y_tr = cargar_datos_y_etiquetas(ruta_train, etiquetas, etiqueta, fillna=True)
    x_tr_te_cols = interseccion_columnas(x_tr.columns.tolist(), x_te.columns.tolist())

    if ruta_modelo_fs != None:
        print('dims antes',len(x_tr_te_cols))
        x_tr_te_cols, imp_values = seleccionar_dimensiones_importantes(x_tr_te_cols,
ruta_modelo=ruta_modelo_fs,
                                                       tipo='xgboost', umbral=0.0001)
        print('dims después',len(x_tr_te_cols))

    y_te = corregir_etiquetas(y_te)
    y_tr = corregir_etiquetas(y_tr).iloc[:n]

    x_te = x_te[x_tr_te_cols]
    x_tr = x_tr[x_tr_te_cols].iloc[:n]
   # y_te = one_hot_encode(y_te)
   # y_tr = one_hot_encode(y_tr)

    scaler_tr = StandardScaler()
    scaler_tr.fit(x_tr)
    x_te = scaler_tr.transform(x_te)
    x_tr = scaler_tr.transform(x_tr)

    print('..')
#    degree, bias = 3, 50  # parametros del kernel
    clf = svm.SVC(kernel=gram_matrix, C=C)  # crea el clasificador
    clf.fit(x_tr, y_tr)  # entrena
    print('Support vectors por clase: ', clf.n_support_)
    y_pr_e = clf.predict(x_te)  # clasifica
    y_pr_t = clf.predict(x_tr)

    c_tr = confusion_matrix(y_tr, aproximar(y_pr_t))
    c_te = confusion_matrix(y_te, aproximar(y_pr_e))

    texto = '{} - recall: {:.3f}% - precision: {:.3f}% - accuracy: {:.3f}% - miss_rate:
{:.3f}% - min pr {:.4f} max{:.4f}'
```

```
    print(texto.format('TRAIN',
                       100 * (c_tr[1][1] / sum(c_tr[1])), 100 * (c_tr[1][1] / (c_tr[0][1] +
c_tr[1][1])),
                       100 * ((c_tr[0][0] + c_tr[1][1]) / (c_tr[0][1] + c_tr[1][0] + c_tr[0][0] +
c_tr[1][1])),
                       100 * (c_tr[1][0] / sum(c_tr[1])), min(y_pr_t), max(y_pr_t)))
    # print()
    print(texto.format('TEST',
                       100 * (c_te[1][1] / sum(c_te[1])), 100 * (c_te[1][1] / (c_te[0][1] +
c_te[1][1])),
                       100 * ((c_te[0][0] + c_te[1][1]) / (c_te[0][1] + c_te[1][0] + c_te[0][0] +
c_te[1][1])),
                       100 * (c_te[1][0] / sum(c_te[1])), min(y_pr_e), max(y_pr_e)))

    print('Confusion matrix TRAIN:\n', c_tr)
    print('Confusion matrix TEST:\n', c_te)


 #   print("Accuracy on test set: ", accuracy_score(y_te, y_pr, normalize=True))

def MLP_2hl_entrenar(ruta_train, ruta_test, learning_rate = 0.001,
    training_epochs = 15,
    batch_size = 100,
    display_step = 1,
    etiqueta ='desercion_regular',
    n_clases = 2,
    n_hidden_1=5000,
    n_hidden_2=5000,
    ruta_modelo_fs = None,
    keepprob=1):

    def one_hot_encode(y, n_classes=2):
        '''
        DESCRIPCION: Función que transforma las clases (variable categórica) en una
codificación numérica.

        INPUT:

        (y)         Vector 1-darray con las clases
        (n_classes)  Cantidad de clases, por defecto 2.

        OUTPUT:
        (encode_y)   Vector n_classes-darray con la codificación para cada ejemplo

        '''
        encode_y = np.zeros((y.shape[0], n_classes), dtype=int)
        for i in range(0, y.shape[0]):
            clase = np.argmax(y[i])
            encode_y[i, y[i]] = 1
```

```
    #  print(encode_y)
      return encode_y

  # Create model
  def next_batch(num, data, labels, n_classes):

      '''
      Return a total of `num` random samples and labels.
      '''
      idx = np.arange(0, len(data))
      np.random.shuffle(idx)
      idx = idx[:num]
      data_shuffle = data[idx]
      labels_shuffle = labels[idx,:]
      labels_shuffle = labels_shuffle.reshape(len(labels_shuffle), n_classes)
      #print(n_classes)

      return data_shuffle, labels_shuffle

  def multilayer_perceptron(x, weights, biases):
      # Hidden fully connected layer with 256 neurons
      layer_1 = tf.nn.relu(tf.add(tf.matmul(x, weights['h1']), biases['b1']))

      # drop_out layer
      drop_out = tf.nn.dropout(layer_1, keep_prob)

      # Hidden fully connected layer with 256 neurons
      layer_2 = tf.nn.relu(tf.add(tf.matmul(drop_out, weights['h2']), biases['b2']))
     # layer_2 = layer_2)

      # Output fully connected layer with a neuron for each class
      out_layer = tf.matmul(layer_2, weights['out']) + biases['out']
      return out_layer

  # Network Parameters
    # 1st layer number of neurons
#  n_hidden_2 = 256  # 2nd layer number of neurons

  etiquetas = ['desercion_regular', 'desercion_global', 'ultimo_mes']
  x_te, y_te = cargar_datos_y_etiquetas(ruta_test, etiquetas, etiqueta, fillna=True)
  x_tr, y_tr = cargar_datos_y_etiquetas(ruta_train, etiquetas, etiqueta, fillna=True)
  x_tr_te_cols = interseccion_columnas(x_tr.columns.tolist(), x_te.columns.tolist())

  if ruta_modelo_fs != None:
      print('dims antes',len(x_tr_te_cols))
      x_tr_te_cols, imp_values = seleccionar_dimensiones_importantes(x_tr_te_cols,
ruta_modelo=ruta_modelo_fs,
                                                     tipo='xgboost', umbral=0.0001)
      print('dims después',len(x_tr_te_cols))
```

