Universidad del Desarrollo
Facultad de Ingeniería

# SISTEMA DE RECOMENDACIÓN DE COLEGIOS BASADO EN FACTORIZACIÓN MATRICIAL

Para facilitar la toma de decisiones de los padres acerca de que establecimiento elegir para sus hijos

POR: JULIO ANDRÉS SOTELO PARRAGUEZ

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

DR. CRISTIAN CANDIA

DRA.(c) MELANIE OYARZÚN

Diciembre 2022

SANTIAGO

Esta tesis está dedicada a mi esposa Beatriz Mafucci por su apoyo incondicional.

AGRADECIMIENTO

Agradezco a mis compañeros del Magíster en Data Science, Sebastián Becerra y Oscar Cabrera, con quienes partimos esta idea y decidí concretarla como tema de mi proyecto de grado.

TABLA DE CONTENIDO

## Resumen

Escoger dónde estudiarán nuestros hijos no es una tarea fácil, son muchos los factores que determinan la elección final y en más de algún caso, esto puede causar gran ansiedad y estrés en padres e hijos. Normalmente la decisión de que colegio elegir está basada en la distancia al colegio, más la información subjetiva entregada por un familiar o un conocido. Por lo tanto, existe la necesidad de desarrollar un sistema de recomendación de colegios basado en información objetiva y confiable. En este trabajo, se propone un sistema de recomendación de colegios tanto públicos y privados de la Región Metropolitana, basado en modelo hibrido de factorización matricial, para distintos niveles de enseñanza (1ro básico a 4to medio), haciendo uso de múltiples parámetros extraídos tanto de las bases de datos públicas del ministerio de educación, como del departamento de evaluación, medición y registro educacional. El sistema de recomendación desarrollado permitió entregar una nueva alternativa de colegio, más objetiva, con un mayor puntaje en base a métricas de desempeño, que los seleccionados en por usuarios en la base de datos del sistema de admisión escolar. Esta herramienta será de gran ayuda para los padres que estén en búsqueda de un colegio para sus hijos, con el fin de seleccionar una alternativa más objetiva de selección de colegios, que simplemente basándose en recomendaciones subjetivas, facilitando la toma de decisiones, y evitando el estrés que esto pueda generar.

# 1. Introducción

Escoger dónde estudiarán nuestros hijos no es una tarea fácil, son muchos los factores que determinan la elección final y en más de algún caso, esto puede causar gran ansiedad y estrés en padres e hijos [1-3] más si existe un desconocimiento del proceso [4]. Normalmente la decisión de que colegio elegir está basada en la distancia al colegio, más la información subjetiva entregada por un familiar o un conocido. Sin embargo, la selección de colegios debe considerar otros criterios como, rendimiento, deserción escolar, acceso a la universidad, entre otros factores, que pueden ser claves para los padres a la hora de seleccionar un establecimiento para sus hijos. Debido a esto existe la necesidad de diseñar un sistema de recomendación de colegio que este basado en información cuantitativa, geográfica y estadística de los establecimientos, con la finalidad de ayudar a los padres a tomar una decisión más objetiva que simplemente basándose en recomendaciones subjetivas y de esta forma facilitar la toma de decisiones acerca de que establecimiento elegir.

Aunque existen herramientas para localizar colegios como: (https://localizar.agenciaeducacion.cl/), (https://colegiosenchile.cl/) y (https://www.buscacolegios.cl/), estas no permiten entregar una recomendación objetiva a los padres que están en búsqueda de un colegio, y, por otro lado, solo hacen uso del Sistema de Medición de la Calidad de la Educación (SIMCE) para entregar una información del desempeño del colegio, siendo otros factores igualmente relevantes.

Los algoritmos más comunes de recomendación corresponden a; algoritmos de filtrado colaborativo, algoritmos basados en contenido, y híbridos que mezcla la información de ambos [5-6]. Considerando la idea de generar un sistema de recomendación de colegios, la descripción de cada uno de los algoritmos seria la siguiente:

- Contenido: Las sugerencias de colegios entregadas por este tipo de algoritmo, se basan en características mismas de los colegios, buscando relaciones entre los colegios con características similares. Es un modelo sencillo porque solamente se requiere la información de los colegios para armarlo, pero no toma en consideración la información proveniente del usuario lo que dificulta su mejora.
- Colaborativo: El sistema entrega recomendaciones de colegios procesando las calificaciones/ranquin otorgadas por los usuarios. Los resultados se interpretan por métodos basados en la memoria, encontrando usuarios con perfiles parecidos para dar sugerencias de acuerdo con lo que ellos han comprado o les ha gustado, o en modelos que realizan predicciones como lo haría una máquina cualquiera.
- Híbrido: El sistema procesa tanto las características del colegio como las calificaciones/ranquin realizadas para dar mejores recomendaciones al usuario. El sistema de Netflix basado en factorización matricial es un excelente ejemplo puesto que da sugerencias basadas en lo que las personas ven, sus hábitos de búsqueda y características similares entre las películas.

Para poder aplicar un método de recomendación hibrido, necesitamos disponer de una base de datos que contenga las calificaciones/ranquin de los usuarios. Esta base de

datos puede ser generada agrupando colegios en base a sus características (Similar a un modelo de contenido) o también es posible hacer uso de clúster [7].

Un trabajo similar al que se propone en este proyecto es la tesis desarrollada por Icaran R., [8] donde el autor hace uso de un algoritmo hibrido para recomendar colegios, a partir de los datos del sistema de admisión escolar (SAE), para un año y un nivel especifico de estudio (1ro básico). Sin embargo, esta base de datos tiene múltiples limitantes, como por ejemplo; poca información disponible para entrenamiento, la localización geográfica del alumno puede estar sesgada cuando esta se desconoce y no incluye colegios privados. Adicionalmente, solo se incluyen como características de desempeño la información del SIMCE.

Por lo tanto, a partir de lo descrito en los párrafos anteriores, la hipótesis de este trabajo es la siguiente: El colegio sugerido mediante el desarrollo de un algoritmo híbrido de recomendación, basado en datos cuantitativos disponible en el MINEDUC (Ministerio de Educación) y DEMRE (Departamento de Evaluación, Medición y Registro Educacional), posee un mejor desempeño en comparación a la selección subjetiva que realizan los padres actualmente.

En este proyecto de grado mi objetivo será crear un sistema de recomendación de colegios tanto públicos y privados de la Región Metropolitana, basado en modelo hibrido de factorización matricial, para distintos niveles de enseñanza (1ro básico a 4to medio), haciendo uso de múltiples parámetros extraídos tanto de las bases de datos públicas del Mineduc, como del DEMRE, con el fin de entregar una recomendación objetiva de selección de colegio.

Para corroborar mi hipótesis, este desarrollo se realizará inicialmente con los colegios municipales y particulares subvencionados, con el fin de comparar nuestra recomendación con el ranquin generado por el usuario en el Sistema de Admisión Escolar (SAE) entre los años 2019 al 2020. Al comparar mis resultados con los del SAE espero entregar una nueva alternativa de colegio que las elegidas por el postulante, que este a una distancia menor que 4 kilómetros del hogar, usando parámetros relacionados con el desempeño de los alumnos en el colegio.

Posteriormente se entrenará un nuevo modelo con todos los colegios (municipales, subvencionados y particulares). Adicionalmente, se entregará un análisis estadístico descriptivo de los colegios recomendados, incluyendo las características de desempeño de los alumnos en el colegio, entre otros parámetros descriptivos.

## 2. Trabajo Relacionado

Los sistemas de recomendación se han implementado ampliamente en aplicaciones de Internet para reducir los costos de información [9]. Cuando buscamos una película en Netflix, un algoritmo basado en factorización matricial está detrás de escena tratando de sugerir la mejor película, en base a ciertos criterios de recomendación en base a contenido y colaborativa [10]. Asimismo, en el momento en que nos cansamos de una canción en particular, Spotify recomienda una canción que un algoritmo predice que es adecuada para nosotros [11]. Amazon o Mercado libre hacen lo mismo cuando intenta vendernos artículos que personas como nosotros compran con frecuencia [12]. Así se podrían describir un sinnúmero de ejemplo similares de sistemas de recomendación.

Si nos movemos al contexto de los colegios, el desempeño de los colegios es un parámetro importante en la toma de decisiones de que establecimiento elegir, sin embargo, los padres no siempre cuentan con esta información [4], adicionalmente la distancia al establecimiento [13] y la religión [3] pueden jugar un rol casi tan importante como el desempeño del colegio. En contraste a lo descrito anteriormente, Farias M. [14] y Valentine D [15] mencionan que el estatus socioeconómico, los valores culturales, la presión del ambiente, las expectativas de los padres, la autopercepción y la personalidad del estudiante, se correlacionan con la elección del colegio. Es importante destacar esto último, ya que todas las variables mencionadas corresponden a variables subjetivas, que no necesariamente están relacionadas con el desempeño del colegio.

De acuerdo con la búsqueda realizada casi ningún sistema de recomendación ha sido implementado en un contexto de recomendación de colegios, a excepción de los trabajos desarrollados por Icaran R., [8] quien también menciona la carencia de sistemas de recomendación en este contexto, y el trabajo desarrollado por Limanto S., [16].

El trabajo de Icaran R., [8] estudia los efectos de implementar un sistema de recomendación en el contexto del Sistema de Admisión Escolar (SAE) de Chile. Desarrolló un algoritmo basado en un Modelo Hibrido, utilizando la librería LightFM de Python, para sugerir colegios a los estudiantes. El autor hace uso de la base de datos SAE del 2019, y un solo nivel de enseñanza que es primero básico. Los resultados propuestos sugieren que los impactos de un sistema de asignaciones como el usado en el SAE pueden ser contradictorios y poco beneficiosos para algunos estudiantes. El trabajo desarrollado

por Limanto S., [16], habla del desarrollo de una aplicación que permite ayudar a los padres a elegir la mejor escuela primaria de acuerdo con ciertos criterios establecidos. La recomendación se determina utilizando el método Analytical Hierarchy Process (AHP). Con base en los resultados de la validación de este trabajo, más del 96 % de los encuestados afirmó que la aplicación podría proporcionar información detallada sobre el colegio lo que permite poder realizar una elección más objetiva.

Como bien se describió anteriormente solo 2 trabajos están basados en el desarrollo de sistemas de recomendación de colegios, es debido a esto que existe una necesidad de generar mayor desarrollo en esta área.