```
#if etiqueta != 'ultimo_mes':
 #   y_te = corregir_etiquetas(y_te)
  #  y_tr = corregir_etiquetas(y_tr)


x_te = x_te[x_tr_te_cols]
x_tr = x_tr[x_tr_te_cols]
y_te = one_hot_encode(y_te)
y_tr = one_hot_encode(y_tr)

scaler_tr = StandardScaler()
scaler_tr.fit(x_te)
x_te = scaler_tr.transform(x_te)
x_tr = scaler_tr.transform(x_tr)
# y_tr = tf.convert_to_tensor(y_te)
# y_te = tf.convert_to_tensor(y_tr)

n_input = len(x_tr_te_cols)  #

#if etiqueta == 'desercion_regular':
n_classes = n_clases  # MNIST total classes (0-9 digits)
funcion_perdida = tf.nn.softmax_cross_entropy_with_logits
funcion_pred = tf.nn.softmax

# tf Graph input
X = tf.placeholder("float", [None, n_input])
Y = tf.placeholder("float", [None, n_classes])
keep_prob = tf.placeholder("float")

# Store layers weight & bias
weights = {
    'h1': tf.Variable(tf.random_normal([n_input, n_hidden_1])),
    'h2': tf.Variable(tf.random_normal([n_hidden_1, n_hidden_2])),
    'out': tf.Variable(tf.random_normal([n_hidden_2, n_classes]))
}
biases = {
    'b1': tf.Variable(tf.random_normal([n_hidden_1])),
    'b2': tf.Variable(tf.random_normal([n_hidden_2])),
    'out': tf.Variable(tf.random_normal([n_classes]))
}


# Construct model
logits = multilayer_perceptron(X, weights, biases)

# Define loss and optimizer
optimizer = tf.train.AdamOptimizer(learning_rate=learning_rate)
loss_op = tf.reduce_mean(funcion_perdida(logits=logits, labels=Y))
```

```
    train_op = optimizer.minimize(loss_op)
    # Initializing the variables
    init = tf.global_variables_initializer()

    with tf.Session() as sess:
        sess.run(init)

        # Training cycle
        for epoch in range(training_epochs):
            avg_cost = 0.
            total_batch = int(len(x_tr) / batch_size)
            # Loop over all batches
            for i in range(total_batch):
                batch_x, batch_y = next_batch(batch_size, x_tr, y_tr, n_clases)
                #batch_x = scaler_tr.transform(batch_x)
             #   print(batch_x.mean(0))

             #   batch_x = tf.convert_to_tensor(x.astype(np.float32).values)
              #  batch_y = tf.convert_to_tensor(y.astype(np.float32).values)

                # Run optimization op (backprop) and cost op (to get loss value)
                _, c = sess.run([train_op, loss_op], feed_dict={X: batch_x, Y: batch_y,
keep_prob : keepprob})
                # Compute average loss
                avg_cost += c / total_batch
            # Display logs per epoch step
            if epoch % display_step == 0:
                print("Epoch:", '%04d' % (epoch + 1), "cost={:.9f}".format(avg_cost))
                y_pr_t = funcion_pred(sess.run(logits, feed_dict={X: x_tr,keep_prob :
1})).eval()
                y_pr_e = funcion_pred(sess.run(logits, feed_dict={X: x_te,keep_prob :
1})).eval()

              #  y_pr_t = y_pr_t.reshape(n_classes, len(y_pr_t))[0]
               # y_pr_e = y_pr_e.reshape(n_classes, len(y_pr_e))[0]
                y_pr_t_ = np.argmax(y_pr_t, axis=1)
                y_pr_e_ = np.argmax(y_pr_e, axis=1)
                y_te_ = np.argmax(y_te, axis=1)
                y_tr_ = np.argmax(y_tr, axis=1)
                #  y_te = np.asarray(y_te.values.reshape(len(y_te), 1))

                print_metricas(y_tr=y_tr_,y_te=y_te_,y_pr_t= y_pr_t_,y_pr_e=y_pr_e_)


        print("Optimization Finished!")

        # Test model
        pred = funcion_pred(logits)  # Apply softmax to logits
        correct_prediction = tf.equal(tf.argmax(pred, 1), tf.argmax(Y, 1))
```

```
        # Calculate accuracy
        accuracy = tf.reduce_mean(tf.cast(correct_prediction, "float"))

        y_te =y_te.reshape(len(y_te), n_classes)
        print("Accuracy:", accuracy.eval({X: x_te, Y: y_te, keep_prob : 1}))
    return



def print_metricas(y_tr, y_te, y_pr_t,y_pr_e):
#    print(y_tr, y_pr_t)
    etr = error(y_tr, y_pr_t)
    # imprime pérdida de batch
    c_te = confusion_matrix(y_te, y_pr_e)
    c_tr = confusion_matrix(y_tr, y_pr_t)
    print('Confusion matrix TRAIN:\n', c_tr)
    print('Confusion matrix TEST:\n', c_te)

    texto = '{} loss[{}]: {:.3f} - recall: {:.3f}% - precision: {:.3f}% - accuracy: {:.3f}% -
miss_rate: {:.3f}%'
  #  print('batch - {}'.format(ruta_train.split("/")[-1]))
    #        print()

    recall_tr = 100 * (c_tr[1][1] / sum(c_tr[1]))
    precision_tr = 100 * (c_tr[1][1] / (c_tr[0][1] + c_tr[1][1]))
    accuracy_tr = 100 * ((c_tr[0][0] + c_tr[1][1]) / (c_tr[0][1] + c_tr[1][0] + c_tr[0][0] +
c_tr[1][1]))
    miss_tr = 100 * (c_tr[1][0] / sum(c_tr[1]))

    print(texto.format('TRAIN', 'error', etr, recall_tr, precision_tr, accuracy_tr, miss_tr))
    # print()

    recall_te = 100 * (c_te[1][1] / sum(c_te[1]))
    precision_te = 100 * (c_te[1][1] / (c_te[0][1] + c_te[1][1]))
    accuracy_te = 100 * ((c_te[0][0] + c_te[1][1]) / (c_te[0][1] + c_te[1][0] + c_te[0][0]
+ c_te[1][1]))
    miss_te = 100 * (c_te[1][0] / sum(c_te[1]))
    ete = error(y_te, y_pr_e)
    print(texto.format('TEST', 'error', ete, recall_te, precision_te, accuracy_te, miss_te))

    # imp =
pd.DataFrame({'score':pd.Series(adab.feature_importances_())},index=x_tr_te_cols)
    #    tipos_imp = ['gain', 'cover', 'weight']
    #   imp[t] = , index=x_tr_te_cols)

    # imprime pérdida del modelo final
    #    print(params['perdida_str']+' final: TRAIN {}, TEST {}'.format(i,
params['perdida_f'](y_tr, y_pr), params['perdida_f'](y_te, y_pr)))
    #print('Confusion matrix TRAIN:\n', confusion_matrix(y_tr, aproximar(y_pr_t)))
```

```
    #print('Confusion matrix TEST:\n', confusion_matrix(y_te, aproximar(y_pr_e)))


    return


def randomforest_entrenar(ruta_train, ruta_test, ruta_modelo_fs=None,
etiqueta='desercion_regular', ns_estimators=[10], max_depth=None,
min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0,
max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0,
min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=-1,
random_state=None, verbose=0, warm_start=False, class_weight=None, criterio=
'error'):

    """antideseritin randomforest"""
   # randfor = RandomForestClassifier(n_estimators=10, max_depth = None,
min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0,
max_features ='auto', max_leaf_nodes = None, min_impurity_decrease = 0.0,
min_impurity_split = None, bootstrap = True, oob_score = False, n_jobs = 1, random_state
= None, verbose = 0, warm_start = False, class_weight = None)
    etiquetas = ['desercion_regular', 'desercion_global', 'ultimo_mes']
    x_te, y_te = cargar_datos_y_etiquetas(ruta_test, etiquetas, etiqueta, fillna=True)
    x_tr, y_tr = cargar_datos_y_etiquetas(ruta_train, etiquetas, etiqueta, fillna=True)
    x_tr_te_cols = interseccion_columnas(x_tr.columns.tolist(), x_te.columns.tolist())
    if ruta_modelo_fs !=None:
        x_tr_te_cols, imp_values = seleccionar_dimensiones_importantes(x_tr_te_cols,
ruta_modelo=ruta_modelo_fs, tipo='xgboost', umbral=0.0001)
        auximp = pd.DataFrame({'dimensiones': x_tr_te_cols, 'importancia':imp_values})
        auximp.to_csv('importancia feature selection')
    if etiqueta != 'ultimo_mes':
        y_te = corregir_etiquetas(y_te)
        y_tr = corregir_etiquetas(y_tr)


    randomforests = []
    ls = []
    ns = []
    recalls_tr = []
    recalls_te = []
    precisions_te = []
    precisions_tr = []
    accuracys_te = []
    accuracys_tr = []
    etes = []
    etrs = []
    for n in ns_estimators:
          randfor = RandomForestClassifier(n_estimators=n,
                                           max_depth=max_depth,
                                           min_samples_split=min_samples_split,
                                           min_samples_leaf=min_samples_leaf,
```

```
                                    min_weight_fraction_leaf=min_weight_fraction_leaf,
                                    max_features=max_features,
                                    max_leaf_nodes = max_leaf_nodes,
                                    min_impurity_decrease = min_impurity_decrease,
                                    min_impurity_split = min_impurity_split,
                                    bootstrap = bootstrap,
                                    oob_score = oob_score,
                                    n_jobs = n_jobs,
                                    random_state = random_state,
                                    verbose = verbose,
                                    warm_start = warm_start,
                                    class_weight = class_weight)


        randfor.fit(x_tr[x_tr_te_cols], y_tr)

        y_pr_t = randfor.predict(x_tr[x_tr_te_cols])
        y_pr_e = randfor.predict(x_te[x_tr_te_cols])
        etr = error(y_tr, y_pr_t)

        # imprime pérdida de batch
        c_te = confusion_matrix(y_te, aproximar(y_pr_e))
        c_tr = confusion_matrix(y_tr, aproximar(y_pr_t))
        print('N_estimators {}'.format(n))
        texto = '{} loss[{}]: {:.3f} - recall: {:.3f}% - precision: {:.3f}% - accuracy:
{:.3f}% - miss_rate: {:.3f}% - min {:.4f} max{:.4f}'
        print('batch - {}'.format(ruta_train.split("/")[-1]))
        #       print()

        recall_tr = 100 * (c_tr[1][1] / sum(c_tr[1]))
        precision_tr = 100 * (c_tr[1][1] / (c_tr[0][1] + c_tr[1][1]))
        accuracy_tr = 100 * ((c_tr[0][0] + c_tr[1][1]) / (c_tr[0][1] + c_tr[1][0] +
c_tr[0][0] + c_tr[1][1]))
        miss_tr = 100 * (c_tr[1][0] / sum(c_tr[1]))

        print(texto.format('TRAIN', 'error', etr, recall_tr, precision_tr, accuracy_tr,
miss_tr, min(y_pr_t),
                           max(y_pr_t)))
        # print()

        recall_te = 100 * (c_te[1][1] / sum(c_te[1]))
        precision_te = 100 * (c_te[1][1] / (c_te[0][1] + c_te[1][1]))
        accuracy_te = 100 * ((c_te[0][0] + c_te[1][1]) / (c_te[0][1] + c_te[1][0] +
c_te[0][0] + c_te[1][1]))
        miss_te = 100 * (c_te[1][0] / sum(c_te[1]))
        ete = error(y_te, y_pr_e)

        print(texto.format('TEST', 'error', ete, recall_te, precision_te, accuracy_te,
miss_te, min(y_pr_e),
```

```
                        max(y_pr_e)))

            randomforests.append(randfor)
            ns.append(n)
            recalls_tr.append(recall_tr)
            recalls_te.append(recall_te)
            precisions_te.append(precision_te)
            precisions_tr.append(precision_tr)
            accuracys_te.append(accuracy_te)
            accuracys_tr.append(accuracy_tr)
            etes.append(ete)
            etrs.append(etr)

#           x_tr_te_cols = interseccion_columnas(x_tr.columns.tolist(), x_te.columns.tolist())

            # imp =
pd.DataFrame({'score':pd.Series(adab.feature_importances_())},index=x_tr_te_cols)
            #    tipos_imp = ['gain', 'cover', 'weight']
            #   imp[t] = , index=x_tr_te_cols)