# 3. Hipótesis y Objetivos

### Hipótesis

El colegio sugerido mediante el desarrollo de un algoritmo híbrido de recomendación, basado en datos cuantitativos disponible en el MINEDUC (Ministerio de Educación) y DEMRE (Departamento de Evaluación, Medición y Registro Educacional), posee un mejor desempeño en comparación a la selección subjetiva que realizan los padres actualmente.

### Objetivo General

Crear un sistema de recomendación de colegios tanto públicos y privados de la Región Metropolitana, basado en modelo hibrido de factorización matricial, para distintos niveles de enseñanza (1ro básico a 4to medio), haciendo uso de múltiples parámetros

extraídos tanto de las bases de datos públicas del Mineduc, como del DEMRE, con el fin de entregar una recomendación objetiva de selección de colegio.

**Objetivos Específicos**

a. Selección y limpieza de las bases de datos disponibles en el MINEDUC (Directorio de colegios, Rendimiento por estudiantes, SAE, SIMCE) y del DEMRE (Matrícula universitaria).
b. Cuantificación de características relevantes para cada uno de los colegios de la región metropolitana, a partir de las bases de datos previamente limpiadas.
c. Desarrollo de algoritmo en base a contenido, para poder generar una base de datos ficticia de selección de colegios, que servirá de entrada para el sistema de recomendación.
d. Desarrollo de algoritmo de recomendación basado en factorización matricial, utilizando librerías de TensorFlow en Python.
e. Validación del algoritmo desarrollado, con la selección de colegios proporcionada por el SAE, para 120 postulantes de 1ro a 4to medio, 10 estudiantes por nivel.
f. Programar visualizaciones geográficas con los resultados de los algoritmos desarrollados.

# 4. Datos y Metodología

## 4.1. Datos

Para llevar a cabo este proyecto te utilizaron 5 bases de datos distintas, 4 relacionadas al MINEDUC (Directorio de colegios, Rendimiento por estudiantes, SAE, SIMCE) y una relacionada al DEMRE (Matrícula universitaria). A continuación, se describirá cada una de las bases de datos utilizadas.

- *Directorio de colegios – MINEDUC*: Esta base de datos corresponde al registro del directorio oficial de establecimientos educacionales a nivel nacional del año 2021. Donde se registran 16498 establecimientos, de los cuales 11285 están en funcionamiento con matrícula. De esta base de datos vamos a considerar solo los establecimientos de la región metropolitana que estén en funcionamiento o estén autorizados de funcionar sin matricula, y cuyas coordenadas geográficas (Latitud - Longitud) estén disponibles. Como resultado final trabajaremos con 2049 colegios presentes en la región metropolitana.
- *Rendimiento por estudiantes – MINEDUC*: Esta base de datos corresponde al registro de rendimiento escolar entre los años 2004 al 2020 por estudiante registrado en las bases de datos publicas del MINEDUC por MRUN (Digito identificador del estudiante). De esta base de datos se considerarán solo los estudiantes de los colegios de la región metropolitana disponibles en la base de datos *Directorio de colegios* previamente filtrada. A partir de esta base de datos podremos extraer información del rendimiento de los alumnos, nivel de enseñanza

(1ro a 4to) que imparte el colegio, genero del alumno, porcentaje de asistencia, y deserción al cierre del año escolar.

- *Sistema de Medición de la Calidad de la Educación (SIMCE) – MINEDUC*: Esta base de datos corresponde a la tabla de puntajes promedio por establecimiento del proceso SIMCE entre los años 2012 a 2018. De esta base de datos se considerarán solo los establecimientos de la región metropolitana disponibles en la base de datos *Directorio de colegios* previamente filtrada. A partir de esta base de datos podremos extraer información del puntaje promedio en las pruebas de lectura y matemática para niveles de enseñanza 4 básico y 2do medio. La decisión de considerar solo estos dos niveles es a que ambos corresponden a la mitad de cada uno de los ciclos formativos del estudiante (Básica y Media).
- *Matrícula universitaria – DEMRE*: Esta base de datos nos muestra a los alumnos matriculados en la educación superior entre los años 2005 al 2020, cada alumno está debidamente identificado con digito identificador (MRUN) registrado en las bases de datos públicas del MINEDUC. Adicionalmente tenemos información de la carrera y universidad a la cual postula, puntaje ponderado del alumno. De esta base de datos se considerarán solo los alumnos que se hayan graduado de los establecimientos educacionales de la región metropolitana disponibles en la base de datos *Directorio de colegios* previamente filtrada.
- *Sistema de Admisión Escolar (SAE) – MINEDUC*: Esta base de datos corresponde al registro del Sistema de Admisión Escolar entre los años 2019 al 2021, por postulante (con digito identificador MRUN) registrado en las bases de datos

públicas del MINEDUC. Esta base de datos posee el orden de las preferencias de los postulantes tanto a colegios municipales como subvencionados, junto con la ubicación geográfica del postulante y el nivel de enseñanza al cual postula. Esto nos permitirá poder sugerir una nueva alternativa de colegio, de acuerdo con las preferencias del postulante. De esta base de datos se considerarán solo los establecimientos de la región metropolitana disponibles en la base de datos *Directorio de colegios* previamente filtrada.

## 4.2. Metodología

**Limpieza y procesamiento de las bases de datos disponibles.**

***Directorio de colegios – MINEDUC*:** Se realizo el siguiente procedimiento en serie para poder limpiar esta base de datos:

I. Se filtro por código de región (COD_REG_RBD = 13), considerando solo la región metropolitana (13).
II. Se conservaron solo los establecimientos con código de dependencia (COD_DEPE2 = 1, 2, 3) correspondientes a colegios municipales (1), particular subvencionado (2) y particular pagado (3).
III. Se conservaron solo los establecimientos con un estado de establecimiento (ESTADO_ESTAB = 1, 4), que corresponde a colegios que están funcionando (1) y que están autorizados, pero sin matricula (4).
IV. Se conservaron solo los establecimientos con código de enseñanza (ENS_01 = 10, 110, 310, 410, 510, 610, 710, 810, 910), correspondientes a colegios con educación

parvularia incluidos (10), este parámetro se incluyo debido a que colegios importantes de la región metropolitana (sobre todo privados) estaban en esta categoría. Colegios con enseñanza básica (110), colegios con enseñanza media humanista-científico niños y jóvenes (310), colegios con enseñanza media técnico-profesional comercial niños y jóvenes (410) , colegios con enseñanza media técnico-profesional industrial niños y jóvenes (510), colegios con enseñanza media técnico-profesional técnica niños y jóvenes (610), colegios con enseñanza media técnico-profesional agrícola niños y jóvenes (710), colegios con enseñanza media técnico-profesional marítima niños y jóvenes (810), colegios con enseñanza media artística niños y jóvenes (910).

V. Esta base de datos posteriormente fue guardada conservando solo las siguientes columnas:

- RBD = Rol base de datos del establecimiento.
- DGV_RBD = Dígito Verificador del RBD.
- NOM_RBD = Nombre del Establecimiento.
- COD_REG_RBD = Código de región en que se ubica el establecimiento.
- COD_DEPE2 = Código de Dependencia del Establecimiento (agrupado).
- RURAL_RBD = Índice de ruralidad del establecimiento.
- LATITUD = Coordenada de Latitud para el establecimiento.
- LONGITUD = Coordenada de Longitud para el establecimiento.
- ENS_01 = Código de enseñanza 01.
- MATRICULA = Establecimiento con matrícula al 30 de abril.

- ESTADO_ESTAB = Estado del establecimiento.
- ORI_RELIGIOSA = Orientación religiosa del establecimiento.
- PAGO_MATRICULA = Pago de matrícula en el establecimiento.
- PAGO_MENSUAL = Pago mensual en el establecimiento.

VI. Finalmente, errores ortográficos que no pudieron ser detectados por líneas de comando tuvieron que ser corregidos a mano, al igual que la corrección de la latitud y longitud de algunos establecimientos.

***Rendimiento por estudiantes – MINEDUC*:** Se realizo el siguiente procedimiento en serie para poder limpiar esta base de datos:

I. Se filtro por código de región (COD_REG_RBD =13), considerando solo la región metropolitana, para cada una de las bases de datos de rendimiento desde el 2004 al 2020.

II. Para cada uno de los establecimientos (RBD), en las bases de datos del 2004 al 2020, y nivel de enseñanza (COD_ENSE2) (1ro básico a 4to medio) extraemos una serie de parámetros que serán descritos a continuación:
    - Promedio de notas general entre los años 2004 al 2020 (PROM_GRAL), expresado en notas de 1 a 7.
    - Porcentaje de asistencia promedio entre los años 2004 al 2020 (ASISTENCIA), expresado en porcentaje de 0 a 100%.

- Porcentaje de alumnos aprobados (promovidos), entre los años 2004 al 2020, expresado en porcentaje de 0 a 100%. Para eso utilizamos la variable (SIT_FIN_R) correspondiente a la situación de promoción del alumno al cierre del año escolar. Este parámetro se calcula de la siguiente manera:

$$PA\% = \left(\frac{P - R}{P + R}\right) 100\% \quad (1)$$

  Donde $P$ corresponde a los alumnos promovidos y $R$ corresponde a los alumnos reprobados.

- Porcentaje de no deserción escolar, calculado utilizando la variable (SIT_FIN_R), entre los años 2004 al 2020, expresado en porcentaje de 0 a 100%. Este parámetro se calcula de la siguiente manera:

$$PR\% = \left(\frac{P + R - Y - T}{P + R + Y + T}\right) 100\% \quad (2)$$

  Donde $P$ corresponde a los alumnos promovidos, $R$ corresponde a los alumnos reprobados, $Y$ a los alumnos retirados y $T$ a los alumnos trasladados.

- Pendiente de la regresión lineal, del promedio de notas (PROM_GRAL) anual entre los años 2004 al 2020.

III. Para cada uno de los establecimientos (RBD), del año 2020, y nivel de enseñanza (COD_ENSE2) (1ro básico a 4to medio) extraemos los siguientes parámetros:

- Cantidad de hombres (GEN_ALU=1).
- Cantidad de mujeres (GEN_ALU=2).

- Porcentaje de hombres con respecto al total (hombres y mujeres) para cada uno de los niveles de enseñanza.
- Cantidad de alumnos para cada uno de los niveles de enseñanza.
- Cantidad de cursos, utilizando la variable correspondiente a la letra del curso (LET_CUR), si existe mas de una letra es que hay mas de un curso para ese nivel de enseñanza.