            # imprime pérdida del modelo final
            #    print(params['perdida_str']+' final: TRAIN {}, TEST {}'.format(i,
params['perdida_f'](y_tr, y_pr), params['perdida_f'](y_te, y_pr)))
            print('Confusion matrix TRAIN:\n', confusion_matrix(y_tr, aproximar(y_pr_t)))
            print('Confusion matrix TEST:\n', confusion_matrix(y_te, aproximar(y_pr_e)))

    #  print(model.eva)

    rtr = np.argmax(recalls_tr)
    rte = np.argmax(recalls_te)
    pte = np.argmax(precision_te)
    ptr = np.argmax(precision_tr)
    ate = np.argmax(accuracy_te)
    atr = np.argmax(accuracy_tr)
    ete = np.argmin(etes)
    etr = np.argmin(etrs)

    #criterio = ['error', 'recall', 'precision', 'accuracy']
    ### cuál modelo guardar, elegir
    if criterio == 'recall':
        x = rtr
    if criterio == 'error':
        x = ete
    if criterio == 'precision':
        x = pte
    if criterio == 'accuracy':
        x = ate

    print(
```

```
        'mejor test {}## error: {} - n_estimators: {} - recall: {} - precision: {} - accuracy:
{}'.format(
            criterio, etes[x], ns[x], recalls_te[x], precisions_te[x], accuracys_te[x]))

    stamp = time.time()
    ll = etiqueta

    try:
        os.makedirs('resultados/modelos/predictores/randomforest')
    except:
        pass
    try:
        os.makedirs('resultados/datos/predictores/randomforest')
    except:
        pass

    randomforests[x].columnas_base_ = x_tr_te_cols
    with open('resultados/modelos/predictores/randomforest/randomforest_max-' + criterio
+ '_n-' + str(ns[x]) + '_' + ll + '_' + str(stamp) + '.model' + '.pkl',
              'wb') as output:
        pickle.dump(randomforests[x], output, pickle.HIGHEST_PROTOCOL)

    visualizar_score_feature(x_tr_te_cols, model=randomforests[x], modelo='randomforest',
stamp=stamp, ll=ll, tipos_metricas=['score'], fs=3)

    return randomforests[x]

def adaboost_entrenar(ruta_train, ruta_test,ruta_modelo_fs=None, etiqueta =
'desercion_regular', learning_rates=[1], ns_estimadores=[50], criterio='recall'):
    """antideseritin adaboost"""
    etiquetas = ['desercion_regular', 'desercion_global', 'ultimo_mes']
    x_te, y_te = cargar_datos_y_etiquetas(ruta_test, etiquetas, etiqueta, fillna= True)
    x_tr, y_tr = cargar_datos_y_etiquetas(ruta_train, etiquetas, etiqueta, fillna=True)
    x_tr_te_cols = interseccion_columnas(x_tr.columns.tolist(), x_te.columns.tolist())
    if ruta_modelo_fs !=None:
        x_tr_te_cols, imp_values = seleccionar_dimensiones_importantes(x_tr_te_cols,
ruta_modelo=ruta_modelo_fs, tipo='xgboost', umbral=0.0001)

    if etiqueta != 'ultimo_mes':
        y_te = corregir_etiquetas(y_te)
        y_tr = corregir_etiquetas(y_tr)

    adaboosts = []
    ls = []
    ns = []
    recalls_tr = []
    recalls_te = []
    precisions_te = []
    precisions_tr = []
```

```
    accuracys_te = []
    accuracys_tr = []
    etes = []
    etrs = []
    for l in learning_rates:
      for n in ns_estimadores:
        adab = AdaBoostClassifier(n_estimators=n, learning_rate=l)
        adab.fit(x_tr[x_tr_te_cols], y_tr)

        y_pr_t = adab.predict(x_tr[x_tr_te_cols])
        y_pr_e= adab.predict(x_te[x_tr_te_cols])


        etr = error(y_tr, y_pr_t)

    # imprime pérdida de batch
        c_te = confusion_matrix(y_te, aproximar(y_pr_e))
        c_tr = confusion_matrix(y_tr, aproximar(y_pr_t))
        print('Learning_rate: {}, N_estimators {}'.format(l,n))
        print('batch - {}'.format(ruta_train.split("/")[-1]))
    #       print()

        recall_tr = 100 * (c_tr[1][1] / sum(c_tr[1]))
        precision_tr = 100 * (c_tr[1][1] / (c_tr[0][1] + c_tr[1][1]))
        accuracy_tr = 100 * ((c_tr[0][0] + c_tr[1][1]) / (c_tr[0][1] + c_tr[1][0] + c_tr[0][0]
+ c_tr[1][1]))
        miss_tr = 100 * (c_tr[1][0] / sum(c_tr[1]))

        texto = '{} loss[{}]: {:.3f} - recall: {:.3f}% - precision: {:.3f}% - accuracy: {:.3f}%
- miss_rate: {:.3f}% - min {:.4f} max{:.4f}'
        print('TRAIN', 'error',
                        etr,
                        recall_tr,
                        precision_tr,
                        accuracy_tr,
                        miss_tr,
                        min(y_pr_t),
                        max(y_pr_t))
        print(texto.format('TRAIN',
                        'error',
                        etr,
                        recall_tr,
                        precision_tr,
                        accuracy_tr,
                        miss_tr,
                        min(y_pr_t),
                        max(y_pr_t)))
    # print()
```

```
        recall_te = 100 * (c_te[1][1] / sum(c_te[1]))
        precision_te = 100 * (c_te[1][1] / (c_te[0][1] + c_te[1][1]))
        accuracy_te = 100 * ((c_te[0][0] + c_te[1][1]) / (c_te[0][1] + c_te[1][0] +
c_te[0][0] + c_te[1][1]))
        miss_te = 100 * (c_te[1][0] / sum(c_te[1]))
        ete = error(y_te, y_pr_e)

        print(texto.format('TEST',  'error', ete ,recall_te, precision_te, accuracy_te, miss_te,
min(y_pr_e), max(y_pr_e)))

        adaboosts.append(adab)
        ls.append(l)
        ns.append(n)
        recalls_tr.append(recall_tr)
        recalls_te.append(recall_te)
        precisions_te.append(precision_te)
        precisions_tr.append(precision_tr)
        accuracys_te.append(accuracy_te)
        accuracys_tr.append(accuracy_tr)
        etes.append(ete)
        etrs.append(etr)




    # imp =
pd.DataFrame({'score':pd.Series(adab.feature_importances_())},index=x_tr_te_cols)
#     tipos_imp = ['gain', 'cover', 'weight']
 #    imp[t] = , index=x_tr_te_cols)

# imprime pérdida del modelo final
#     print(params['perdida_str']+' final: TRAIN {}, TEST {}'.format(i,
params['perdida_f'](y_tr, y_pr), params['perdida_f'](y_te, y_pr)))
        print('Confusion matrix TRAIN:\n', confusion_matrix(y_tr, aproximar(y_pr_t)))
        print('Confusion matrix TEST:\n', confusion_matrix(y_te, aproximar(y_pr_e)))

    #  print(model.eva)


    rtr = np.argmax(recalls_tr)
    rte = np.argmax(recalls_te)
    pte = np.argmax(precisions_te)
    ptr = np.argmax(precisions_tr)
    ate = np.argmax(accuracys_te)
    atr = np.argmax(accuracys_tr)
    ete = np.argmax(etes)
    etr = np.argmax(etrs)


    criterio = 'error'#['error','recall','precision','accuracy']
```

```
### cuál modelo guardar, elegir
    x = 0

    if criterio == 'recall':
        x = rtr
    if criterio == 'error':
        x = ete
    if criterio == 'precision':
        x = pte
    if criterio == 'accuracy':
        x = ate


    print('max test {}## error: {} - learning_rate: {} - n_estimators: {} - recall: {} -
precision: {} - accuracy: {}'.format(criterio, etes[x], ls[x],ns[x], recalls_te[x],
precisions_te[x],accuracys_te[x]))

    stamp = time.time()
    ll = etiqueta

    try:
        os.makedirs('resultados/modelos/predictores/adaboost')
    except:
        pass
    try:
        os.makedirs('resultados/datos/predictores/adaboost')
    except:
        pass

    adaboosts[x].columnas_base_ = x_tr_te_cols
    with open('resultados/modelos/predictores/adaboost/adaboost_max-'+criterio+'_lr-
'+str(ls[x])+'_n-'+str(ns[x])+'_' + ll + '_' + str(stamp) + '.model' + '.pkl',          'wb') as
output:
        pickle.dump(adaboosts[x], output, pickle.HIGHEST_PROTOCOL)

    visualizar_score_feature(x_tr_te_cols, model=adaboosts[x],modelo='adaboost',
stamp=stamp, ll=ll, tipos_metricas=['score'], fs=5)

    return adaboosts[x]




def xgboost_iterativo_incremental(lista_rutas_batchs, ruta_test,
ruta_modelo_fs=None,n_mostrar=100,etiqueta = 'desercion_regular', batchs='full',
epochs=1, tipo= 'binlog', l_r=0.1, reg_alpha = 0, reg_lambda= 1, max_depth=6,
n_estimators=500, silent=True, gamma=0,colsample_bytree = 1,subsample =
1,colsample_bylevel=1,l_r_batch_decay =1,earlysr = 50):
    """antideseritin"""
```

```
   etiquetas= ['ultimo_mes', 'desercion_regular', 'desercion_global']

## carga dates de test
   x_te, y_te = cargar_datos_y_etiquetas(ruta_test, etiquetas, etiqueta)

## define número de batchs
   if batchs is 'full':
      batchs = len(lista_rutas_batchs)

## parámetros que definen al algoritmo
   if tipo == 'binlog':
      objetivo = 'binary:logistic'
      eval_metric =  'error'#'logloss'#
      perdida_str = 'ERROR' #'logloss'#
      perdida_f = error

   params = {
      'objective': objetivo,
#       'eval_set': eval_set,
      'eval_metric': eval_metric,
      'learning_rate': l_r,
     # 'updater': 'refresh',
    #  'process_type': 'update',
      'refresh_leaf': True,
      'reg_lambda': reg_lambda,  # L2
      'reg_alpha': reg_alpha,  # L1
       'silent': silent,
       'verbose': False,
      'perdida_str': perdida_str,
      'perdida_f': perdida_f,
      'max_depth': max_depth,
   #    'n_estimators': n_estimators,
      'gamma':gamma,
      'colsample_bytree': colsample_bytree,
      'subsample': subsample,
      'colsample_bylevel':colsample_bylevel

   }

### Inicialización el modelo
   model = None
   # inicio de época
   etiquetas = ['ultimo_mes', 'desercion_regular','desercion_global']
   etiqueta = 'desercion_regular'
   l = list(range(batchs))
   print(lista_rutas_batchs)

   print(params)
      # inicio de batch
```

```
    # shuffle(l)
  M = []
  for j in l:
     x_tr, y_tr = cargar_datos_y_etiquetas(lista_rutas_batchs[j], etiquetas, etiqueta)


  #    if etiqueta != 'ultimo_mes':
   #       y_te = corregir_etiquetas(y_te)
    #      y_tr = corregir_etiquetas(y_tr)

     if j == 0:
        x_tr_te_cols = interseccion_columnas(x_tr.columns.tolist(), x_te.columns.tolist())
       # print(x_tr_te_cols)
        if ruta_modelo_fs != None:
           x_tr_te_cols_ = x_tr_te_cols
           x_tr_te_cols = seleccionar_dimensiones_importantes(ruta_modelo_fs)
           print('....')
 #         with open('pickle.test', 'wb') as f:
#          print(x_tr_te_cols)
 #         pickle.dump(x_tr_te_cols, f)

     for i in range(epochs):
        # carga datos de entrenamiento por batch
       # train_set = [(,)]
        dtrain = xgb.DMatrix(x_tr[x_tr_te_cols],  y_tr)
        dtest = xgb.DMatrix(x_te[x_tr_te_cols], y_te)
        evals_set =  [(dtrain,'train'),(dtest,'eval')]
        evals_result =  {'eval_metric': 'error'}

        # entrena el modelo con los params y x_tr, y_tr, inicializado con el modelo anterior
iterado (inicial model=None).
        model = xgb.train(params, dtrain=dtrain, num_boost_round=n_estimators , evals=
evals_set, evals_result= evals_result, early_stopping_rounds= earlysr,xgb_model=model)

        print(model)

        # predecir con el nuevo modelo entrenado
        y_pr_t = model.predict(xgb.DMatrix(x_tr[x_tr_te_cols]))
        y_pr_e = model.predict(xgb.DMatrix(x_te[x_tr_te_cols]))