IV. Posteriormente se realizó una copia de la base de datos de directorio, por cada uno de los niveles de enseñanza (1ro básico a 4to medio), para incluir cada uno de los siguientes parámetros por nivel de enseñanza:

- PROM = Promedio de notas general entre los años 2004 al 2020.
- ASIS = Porcentaje de asistencia promedio entre los años 2004 al 2020.
- APRB = Porcentaje de alumnos aprobados entre los años 2004 al 2020.
- NOND = Porcentaje de no deserción escolar entre los años 2004 al 2020.
- PEND = Pendiente de la regresión lineal del promedio de notas entre los años 2004 al 2020.
- PHOM = Porcentaje de hombres vs total, para el año 2020.
- CALU = Cantidad de alumnos, para el año 2020.
- CCUR = Cantidad de cursos, para el año 2020.
- CHOM = Cantidad de hombres, para el año 2020.
- CMUJ = Cantidad de mujeres, para el año 2020.

V. Finalmente, para cada uno de los directorios creados, solo se conservan los establecimientos que contengan el nivel de enseñanza correspondientes a la base

de datos de directorio creada. Por ejemplo, el Instituto Nacional no estará contenido en esta base de datos correspondiente a 1ro básico ya que parte desde 7mo básico.

***Sistema de Medición de la Calidad de la Educación (SIMCE) – MINEDUC***: Se realizo el siguiente procedimiento en serie para poder limpiar esta base de datos:

I. Se filtro por código de región (COD_REG_RBD =13), considerando solo la región metropolitana, para cada una de las bases de datos del SIMCE desde el 2012 al 2018, para los cursos de 4to Básico y 2do Medio.
II. Se calculo el puntaje promedio de la prueba de lectura, entre los años 2012 al 2018, para los cursos de 4to Básico (PROM_LECT4B_RBD) y 2do Medio (PROM_LECT2M_RBD).
III. Se calculo el puntaje promedio de la prueba de matemáticas, entre los años 2012 al 2018, para los cursos de 4to Básico (PROM_MATE4B_RBD) y 2do Medio (PROM_ MATE2M_RBD).
IV. Pendiente de la regresión lineal, del promedio de la prueba de lectura, entre los años 2012 al 2018, para los cursos de 4to Básico (PROM_LECT4B_RBD) y 2do Medio (PROM_LECT2M_RBD)
V. Pendiente de la regresión lineal, del promedio de la prueba de matemáticas, entre los años 2012 al 2018, para los cursos de 4to Básico (PROM_MATE4B_RBD) y 2do Medio (PROM_ MATE2M_RBD).

VI. Posteriormente para las bases de datos de directorio entre 1ro a 8vo básico, incluimos los siguientes parámetros:

- SIMCE-LEC: Promedio de la prueba de lectura SIMCE, entre los años 2012 al 2018, para 4to Básico.
- SIMCE-MAT: Promedio de la prueba de matemática SIMCE, entre los años 2012 al 2018, para 4to Básico.
- PEND-LEC: Pendiente de la regresión lineal del promedio de la prueba de lectura SIMCE, entre los años 2012 al 2018, para 4to Básico.
- PEND-MAT: Pendiente de la regresión lineal del promedio de la prueba de matemática SIMCE, entre los años 2012 al 2018, para 4to Básico.

VII. Posteriormente para las bases de datos de directorio entre 1ro a 4to medio, incluimos los siguientes parámetros:

- SIMCE-LEC: Promedio de la prueba de lectura SIMCE, entre los años 2012 al 2018, para 2do Medio.
- SIMCE-MAT: Promedio de la prueba de matemática SIMCE, entre los años 2012 al 2018, para 2do Medio.
- PEND-LEC: Pendiente de la regresión lineal del promedio de la prueba de lectura SIMCE, entre los años 2012 al 2018, para 2do Medio.
- PEND-MAT: Pendiente de la regresión lineal del promedio de la prueba de matemática SIMCE, entre los años 2012 al 2018, para 2do Medio.

***Matrícula universitaria – DEMRE***: Se realizo el siguiente procedimiento en serie para poder limpiar esta base de datos:

I. Se realizó la unión por MRUN entre las bases de datos de rendimiento y matricula universitaria, para años consecutivos. Por ejemplo, rendimientos 2004 con matricula universitaria 2005. Con el fin de saber que estudiantes de 4to medio se matricularon en la universidad al año siguiente de graduarse. Una vez, obtenidos estos registros para cada establecimiento que posee 4to medio, calculamos lo siguientes parámetros:
    - Promedio de puntaje obtenido al ingresar a la universidad entre los años 2005 al 2020.
    - Promedio de matriculados en la universidad, por curso entre los años 2005 al 2020.
    - Porcentaje de matriculados con respecto a la cantidad de alumnos por curso en cuarto medio, entre los años 2005 y 2020.

II. Se realizó la unión por MRUN entre las bases de datos de rendimiento y matrícula universitaria, para tres años consecutivos. Por ejemplo, rendimientos 2004 con matrícula universitaria 2005, y rendimientos 2004 con matrícula universitaria 2006. Con el fin de saber que estudiantes de 4to medio se matricularon en la universidad al año siguiente de graduarse, y posteriormente se matricularon nuevamente al año subsiguiente. Una vez, obtenidos estos registros para cada establecimiento que posee 4to medio, calculamos lo siguientes parámetros:

- Porcentaje de alumnos que no desertan al primer año de ingresar a la universidad entre los años 2005 al 2020. Por lo tanto, si este parámetro es 100%, esto quiere decir que ningún alumno de los que ingreso a la universidad el año anterior, se matriculo en otra universidad.

III. Posteriormente para las bases de datos de directorio entre 1ro a 4to medio, incluimos los siguientes parámetros:

- PROM-PUNT: Promedio de puntaje PSU obtenido, entre los años 2005 al 2020.
- PROM-POST: Promedio de postulantes matriculados en la universidad, entre los años 2005 al 2020.
- PORC-MATR: Porcentaje de matriculados con respecto a la cantidad de alumnos por curso en cuarto medio, entre los años 2005 y 2020.
- DE-F-YEAR: Porcentaje de alumnos que no desertan al primer año de ingresar a la universidad entre los años 2005 al 2020.

Una vez limpiadas las bases de datos de rendimiento escolar, SIMCE y matrícula universitaria, se unieron todas las bases de datos de directorio por nivel de enseñanza, en una sola gran base de dato (df_directorio_all.csv), donde se incluyo una columna denominada ID_NIVEL, que corresponde al nivel se enseñanza de 1ro básico a 4to medio (con valores que van del 1 a 12). Adicionalmente, se revisó nuevamente toda la base de datos de directorio y si el establecimiento no posee alumnos el colegio se descarta, ya que puede ser un error en el registro de las bases de datos de rendimiento, que fue de donde se extrajo esta información.

Por otro lado, se encontró que algunos colegios poseían el mismo nombre, pero distinto RBD, es debido a esto que incluimos en los nombres repetidos un número consecutivo para poder diferenciar el colegio por nombre (NOM_RBD), no solo por RBD. En la tabla 1 del ANEXO, se describe cada una de las columnas de la base de datos final de directorio utilizada en este proyecto.

***Sistema de Admisión Escolar (SAE) – MINEDUC***: Se realizo el siguiente procedimiento en serie para poder limpiar esta base de datos:

I. Se realizó la unión de las postulaciones SAE de los años 2019 al 2021, con el fin de aumentar la base de datos. Se incluyeron solo las postulaciones que corresponden a los RBD contenidos en la base de datos de directorio de la RM, que se limpió previamente.

II. Adicionalmente se calculo la distancia de la dirección reportada por el postulante al establecimiento al cual esta postulando, utilizando la fórmula de Haversine descrita a continuación,

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (3)$$

Donde $r$ es el radio de la tierra 6371 kilómetros, $\varphi_1$ y $\varphi_2$ con las coordenadas de latitud de la coordenada 1 y la coordenada 2 respectivamente, $\lambda_1$ y $\lambda_2$ con las coordenadas de longitud de la coordenada 1 y la coordenada 2 respectivamente.

En la tabla 2 del ANEXO, se describe cada una de las columnas de la base de datos final de SAE utilizada en este proyecto para poder validar el algoritmo propuesto.

**Desarrollo de algoritmo en base a contenido.**

Uno de los requisitos para que el algoritmo de recomendación hibrido de factorización matricial funcione de forma adecuada, es el uso de una base de datos de recomendación, donde se entregue un ranquin para cada uno de los colegios que estén a una cierta distancia. Para esto no podemos considerar la base de datos del SAE debido a que es una selección bastante subjetiva de los colegios, donde el parámetro más relevante es la distancia al colegio, sin considerar que tan bueno es este. Un ejemplo de como debe ser la base de datos de recomendación de colegios se muestra en la Figura 1.

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 |
|---|---|---|---|---|---|---|---|---|
| P1 | 1 | 2 | - | 5 | 2 | - | 1 | - |
| P2 | - | 2 | 4 | 5 | - | 3 | - | 3 |
| P3 | 5 | 2 | - | - | 2 | 3 | 1 | - |
| P4 | 5 | - | 4 | 5 | 2 | 3 | - | - |
| P5 | 2 | 1 | 4 | 3 | 1 | - | - | 4 |
| P6 | 5 | - | - | 5 | 2 | 3 | 3 | - |
| P7 | - | 3 | 3 | - | 3 | 1 | 2 | - |

**Figura 1:** Ejemplo de una base de datos de recomendación de colegios, donde C1 al C8 corresponden a 8 colegios, y P1 al P7 corresponde a 7 postulantes. Los valores dentro de la tabla corresponden a los distintos ranquin (1 a 5) asignados por los postulantes a cada uno de los colegios.

Adicionalmente, debemos tener en consideración que esta base de datos de recomendaciones se generará cada ves que ejecutemos nuestro algoritmo, ya que para cada consulta que realice el usuario, tendremos que generar una base de datos de recomendación, que contendrá solo los colegios que estén a una cierta distancia (kilómetros) sugerida por el usuario. En la Figura 2, se muestra un ejemplo de una dirección consultada en la comuna de Las Condes.