        # imprime pérdida de batch
        c_te=confusion_matrix(y_te, aproximar(y_pr_e))
        c_tr=confusion_matrix(y_tr, aproximar(y_pr_t))
        texto= '{} loss[{}]: {:.3f} - recall: {:.3f}% - precision: {:.3f}% - accuracy:
{:.3f}% - miss_rate: {:.3f}% - min pr {:.4f} max{:.4f}'
        print('[epoch {}, batch {}] - {}'.format(i,j,lista_rutas_batchs[j].split("/")[-1] ))
  #       print()
        print(texto.format('TRAIN',params['eval_metric'],params['perdida_f'](y_tr, y_pr_t)
,100*(c_tr[1][1]/sum(c_tr[1])), 100*(c_tr[1][1]/(c_tr[0][1]+c_tr[1][1])),
```

```
100*((c_tr[0][0]+c_tr[1][1])/(c_tr[0][1]+c_tr[1][0]+c_tr[0][0]+c_tr[1][1])),100*(c_tr[1][
0]/sum(c_tr[1])),min(y_pr_t),max(y_pr_t)))
        # print()
         print(texto.format('TEST',params['eval_metric'],params['perdida_f'](y_te,
y_pr_e),100*(c_te[1][1]/sum(c_te[1])), 100*(c_te[1][1]/(c_te[0][1]+c_te[1][1])),
100*((c_te[0][0]+c_te[1][1])/(c_te[0][1]+c_te[1][0]+c_te[0][0]+c_te[1][1])),100*(c_te[1
][0]/sum(c_te[1])),min(y_pr_e),max(y_pr_e)))

        imp = pd.DataFrame(index=x_tr_te_cols)
        tipos_imp = ['gain']
        for t in tipos_imp:
           imp[t] = pd.Series(model.get_score(importance_type=t), index=x_tr_te_cols)

        M.append(imp)

      #del x_tr
      #del y_tr

      if len(l)>1:
        params['learning_rate'] = params['learning_rate']*l_r_batch_decay
        params.update({'learning_rate': params['learning_rate']*l_r_batch_decay,
                'process_type': 'update',
                'updater': 'refresh',
                'refresh_leaf': False})

      # imprime pérdida de época
#       print(params['perdida_str']+' epoch{}: TRAIN {}, TEST {}'.format(i,
params['perdida_f'](y_tr, y_pr), params['perdida_f'](y_te, y_pr)))
   # predice con el modelo final
   print(model.attributes())
   print(model.best_ntree_limit,model.attributes()['best_iteration'])

   y_pr_t = model.predict(xgb.DMatrix(x_tr[x_tr_te_cols]),
ntree_limit=model.best_ntree_limit)
   y_pr_e = model.predict(xgb.DMatrix(x_te[x_tr_te_cols]),
ntree_limit=model.best_ntree_limit)
   c_tr= confusion_matrix(y_tr, aproximar(y_pr_t))
   c_te = confusion_matrix(y_te, aproximar(y_pr_e))
   print(texto.format('bntl. TRAIN', params['eval_metric'], params['perdida_f'](y_tr,
y_pr_t),
                100 * (c_tr[1][1] / sum(c_tr[1])), 100 * (c_tr[1][1] / (c_tr[0][1] +
c_tr[1][1])),
                100 * ((c_tr[0][0] + c_tr[1][1]) / (c_tr[0][1] + c_tr[1][0] + c_tr[0][0] +
c_tr[1][1])),
                100 * (c_tr[1][0] / sum(c_tr[1])), min(y_pr_t), max(y_pr_t)))
   # print()
   print(texto.format('bntl. TEST', params['eval_metric'], params['perdida_f'](y_te,
y_pr_e),
```

```
                100 * (c_te[1][1] / sum(c_te[1])), 100 * (c_te[1][1] / (c_te[0][1] +
c_te[1][1])),
                100 * ((c_te[0][0] + c_te[1][1]) / (c_te[0][1] + c_te[1][0] + c_te[0][0] +
c_te[1][1])),
                100 * (c_te[1][0] / sum(c_te[1])), min(y_pr_e), max(y_pr_e)))

    print('Confusion matrix TRAIN:\n', c_tr)
    print('Confusion matrix TEST:\n', c_te)
    ll = etiqueta

    try:
        os.makedirs('resultados/modelos/predictores/xgboost')
    except:
        pass
    try:
        os.makedirs('resultados/datos/predictores/xgboost')
    except:
        pass
    model.columnas_base_ = x_tr_te_cols
    with open('resultados/modelos/predictores/xgboost/xgboost_iteri_' + ll + '_' +
str(stamp) + '.model' + '.pkl','wb') as output:
        pickle.dump(model, output, pickle.HIGHEST_PROTOCOL)

    visualizar_score_feature(x_tr_te_cols, model,stamp, ll=ll, n_mostrar=n_mostrar)
    with
open('resultados/datos/predictores/xgboost/M_xgboost_iteri_'+ll+'_'+str(stamp)+'.model'
+'.pkl', 'wb') as output:
        pickle.dump(M, output, pickle.HIGHEST_PROTOCOL)

    return model


def xgboost(ruta_modelo, ruta_datos, caso='test', etiqueta = 'desercion_regular'):
    etiquetas = ['ultimo_mes', 'desercion_regular','desercion_global']
    #cargar datos a evaluar
    if caso == 'predic':
        x_te = pd.DataFrame()
        y_te = pd.DataFrame()

  #  eval_set = pd.read_csv(ruta_datos)
    if caso == 'test':
        x_te, y_te = cargar_datos_y_etiquetas(ruta_datos, etiquetas, etiqueta)



    with open(ruta_modelo, 'rb') as input:
        bst = pickle.load(input)

    lista = bst.columnas_base_
```

```
    l = x_te.columns.tolist()

    visualizar_score_feature(lista, bst, stamp='testing', ll='etiqueta**')


    faltantes_dims = [i for i in lista if i not in l ]
    for j in faltantes_dims:
        x_te[j] = np.zeros(len(x_te))


    #predecir
    y_pr_e = bst.predict(xgb.DMatrix(x_te[lista]))

    if caso == 'test':
        c_te= confusion_matrix(y_te, aproximar(y_pr_e))
        print('Confusion matrix:\n', c_te)
        texto = '{} loss: {:.3f} - recall: {:.3f}% - precision: {:.3f}% - accuracy: {:.3f}% -
miss_rate: {:.3f}%'
        print(texto.format('TEST', error(y_te, y_pr_e),
                    100 * (c_te[1][1] / sum(c_te[1])), 100 * (c_te[1][1] / (c_te[0][1] +
c_te[1][1])),
                    100 * ((c_te[0][0] + c_te[1][1]) / (c_te[0][1] + c_te[1][0] + c_te[0][0] +
c_te[1][1])),
                    100 * (c_te[1][0] /sum(c_te[1]))))
    return True

def aprox(n):
    i = round(n)
    return i

def aproximar(l):
    return [aprox(i) for i in l]


def error(y_te, y_pr):
    y_error = [abs(y_te[i] - aprox(y_pr[i])) for i in range(len(y_te))]
    return sum(y_error)
```

A continuación se adjunta el código de la implementación de los agrupadores.

```
from sklearn import mixture
import matplotlib.pyplot as plt
import pickle
import os
import time
import numpy as np
from algoritmo.aprendizaje.preparacion import *
from algoritmo.aprendizaje.evaluacion import seleccionar_dimensiones_importantes
from datetime import datetime

####
def GMM_fit(n_componentes, años = [2016,2017], n_init=6,cv_type='diag', m=2,
ruta_modelo_fs= None):
    ruta = []
    rutad = []
    rutas = rutas_datos('resultados/datos/agrupadores/desertores/')
    #print(rutas)
    for año in años:
      for r in rutas:
        if str(año) in r:
            if 'preprocesada' in r:
                ruta.append(r)
                ruta_n = r
            if 'preprocesada' not in r:
                rutad.append(r)
                ruta_d = r
    # print(ruta)
#    basenormalizada = pd.read_csv(ruta_n)


    bases = []
    bases_ = []
    for i, r in enumerate(ruta):
        basenorm = pd.read_csv(r)
        base_ = pd.read_csv(rutad[i])
        bases.append(basenorm)
        bases_.append(base_)

    columnas = []
    for i, b in enumerate(bases):
        if i == 0:
            columnas = b.columns.tolist()
            columnas_ = bases_[i].columns.tolist()
        c = b.columns.tolist()
        c_ =  bases_[i].columns.tolist()
        columnas = interseccion_columnas(columnas, c)
        columnas_ = interseccion_columnas(columnas_,c_)
```

```
    basenormalizada = pd.DataFrame(columns=columnas)
    base = pd.DataFrame(columns= columnas_)
    for i, b in enumerate(bases):
        basenormalizada = basenormalizada.append(b[columnas], ignore_index=True)
        base = base.append(bases_[i][columnas_], ignore_index=True)

    del basenorm
    del base_
    del bases
    del bases_

    x_tr_te_cols = basenormalizada.columns.tolist()
    if ruta_modelo_fs != None:
            x_tr_te_cols, imp_values = seleccionar_dimensiones_importantes(x_tr_te_cols,
ruta_modelo_fs,
                                                    tipo=ruta_modelo_fs.split('/')[-
1].split('_')[0])
            imp_dim = dict(zip(x_tr_te_cols,imp_values))


    x_tr_te_cols = interseccion_columnas(x_tr_te_cols, columnas)
    imp_values = [imp_dim[i] for i in x_tr_te_cols]
    imp_dic = dict(zip(x_tr_te_cols,imp_values))


    clusterer = mixture.GaussianMixture(n_components=n_componentes, n_init=n_init,
covariance_type=cv_type)

    basenormalizada = basenormalizada[x_tr_te_cols]
    base = base[x_tr_te_cols]
    clusterer.fit(basenormalizada)

    pesos = clusterer.weights_.tolist()
    covarianza = clusterer.covariances_
    precision = clusterer.precisions_
    centros = clusterer.means_
    predicc = clusterer.predict(basenormalizada)

    print('...')
    centros_indices = muestras_centros_indices(n=500,
base=basenormalizada,predicc=predicc, centros=centros, precision=precision,
tipo=cv_type)
 #   print(centros_indices)
    stamp = datetime.now()

    dims_criticas = dimensiones_criticas(covarianza, tipo=cv_type, cols= x_tr_te_cols, m=m,
stamp=stamp)

### GUARDAR ARCHIVO RESUMEN
```

```
  #  datos = pd.read_csv(ruta_d)
    archivo = 'resultados/datos/agrupadores/resumen_GMM_m-
'+str(m)+'_c'+str(n_componentes)+'_v' + cv_type+'_'+str(stamp)+'.xlsx'
    writer = pd.ExcelWriter(archivo)


    mediana_gral=base.median(axis=0)
    media_gral=base.mean(axis=0)
    max_gral=base.max(axis=0)
    min_gral=base.min(axis=0)

    for i in np.arange(len(centros_indices)):
      #  for ci in centros_indices[i]:
          aux = base.iloc[centros_indices[i]].T
          c_x = aux.columns.tolist()
          # print(x)
          #print(list(set([i for i in x if x.count(i) > 1])))


          # aux = pd.DataFrame(index=x)

          indices_cluster = np.argwhere(predicc==i)
          indices_cluster = [i[0] for i in indices_cluster]
          #print(indices_cluster)
          datos_cluster = base.iloc[indices_cluster]


          aux = aux.assign(dim_critica = np.full(len(aux), False, dtype=bool),
importancia_predic= [imp_dic[i] if i in x_tr_te_cols else 0 for i in aux.index])
      #     print(aux.index)
       #    print(dims_criticas)
          for c in base.columns.tolist():
             if c in dims_criticas[i]:
               aux.loc[c,'dim_critica'] = True

          print('1')

          aux = aux.assign(centros_cluster= centros[i],
mediana_1000muesta_centros_cluster= aux.median(axis=1))
          print('2')

          aux =
aux.assign(mediana_cluster=datos_cluster.median(axis=0),media_cluster=datos_cluster.me
an(axis=0),max_cluster=datos_cluster.max(axis=0),min_cluster=datos_cluster.min(axis=0
))
          print('3')
          aux = aux.assign(mediana_gral= mediana_gral,media_gral= media_gral,max_gral
= max_gral,min_gral = min_gral)

          print('4')
```

```
#           aux = aux.assign(centros_cluster = )
            aux = aux.drop(c_x,axis=1)
        ########## ????????