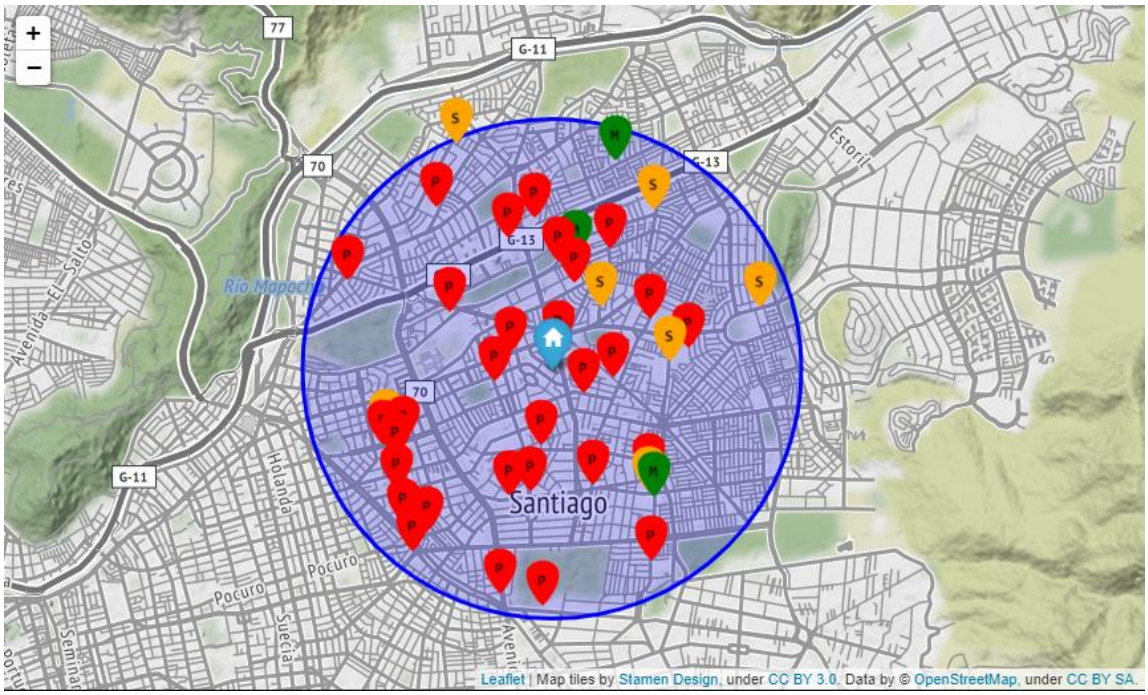



**Figura 2:** Ejemplo de una dirección consultada en la comuna de Las Condes, calle Rosario Sur. En azul se muestra el respectivo rango de distancia consultado correspondiente a 4 kilómetros y los colegios que se deben ranquear dentro de esta distancia, entre particulares (P), subvencionados (S) y municipales (M), dependiendo de la consulta del usuario.

Para comprender adecuadamente los siguientes párrafos, consideraremos al **usuario,** como la persona que realizará la consulta en nuestro sistema, y el **postulante** corresponderá al dato simulado que crearemos para generar nuestra base de datos de recomendación. Para poder generar esta base de datos vamos a utilizar los siguientes parámetros de la base de datos de directorio, para cada nivel de enseñanza (1ro básico a 4to medio).

- PROM: Promedio general por curso, entre los años 2004 al 2020.
- ASIS: Porcentaje de asistencia por curso, entre los años 2004 al 2020
- APRB: Porcentaje de alumnos aprobados respecto al total por curso, entre los años 2004 al 2020
- NOND: Porcentaje de alumnos que no fueron retirados o trasladados con respecto al total de alumnos por curso, entre los años 2004 al 2020
- PEND: Pendiente de la regresión lineal del promedio general por nivel, entre los años 2004 al 2020
- SIMCE-LEC: Valor promedio del puntaje obtenido en el SIMCE lenguaje entre los años 2012 al 2018
- SIMCE-MAT: Valor promedio del puntaje obtenido en el SIMCE matemática entre los años 2012 al 2018
- PEND-LEC: Pendiente de la regresión lineal del valor promedio obtenido en el SIMCE de lenguaje, entre los años 2012 a 2018
- PEND-MAT: Pendiente de la regresión lineal del valor promedio obtenido en el SIMCE de matemática, entre los años 2012 a 2018

- PROM-PUNT: Puntaje promedio obtenido por los alumnos, que se matricularon en la universidad, entre los años 2005 y 2020, solo para los colegios que tengan registro de poseer 4to medio.
- PROM-POST: Número de alumnos promedio que se matriculan en la universidad, entre los años 2005 y 2020, solo para los colegios que tengan registro de poseer 4to medio.
- DE-F-YEAR: Porcentaje de alumnos que no desertaron al primer año de universidad, esto quiere decir alumnos que no ingresaron al año siguiente a otra universidad.
- PORC-MATR: Porcentaje de matriculados con respecto a la cantidad de alumnos por curso en cuarto medio, entre los años 2005 y 2020.

Para generar los ranquin de la base de datos de recomendación, se generaron de forma aleatoria 1000 coordenadas en un rango de distancia de ½ de la distancia consultada por el usuario, que lo asignaremos como **½ U**. Esto quiere decir si el usuario consulto por un rango de 4 kilómetros, el número de coordenadas aleatorias estarán generadas a una distancia de 2 kilómetros, de la dirección consultada por el usuario. Cada una de estas coordenadas simulara la localización geográfica de un postulante, el cual va a poseer el mismo nivel de enseñanza y genero que el usuario que está realizando la consulta. Posteriormente para cada uno de estos postulantes generados de forma aleatoria, se ranquean los colegios que están a una distancia de ½ U de su ubicación. Asignando asi un numero de recomendación de 1 a 5 a los colegios encontrados en ese rango, y cuyas

caracteristicas coincidan con las consultadas por el usuario (p.ej. nivel de enseñanza, tipo de colegio, orientacion religiosa, etc.). El ranquin de 1 a 5 se asignara de acuerdo con la siguiente ecuación que engloba los parametros descritos anteriormente:

$$S(x) = 0.1 * \left(P_r(x) + A_s(x) + A_p(x) + N_d(x) + P_n(x)\right) + \quad (4)$$
$$0.0625 * (S_l(x) + S_m(x) + P_l(x) + P_m(x)) +$$
$$0.0625 * (P_{pu}(x) + P_{po}(x) + D_y(x) + P_{ma}(x))$$

Donde cada una de las variables al interior de la ecuación anterior estan normalizadas entre (0 y 1), y $x$ corresponde a un establecimiento en particular, las variables de la ecuación corresponden a: $P_r$ = PROM, $A_s$ = ASIS, $A_p$ = APRB, $N_d$ = NOND, $P_n$ = PEND, $S_l$ = SIMCE-LEC, $S_m$ = SIMCE-MAT, $P_l$ = PEND-LEC, $P_m$ = PEND-MAT, $P_{pu}$ = PROM-PUNT, $P_{po}$ = PROM-POST, $D_y$ = DE-F-YEAR y $P_{ma}$ = PORC-MATR. Los ultimos 4 terminos solo se incluyen en los niveles de enseñanza de 1ro a 4to medio.

Para cada uno de los 1000 postulantes simulados, el ranquin asignado por el postulante a cada uno de los colegios que estan a una distancia ½ U de su ubicación, estaran dado por los siguientes rangos.

$$R(x) = 1 \quad si \quad \min(S_{tot}) < S(x) < 0.2 * \max(S_{tot}) \quad (5)$$
$$R(x) = 2 \quad si \quad 0.2 * \max(S_{tot}) < S(x) < 0.4 * \max(S_{tot})$$
$$R(x) = 3 \quad si \quad 0.4 * \max(S_{tot}) < S(x) < 0.6 * \max(S_{tot})$$
$$R(x) = 4 \quad si \quad 0.6 * \max(S_{tot}) < S(x) < 0.8 * \max(S_{tot})$$
$$R(x) = 5 \quad si \quad 0.8 * \max(S_{tot}) < S(x) < \max(S_{tot})$$

Donde $R(x)$ es el ranquin asignado al colegio $x$ y $S_{tot}$ es un vector que contine la informacion de todos los $S(x)$, que se encuentran a una distancia ½ U del postulante. En la Figura 3 se muestra un ejemplo de las distancias y rangos de busquedas asignados a los postulantes.

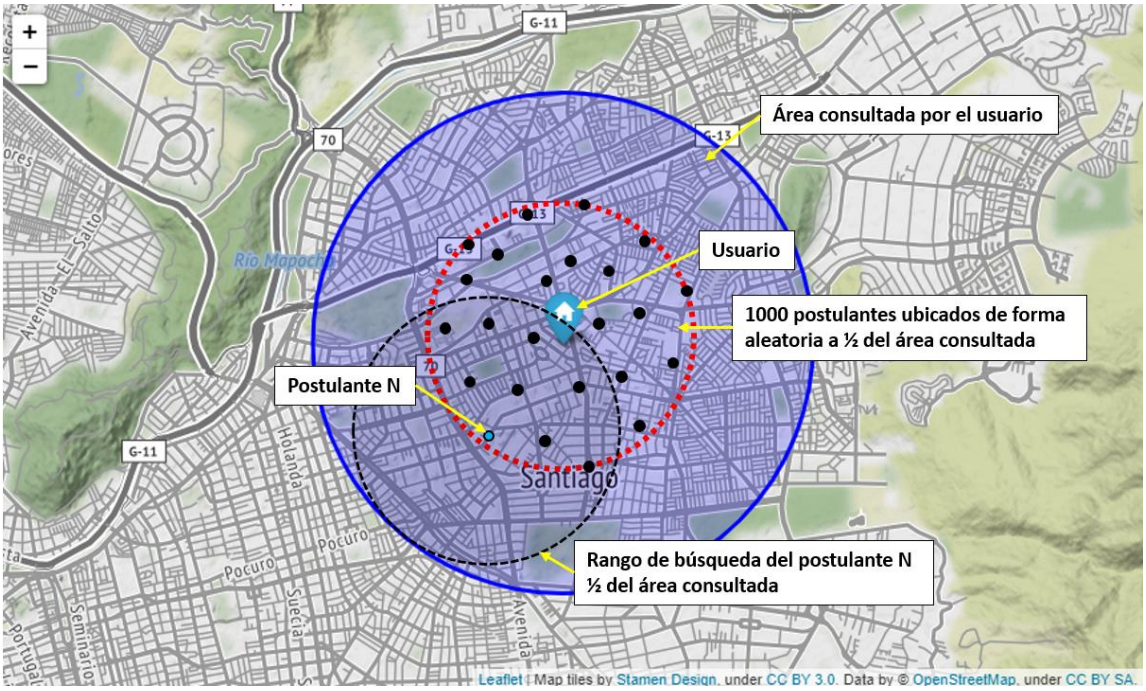



**Figura 3:** Generación aleatoria de postulantes en un rango de ½ U. Adicionalmente se muestra el rango de búsqueda de un postulante N, que también coincide con ½ U. Dentro de este rango de búsqueda el postulante N, ranqueara los establecimientos que se encuentren en ese rango, con un ranquin de 1 a 5, a partir de las ecuaciones (4) y (5).