            sn = 'cluster '+str(i)
            aux.to_excel(writer, sheet_name=sn, index=True)
    sn = 'pesos_clusters'
    aux = pd.DataFrame({'cluster':range(0,len(centros_indices)),'pesos': pesos})
    aux.to_excel(writer, sheet_name=sn, index=False)
    writer.save()

    try:
        os.makedirs('resultados/agrupadores')
    except:
        pass

### guardar nuevo parámetro de la clase con la lista de columnas usadas en orden
    clusterer.columnas_base_ = columnas



    with open('resultados/agrupadores/clusterer_GMM_c'+str(n_componentes)+'_v' +
cv_type +'_'+str(stamp)+'.pkl', 'wb') as output:
        pickle.dump(clusterer, output, pickle.HIGHEST_PROTOCOL)
    return True

def muestras_centros_indices(n, base, predicc, centros, precision, tipo):
    uni = np.sort(np.unique(predicc))
    c = []
    for u in range(len(uni)):
        y, = np.where(predicc == u)
        d = []
        for i in y:
            dist_maha = maha_dist(centros[u],base.iloc[i].values, precision[u], tipo=tipo)
            d.append(dist_maha)
           # print('d:',dist_maha)

        aux_ = pd.DataFrame({'y': y.tolist(),'dd': d})
    #    print(aux.columns.tolist())
     #   print(d)
     #   print(aux_.iloc[:20])
        aux_ = aux_.sort_values(by=['dd'])
        aux_.reset_index(inplace=True)
        c.append(aux_['y'].iloc[:n].tolist())
    return c

def GMM_predict(model_ruta, data_ruta):
```

```
    with open(model_ruta, 'rb') as input:
        model = pickle.load(input)

    cols = model.columnas_base_

    data = pd.read_csv(data_ruta)
    y = model.predict(data[cols])
    etiqueta = data_ruta.split('/')[-1] + '_predicciones'
    ruta = '/'.join(data_ruta.split('/')[:-1]) + etiqueta
    y.to_csv(ruta, index=False)
    return True

def dimensiones_criticas(cov, tipo, cols, m, stamp):
    if tipo=='diag':
        dims = []
        for n, i in enumerate(cov):
            inv_var = [1/x for x in i]
            plt.plot(range(0,len(inv_var)),inv_var, '*')
            plt.xlabel('x: dimensiones')
            plt.ylabel('1/var(x)')
            plt.title('dims vs 1/var')

plt.savefig('resultados/visualizaciones/agrupadores/GMs_var_dims_(inv_var)_c'+str(len(co
v))+'_n'+str(n)+'_vm'+tipo+'_'+str(stamp)+'.png', dpi=400)
            plt.close()

            plt.hist((inv_var), density=False, bins=16)
            plt.xlabel('(1/var)')
            plt.ylabel('n')
            plt.title('histograma (1/var)')
            plt.savefig(
                'resultados/visualizaciones/agrupadores/GMs_var_dims_hist_(inv_var)_c' +
str(len(cov)) + '_n' + str(
                    n) + '_vm' + tipo+'_'+str(stamp)+'.png', dpi=400)
            plt.close()

            plt.plot(range(0,len(inv_var)),np.log(inv_var), '*')
            plt.xlabel('x: dimensiones')
            plt.ylabel('log(1/var(x))')
            plt.title('dims vs log(1/var)')

plt.savefig('resultados/visualizaciones/agrupadores/GMs_var_dims_log(inv_var)_c'+str(len(
cov))+'_n'+str(n)+'_vm'+tipo+'_'+str(stamp)+'.png', dpi=400)
            plt.close()

            plt.hist(np.log(inv_var), bins= 16)
            plt.xlabel('log(1/var)')
            plt.ylabel('n')
            plt.title('histograma log(1/var)')
```

```
plt.savefig('resultados/visualizaciones/agrupadores/GMs_var_dims_hist_log(inv_var)_c' +
str(len(cov)) + '_n' + str(n) + '_vm' + tipo+'_'+str(stamp)+'.png', dpi=400)
        plt.close()


        plt.plot(range(0,len(i)),i, '*')
        plt.xlabel('x: dimensiones')
        plt.ylabel('var(x)')
        plt.title('dims vs var(x)')

plt.savefig('resultados/visualizaciones/agrupadores/GMs_var_dims_var_c'+str(len(cov))+'_
n'+str(n)+'_vm'+tipo+'_'+str(stamp)+'.png', dpi=400)
        plt.close()

        plt.plot(range(0,len(i)),np.log(i), '*')
        plt.xlabel('x: dimensiones')
        plt.ylabel('log(var(x))')
        plt.title('dims vs log(var)')

plt.savefig('resultados/visualizaciones/agrupadores/GMs_var_dims_log(var)_c'+str(len(cov)
)+'_n'+str(n)+'_vm'+tipo+'_'+str(stamp)+'.png', dpi=400)
        plt.close()


        plt.hist(i, bins=16)
        plt.xlabel('var')
        plt.ylabel('n')
        plt.title('histograma var')
        plt.savefig('resultados/visualizaciones/agrupadores/GMs_var_dims_hist_var_c' +
str(len(cov)) + '_n' + str(n) + '_vm' + tipo+'_'+str(stamp)+'.png', dpi=400)
        plt.close()

        plt.hist(np.log(i), bins= 16)
        plt.xlabel('log(var)')
        plt.ylabel('n')
        plt.title('histograma log(var)')
        plt.savefig('resultados/visualizaciones/agrupadores/GMs_var_dims_hist_log(var)_c'
+ str(len(cov)) + '_n' + str(n) + '_vm' + tipo+'_'+str(stamp)+'.png', dpi=400)
        plt.close()
        #m = (np.max(inv_var)/2)
        #m = 2
        d = []
        dd = []
        for i, v in enumerate(np.log(inv_var)):
            if v >= m:
                d.append(cols[i])
        dims.append(d)
    #print(dims)
    return dims
```

```
def maha_dist(x,y,precision, tipo):
    w = x - y
    if tipo == 'diag':
        dist = 0
        for i, p in enumerate(precision):
                dist += p*(w[i]**2)
        dist = np.sqrt(dist)

    if tipo == 'full':
        #precision = np.diag(precision)
        dist = np.sqrt(w.T*precision*w)
    return dist
```