Una ves ranqueados los colegios, para cada uno de los postulantes simulados, debemos incluir una fila a nuestra matriz de recomendación que corresponda al usuario que realiza la consulta, como se muestra en la Figura 4. Esto es debido a que la

información del usuario también debe estar contenida dentro de esta matriz de recomendación, para que el algoritmo pueda realizar la recomendación de forma adecuada.

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 |
|---|---|---|---|---|---|---|---|---|
| U | - | 1 | - | 4 | 3 | 1 | - | - |
| P1 | 1 | 2 | - | 5 | 2 | - | 1 | - |
| P2 | - | 2 | 4 | 5 | - | 3 | - | 3 |
| P3 | 5 | 2 | - | - | 2 | 3 | 1 | - |
| P4 | 5 | - | 4 | 5 | 2 | 3 | - | - |
| P5 | 2 | 1 | 4 | 3 | 1 | - | - | 4 |
| P6 | 5 | - | - | 5 | 2 | 3 | 3 | - |
| P7 | - | 3 | 3 | - | 3 | 1 | 2 | - |

**Figura 4:** Ejemplo de una base de datos de recomendación de colegios, donde U corresponde al usuario que realiza la consulta en nuestro sistema. El resto de las variables están descritas en la figura 1.

Los ranquin asignados por el usuario se realizan de la siguiente manera: a) Se ranquean todos los colegios que estén dentro de la distancia consultada por el usuario (p.ej. 4 kilómetros), utilizando las ecuaciones 4 y 5 descritas anteriormente. b) Una vez ranqueados los colegios, se seleccionan los 5 colegios mas cercanos a la dirección consultada por el usuario, y cuyas caracteristicas coincidan con las consultadas (p.ej. nivel de enseñanza, tipo de colegio, orientacion religiosa, etc.). c) Los ranquin de estos 5 colegios son posteriormente asignacion a la matriz de recomendación, como si fuesen las preferencias del usuario.

Por otro lado, para la etapa de validación del sistema propuesto de recomendación se utilizará la información contenida en el SAE. Se seleccionarán de forma aleatoria 120 alumnos postulantes del SAE, 10 por cada uno de los niveles de enseñanza. Cada uno de estos alumnos corresponderá a un "usuario SAE" en nuestro sistema de recomendación. Para cada usuario SAE, se simularán 1000 postulantes, y se reconstruirá la matriz de recomendación al igual que los pasos descritos en los párrafos anteriores. Sin embargo, el ranquin de los colegios asignados por el usuario SAE, se realizarán de forma distintas a las descritas en el párrafo anterior, ya que en este caso se conocen las preferencias del usuario SAE.

Los colegios seleccionados por el usuario SAE serán ranqueados en base al siguiente procedimiento: a) Se ranquean todos los colegios que estén dentro de la distancia de 4 kilómetros de la localización geográfica del usuario SAE, utilizando las ecuaciones 4 y 5 descritas anteriormente. b) Una vez ranqueados los colegios, se seleccionan los colegios al cual el usuario SAE desea postular, información entregada por el RBD en la base de datos SAE, y se extraen las caracteristicas de los colegios seleccionados (p.ej. nivel de enseñanza, tipo de colegio, orientacion religiosa, etc.). c) Posteriormente, los ranquin de estos colegios son asignados a la matriz de recomendación, siendo estas la preferencias del usuaro SAE.

**Desarrollo de algoritmo de recomendación basado en factorización matricial**

El filtrado colaborativo es la aplicación de la factorización matricial para identificar la relación entre las entidades de los colegios y los usuarios o postulantes. Con la entrada del ranquin de los usuarios, nos gustaría predecir cómo los usuarios calificarían los colegios para que los usuarios puedan obtener la recomendación de colegios basada en la predicción. Por ejemplo, en la Figura 4, los postulantes P2 y P4, tienen preferencias muy similares entre ellos, por lo tanto, si quisiéramos saber el ranquin que el postulante P2 asignará al colegio C5, podríamos sugerir un ranquin de 2, ya que el resto de las predicciones son similares a las del postulante P4, por lo que sus opiniones también serian similares. Ahora bien, ¿Podríamos recomendar el colegio C5 a P2?, la respuesta es **NO**, debido a que el ranquin es demasiado bajo, si el ranquin fuese de 4 o 5 lo recomendaríamos sin dudarlo.

*Concepto matemático de la factorización matricial*

Para describir el concepto de factorización, utilizamos el articulo propuesto por Denise Chen [17], y se adaptó a nuestro propósito que son los colegios. Defina un conjunto de usuarios ($U$), colegios ($D$) y $R$ es una matriz de tamaño (número de usuarios x número de colegios). La matriz $R = |U| * |D|$ incluye todos los ranquin dados por los usuarios (p.ej. la matriz de la Figura 4, sería la matriz $R$). El objetivo es descubrir las características latentes $k$, que corresponden a las características de los usuarios y colegios (que en este caso es lo que se desea estimar). Por lo tanto, la multiplicación de dos matrices P(|U|*k) y Q(|D|*k) traspuesta, generaría como resultado la matriz R.

$$R \approx P \times Q^T = \hat{R} \tag{6}$$

La matriz $P$ representa la asociación entre un usuario y las características latentes, mientras que la matriz $Q$ representa la asociación entre un colegio y las características latentes. A partir de la ecuación anterior, podríamos obtener la predicción de la calificación de un elemento mediante el cálculo del producto escalar de los dos vectores correspondientes a $p_i$ y $q_j$, pertenecientes a los vectores $i$ y $j$ de las matrices P y Q respectivamente.

$$\hat{r}_{ij} = p_i^T q_j = \sum_{k=1}^{k} p_{ik} q_{kj} \tag{7}$$

Para obtener estas matrices $P$ y $Q$, necesitamos inicializar las dos matrices y calcular la diferencia del producto llamado matriz $M = R - \hat{R}$. Las matrices P y Q se pueden calcular de forma automática a través del uso de la librería de TensorFlow recommenders [18]. El paso siguiente una vez inicializadas las matrices corresponde, a minimizar la diferencia entre la matriz original $R$ y la matriz aproximada $\hat{R}$ a través de distintas iteraciones (en nuestro caso llamadas épocas), para este propósito se implementa el uso del optimizador AdaGrad (Optimizados de Gradiente Adaptativo). El método de gradiente adaptativo tiene como objetivo encontrar un mínimo local de la diferencia, entre los parámetros de matriz reconstruida ($\hat{r}_{ij}$), y la matriz original ($r_{ij}$).

$$e_{ij}^2 = \left(r_{ij} - \hat{r}_{ij}\right)^2 = \left(r_{ij} - \sum_{k=1}^{k} p_{ik} q_{kj}\right)^2 \tag{8}$$

Este procedimiento se repite de forma consecutiva para alcanzar un mínimo de iteraciones asignadas al algoritmo que en este caso corresponden a 15 épocas. En la Figura 5 se muestra un diagrama que resume el procedimiento aplicado para calcular la factorización matricial.

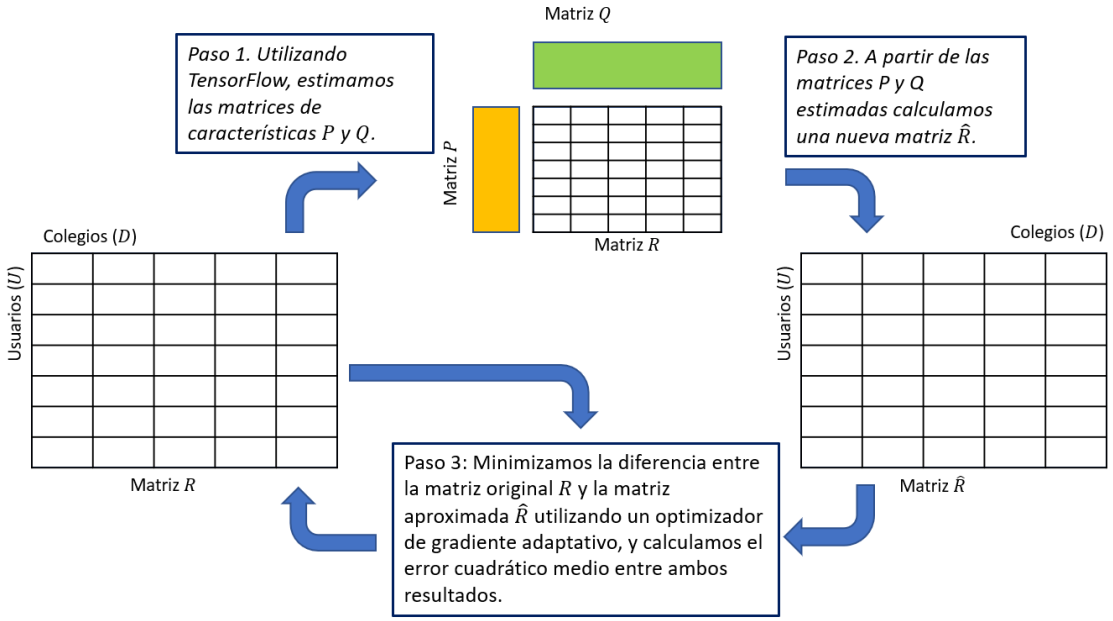



**Figura 5**: Ejemplo del procedimiento para realizar la factorización matricial. Los variables en el diagrama son las mismas que las descritas en las ecuaciones 6, 7 y 8.

El algoritmo propuesto de factorización matricial fue programado en TensorFlow, a través de su librería TensorFlow Recommenders, y se utilizaron los siguientes parámetros descritos en la Tabla 1.

Tabla1: Parámetros utilizados en el algoritmo propuesto.

| Parámetro | Descripción |
|---|---|
| Batch Size | 1024 para entrenamiento y 512 para validación. |
| Epochs | 15 épocas. |
| Loss Function | Error cuadrático medio (RMSE) |
| Optimizers | AdaGrad (Gradiente adaptativo) |
| Learning Rate | 0.1 |
| Latent Features | 64 características latentes tanto para los usuarios, como para los colegios. |
| Model Layers | Dense(256, activation="relu"),<br>Dense(64, activation="relu"),<br>Dense(1) |

En cuanto a los tiempos de cómputo, el modelo se demora alrededor de 11 segundos en ejecutarse y entregar una recomendación después de las 15 épocas, en un computador con un procesador Intel Core i7 11ª generación, 64 RAM, y una NVIDIA GPU RTX 3080 Ti.

Una ves concluido el algoritmo se debe entregar la información de los colegios recomendados, para ello usaremos tanto el formado de DataFrame de pandas y también se mostrarán en un mapa con la localización, desde el que tiene mayor prioridad al que tiene menor prioridad. Un ejemplo del mapa se muestra en la Figura 6, para un usuario que consulto, por la dirección Rosario Sur 600, radio de búsqueda 4 kilómetros, primero básico, hombre y colegios privados.

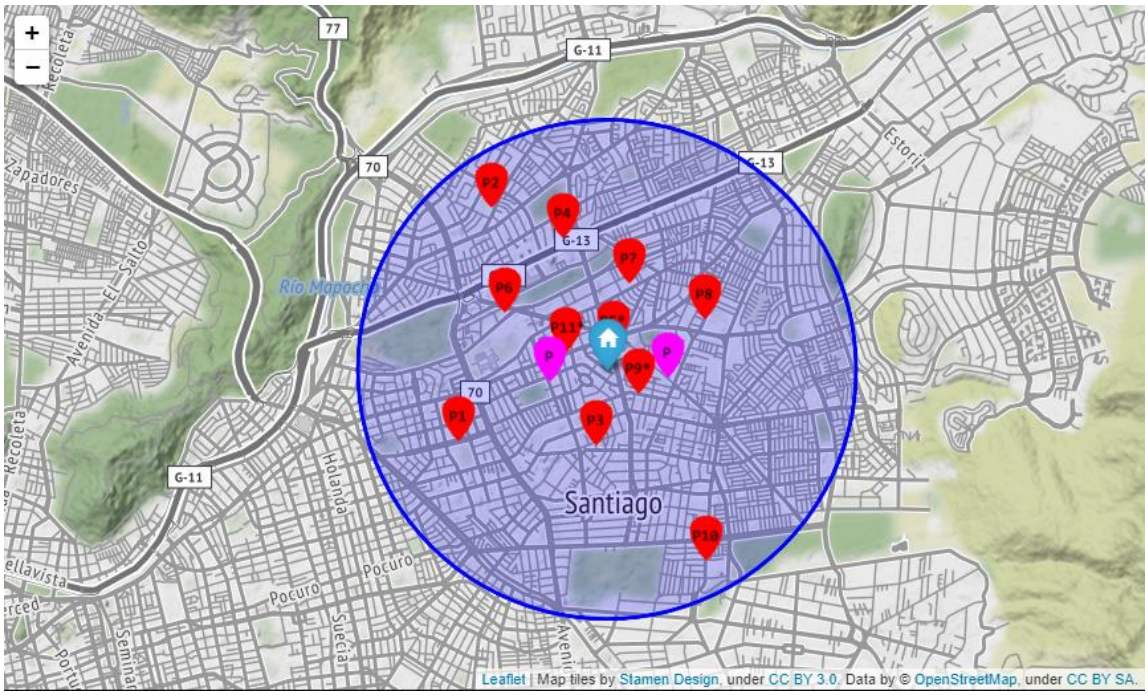



**Figura 6:** Ejemplo de mapa de localización de las recomendaciones sugeridas por el algoritmo, P1 es el colegio con mayor ponderación en cuanto al ranquin sugerido de la matriz $\hat{R}$, seguido de P2, P3 y así sucesivamente. El asterisco (*) indica que ese colegio, fue seleccionado inicialmente por el usuario, utilizado solo la métrica de distancia. Los colegios en magenta marcados con la letra P, son colegios que fueron seleccionados inicialmente pero no fueron sugeridos por el sistema.

**Análisis exploratorio, validación y visualizaciones del algoritmo desarrollado**

Inicialmente se realizará un análisis estadístico descriptivo de las bases de datos filtradas, correspondientes a la base de datos de directorio para todos los niveles de enseñanza (Tabla 1 del ANEXO), y la base de datos de validación SAE (Tabla 2 ANEXO).

La etapa de validación del algoritmo desarrollado consta de la selección aleatoria de 120 postulantes de la base de datos de validación SAE, 10 por cada uno de los niveles de enseñanza 1ro básico a 4to medio. Para cada uno de los postulantes, se aplicará el algoritmo descrito en los párrafos anteriores, y se seleccionaran los 10 colegios mas recomendados, para compararlos con los colegios que postulante selecciono. Se realizarán 3 comparaciones; I) todos los colegios seleccionados por los postulantes, con las 10 mejores recomendaciones del algoritmo, II) el primer colegio seleccionado por el postulante, con la primera recomendación del algoritmo, III) los primeros 3 colegios seleccionados por el postulante, con las primeras 3 recomendaciones dadas por el algoritmo. Se analizará la distancia y el puntaje obtenido por el colegio (ecuación 4), para cada nivel de enseñanza. Se utilizarán gráficos de cajas y también se aplicará el análisis estadístico de Mann-Whitney U test, con una significancia estadística p-value $< 0.5$ o menor.

En términos de visualización geográfica de datos, algunas de ellas fueron previamente mostradas en las figuras 2 y 6 de este documento, pero adicionalmente se programará en el lenguaje de PHP, algunas descripciones de los colegios, con el fin de que sea una visualización interactiva.

# 5. Resultados

### Análisis exploratorio de las bases de datos

En este apartado analizamos las bases de datos de directorio filtradas y también las bases de datos de validación del SAE, para poder entregar una idea de cómo están compuestas ambas bases de datos.

Una descripción básica de lo que contiene la base de datos de directorio se muestra en la Tabla 2. Donde se puede observar que la región metropolitana, posee más colegios particulares subvencionados, que municipales y particulares pagados. De igual forma también hay más alumnos hombres que mujeres en los distintos niveles de enseñanza.

Tabla 2: Parámetros descriptivos básicos de los colegios en la base de datos de directorio.

| Nivel | Cantidad de Colegios Municipales | Cantidad de Colegios P. Subvencionados | Cantidad de Colegios P. Pagados | Cantidad de hombres | Cantidad de Mujeres |
|---|---|---|---|---|---|
| 1ro Básico | 490 | 950 | 267 | 50912 | 47685 |
| 2do Básico | 490 | 949 | 267 | 50189 | 47390 |
| 3ro Básico | 487 | 944 | 263 | 51101 | 47796 |
| 4to Básico | 490 | 946 | 266 | 51049 | 47910 |
| 5to Básico | 489 | 938 | 267 | 50783 | 48237 |
| 6to Básico | 491 | 940 | 266 | 49898 | 47683 |
| 7mo Básico | 512 | 898 | 262 | 49075 | 46651 |
| 8vo Básico | 512 | 899 | 260 | 46240 | 44913 |
| 1ro Medio | 171 | 583 | 243 | 42372 | 41288 |
| 2do Medio | 170 | 580 | 240 | 39901 | 39602 |
| 3ro Medio | 139 | 503 | 239 | 29059 | 29278 |
| 4to Medio | 137 | 496 | 239 | 26854 | 28014 |

En la Figura 7 se muestra un análisis descriptivo de algunos de los parámetros utilizados para calcular el puntaje de cada uno de los colegios (ecuación 4), a partir de la base de datos de directorio. El promedio de notas corresponde a un 5.36, entre todos los colegios de la base de datos, para todos los niveles de enseñanza, la asistencia promedio es de un 86%, el porcentaje de aprobados es de un 92.3%, el porcentaje de alumnos que no desertan es de un 87.9%, el puntaje SIMCE promedio de lectura es de 261.5, y matemáticas es de 260.1, el promedio de puntaje para entrar a la universidad es de 621 en la prueba de selección universitaria PSU, con un promedio de postulantes de 6, equivalentes a un porcentaje promedio del curso de 21%.

| | PROM | ASIS | APRB | NOND | SIMCE-LEC | SIMCE-MAT | PROM-PUNT | PROM-POST | DE-F-YEAR | PORC-MATR |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 17283.000000 | 17283.000000 | 17283.000000 | 17283.000000 | 16761.000000 | 16761.000000 | 9980.000000 | 9980.000000 | 9980.000000 | 9980.000000 |
| **mean** | 5.360171 | 85.842037 | 92.277676 | 87.953901 | 261.544680 | 260.081247 | 621.040866 | 6.372049 | 99.338875 | 21.131541 |
| **std** | 0.642911 | 8.745203 | 7.855439 | 9.962775 | 24.658381 | 32.356632 | 60.313801 | 5.412501 | 7.541022 | 18.616327 |
| **min** | 0.008563 | 0.275000 | 0.000000 | 0.000000 | 124.000000 | 102.666667 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 5.098509 | 83.382748 | 89.348766 | 84.042060 | 243.000000 | 236.142857 | 599.518750 | 2.312500 | 100.000000 | 7.083333 |
| **50%** | 5.416362 | 87.489083 | 94.436718 | 90.416667 | 261.000000 | 255.428571 | 614.744792 | 4.687500 | 100.000000 | 15.056818 |
| **75%** | 5.718215 | 90.432532 | 97.752809 | 94.576035 | 278.857143 | 280.142857 | 640.031897 | 9.104167 | 100.000000 | 30.681818 |
| **max** | 6.966667 | 100.000000 | 100.000000 | 100.000000 | 336.285714 | 386.000000 | 870.303409 | 35.312500 | 100.000000 | 98.684211 |

**Figura 7:** DataFrame con el análisis descriptivo de los parámetros utilizados en la ecuación 4. Para todos los colegios y niveles de enseñanza.

Debido a que los últimos parámetros correspondientes a las postulaciones a la universidad están solo incluidos en los cursos de enseñanza media, analizaremos a través del uso de histogramas y gráficos de dispersión, las variables de promedio de notas (PROM), promedio de puntaje SIMCE lectura (SIMCE-LEC), promedio de puntaje SIMCE matemática (SIMCE-MAT), promedio de puntaje PSU (PROM-PUNT),

porcentaje de matriculados en la universidad (PORC-MATR), solo para los alumnos de 4to medio, entre los distintos tipos de colegio.

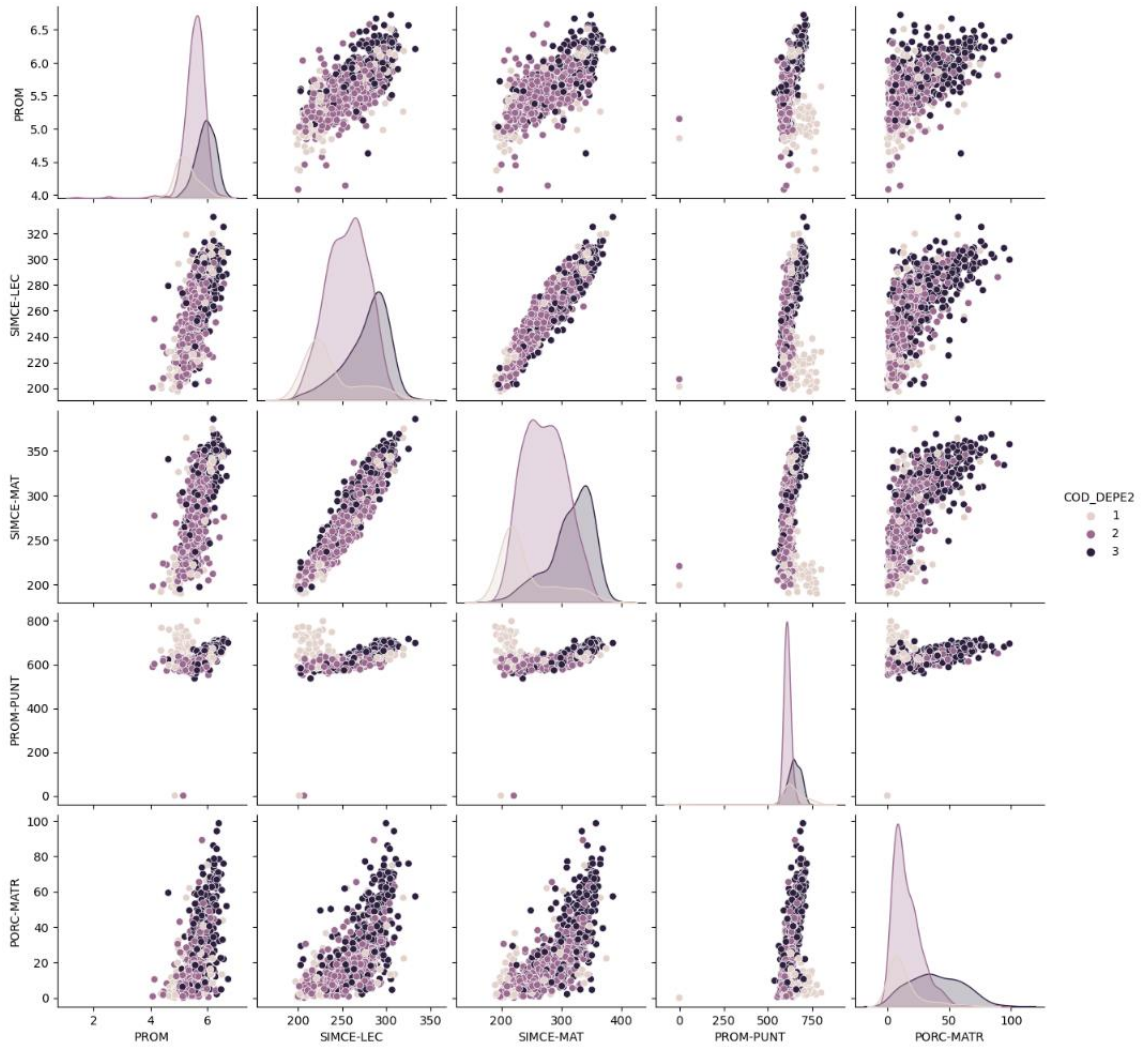



**Figura 8:** Histogramas y gráficos de dispersión, para un grupo de variables de desempeño de los alumnos de 4to medio, contenidas en la base de datos de directorio. Cada color indica si el establecimiento es municipal = 1, particular subvencionado = 2, o particular pagado = 3.

Observando la Figura 8, es evidente que los alumnos de colegios particulares pagados tienen un desempeño mayor con respecto a los colegios municipales y particulares subvencionados, tanto en promedio de notas como SIMCE. Esta ventaja notoria no es tan evidente cuando se analizan los puntajes de la PSU, cuyo histograma es más estrecho entre los distintos establecimientos. Sin embargo, los estudiantes de colegios particulares pagados tienen más opciones de matricularse en la universidad que los estudiantes de los otros colegios.

La base de datos SAE, contiene información acerca de los postulantes (MRUN) y sus preferencias de colegios (PREFERENCIA_POSTULANTE), adicionalmente entrega las coordenadas georreferenciales de los postulantes. Por lo que el primer paso consistió en unir esta base de datos por RBD con la base de datos de directorio, para saber cual es la distancia promedio entre la localización georreferencial del postulante y la localización georreferencial del colegio al cual postula. El resultado de este análisis nos entregó que estudiantes están dispuestos a recorrer una distancia de 3.84 kilómetros en promedio, es por ello por lo que para la validación se estableció una distancia de búsqueda de 4 kilómetros.

Posteriormente limpiamos esta base de datos para quedarnos solo con la base de datos SAE filtrada, que contiene las preferencias de 120 postulantes seleccionados de forma aleatoria, con la condición de que vivan a una distancia menor de 4 kilómetros con respecto a los colegios a los cuales postulan. La descripción básica de esta base de datos se muestra en la Tabla 3. Donde se puede observar que, en los cursos de básica, los alumnos tienen a vivir más cerca de sus casas que en los cursos de media, el promedio de

postulaciones es muy similar para todos los niveles de enseñanza a excepción de 6to básico que aumenta a más de 10 colegios seleccionados por postulante. En cuanto a la cantidad de hombres vs mujeres de nuestra base de datos de validación, se observa que ambos números son bastante similares.

Tabla 3: Parámetros descriptivos básicos de los 120 postulantes seleccionados de forma aleatoria en la base de datos de validación SAE.

| Nivel | Cantidad de Colegios en promedio a los que postulan | Distancia promedio entre el colegio y la ubicación del alumno (km) | Cantidad de hombres | Cantidad de Mujeres |
|---|---|---|---|---|
| 1ro Básico | 7.4 | 1.42 | 4 | 6 |
| 2do Básico | 7.2 | 1.35 | 5 | 5 |
| 3ro Básico | 8.2 | 1.33 | 6 | 4 |
| 4to Básico | 8.2 | 1.72 | 4 | 6 |
| 5to Básico | 7.9 | 1.76 | 6 | 4 |
| 6to Básico | 10.1 | 1.74 | 6 | 4 |
| 7mo Básico | 8.9 | 1.71 | 6 | 4 |
| 8vo Básico | 7.7 | 1.48 | 5 | 5 |
| 1ro Medio | 7.6 | 1.84 | 5 | 5 |
| 2do Medio | 8.7 | 1.96 | 3 | 7 |
| 3ro Medio | 7.5 | 1.67 | 4 | 6 |
| 4to Medio | 7.5 | 1.87 | 5 | 5 |

**Validación del modelo propuesto**

La etapa de validación del algoritmo desarrollado consta de la selección aleatoria de 120 postulantes de la base de datos de validación SAE, 10 por cada uno de los niveles de enseñanza 1ro básico a 4to medio. Estos colegios corresponden solo a colegios municipales y particulares subvencionados.

*Utilizando todos los colegios y todas las recomendaciones:* Este análisis consiste en la comparación de todos los colegios seleccionados por los postulantes, con las 10 mejores recomendaciones del algoritmo, para cada uno de los niveles de enseñanza (Figuras 9 y 10), podemos observar que todos los colegios recomendados están dentro del rango de búsqueda establecido de 4 kilómetros. Existen diferencias significativas a partir del uso de Man-Whitney U test, entre la mayoría de los resultados, a excepción de 8vo básico, 1ro medio y 2do medio. Es de esperar que este parámetro diera significativo ya que la mayor parte de los postulantes busca el colegio más cercano a donde reside, independientemente de las características de desempeño del colegio.

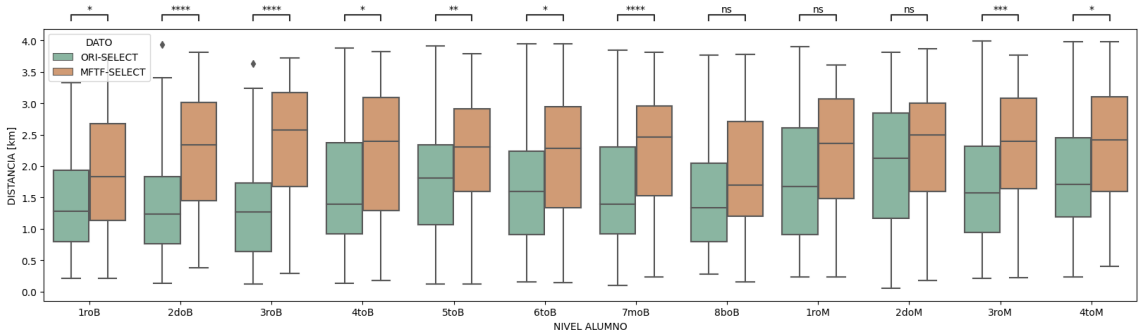



**Figura 9:** Grafico de cajas para la distancia en kilómetros, entre la selección original de los postulantes (ORI-SELECT) y las 10 mejores recomendaciones obtenidas del método de factorización matricial (MFTF-SELECT), para cada uno de los niveles de enseñanza. Las variables **ns** = no significativo, * = 0.05 > p-value > 0.01, ** = 0.01 > p-value > 0.001, *** = 0.001 > p-value > 0.0001, **** = p-value < 0.0001.

Por otro lado, si analizamos el puntaje obtenido por los colegios recomendados, todos los niveles presentan diferencias significativas y se recomiendan alternativas de colegios con un puntaje (ecuación 4) mayor que los seleccionados por los postulantes, podemos ver que el valor promedio de lo reportado por el algoritmo propuesto es superior al valor promedio de los colegios seleccionados por los postulantes. Lo cual nos indica que nuestro algoritmo entrega una alternativa de colegios con mejor desempeño que los que el postulante selecciono en sistema SAE. También podemos observar el grafico de la Figura 10, que los resultados entregados por el algoritmo tienen una variabilidad menor a los resultados entregados por los postulantes.

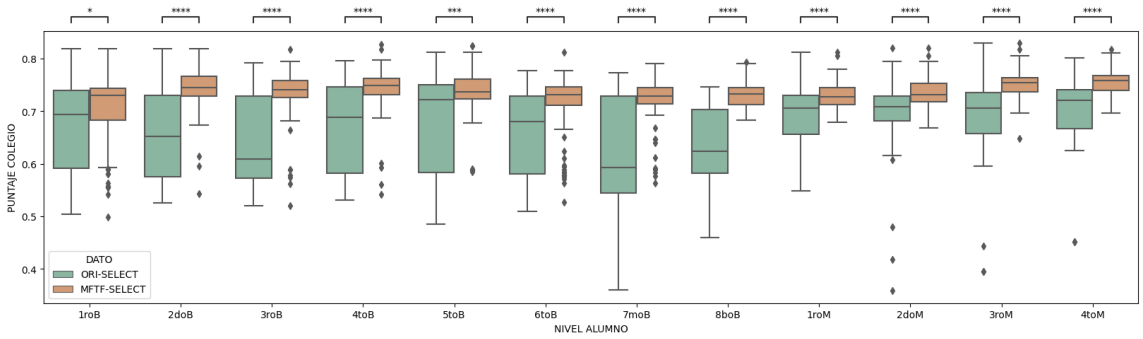



**Figura 10:** Grafico de cajas para el puntaje obtenido por el colegio, entre la selección original de los postulantes (ORI-SELECT) y las 10 mejores recomendaciones obtenidas del método de factorización matricial (MFTF-SELECT), para cada uno de los niveles de enseñanza. Las variables **ns** = no significativo, * = 0.05 > p-value > 0.01, ** = 0.01 > p-value > 0.001, *** = 0.001 > p-value > 0.0001, **** = p-value < 0.0001.

*Utilizando el primer colegio y la primera recomendación:* Podemos observar que para la mayoría de los niveles no existe diferencia significativa en cuanto a distancia al colegio, a excepción de 3ro básico y 8vo básico. Esto quiere decir que la primera recomendación esta a una distancia bastante similar a la del colegio seleccionado por el postulante.

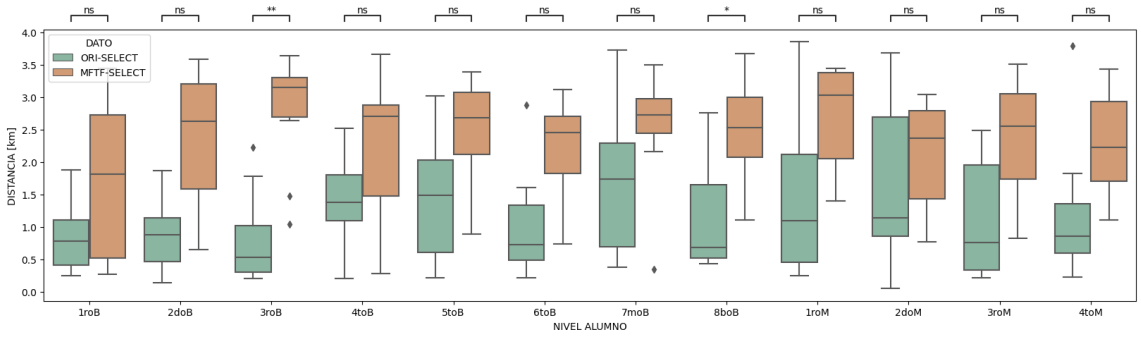



**Figura 11:** Grafico de cajas para la distancia en kilómetros, entre la primera preferencia de los postulantes (ORI-SELECT) y la primera recomendación obtenidas del método de factorización matricial (MFTF-SELECT), para cada uno de los niveles de enseñanza. Las variables **ns** = no significativo, * = 0.05 > p-value > 0.01, ** = 0.01 > p-value > 0.001, *** = 0.001 > p-value > 0.0001, **** = p-value < 0.0001.

Este mismo patrón se aprecia cuando analizamos el puntaje obtenido por el colegio. Sin embargo, nuestras recomendaciones siguen teniendo valores superiores a las seleccionadas por los postulantes y con una menor variabilidad.

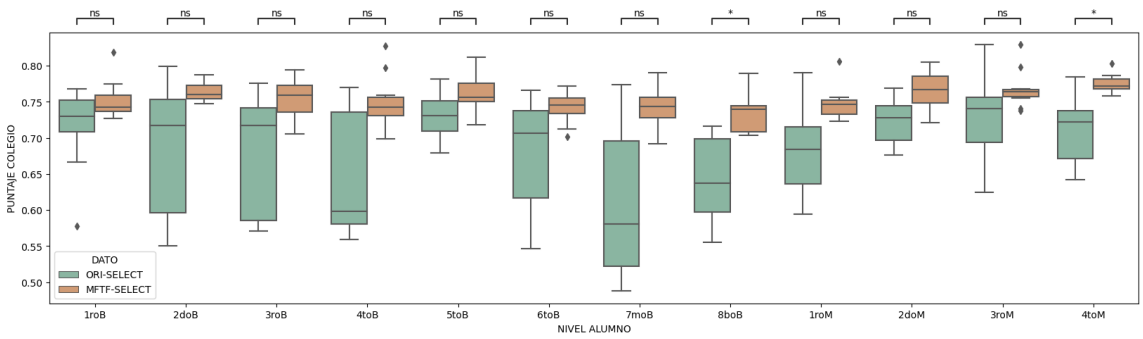



**Figura 12:** Grafico de cajas para el puntaje obtenido por el colegio, entre la primera preferencia de los postulantes (ORI-SELECT) y la primera recomendación obtenidas del método de factorización matricial (MFTF-SELECT), para cada uno de los niveles de enseñanza. Las variables **ns** = no significativo, * = 0.05 > p-value > 0.01, ** = 0.01 > p-value > 0.001, *** = 0.001 > p-value > 0.0001, **** = p-value < 0.0001.

*Utilizando los primeros tres colegios y las primeras tres recomendaciones:* Podemos observar un resultado muy similar al reportado en el primer experimento, la mayoría de los niveles de enseñanza poseen una diferencia significativa, y nuestras recomendaciones están más lejos que las seleccionadas por el postulante, pero dentro del rango sugerido de 4 kilómetros.

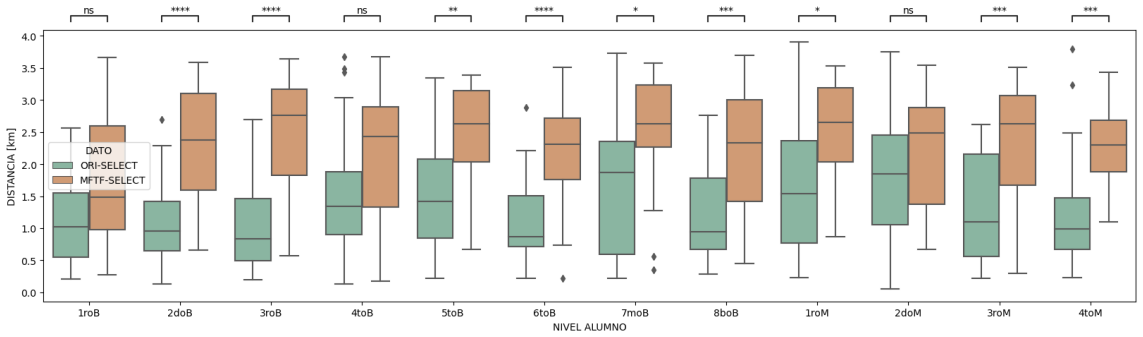


**Figura 13:** Grafico de cajas para la distancia en kilómetros, entre las 3 primeras preferencias selección original de los postulantes (ORI-SELECT) y las 3 mejores recomendaciones obtenidas del método de factorización matricial (MFTF-SELECT), para cada uno de los niveles de enseñanza. Las variables **ns** = no significativo, * = 0.05 > p-value > 0.01, ** = 0.01 > p-value > 0.001, *** = 0.001 > p-value > 0.0001, **** = p-value < 0.0001.

Si analizamos el puntaje obtenido por los colegios recomendados, para todos los niveles presentan diferencias significativas y se recomiendan alternativas de colegios con un puntaje (ecuación 4) mayor que los seleccionados por los postulantes, podemos ver que el valor promedio de lo reportado por el algoritmo propuesto es superior al valor promedio de los colegios seleccionados por los postulantes.

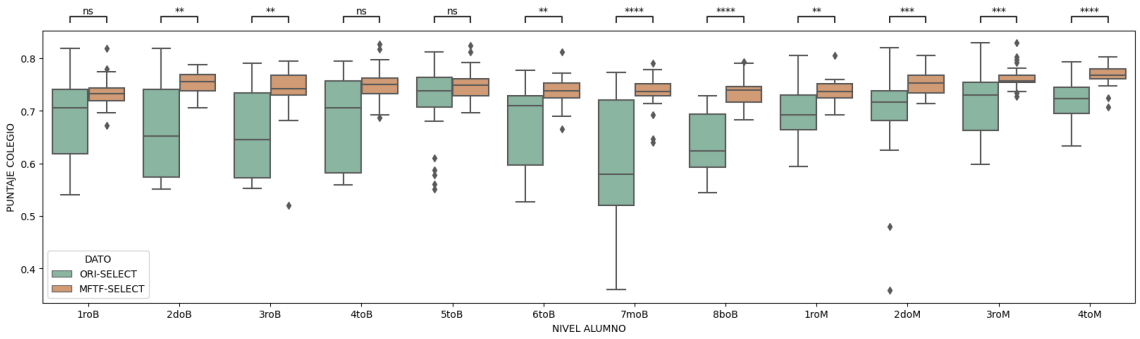



**Figura 14:** Grafico de cajas para el puntaje obtenido por el colegio, entre las 3 primeras preferencias selección original de los postulantes (ORI-SELECT) y las 3 mejores recomendaciones obtenidas del método de factorización matricial (MFTF-SELECT), para cada uno de los niveles de enseñanza. Las variables **ns** = no significativo, * = 0.05 > p-value > 0.01, ** = 0.01 > p-value > 0.001, *** = 0.001 > p-value > 0.0001, **** = p-value < 0.0001.

El evaluar el comportamiento promedio de la función de perdida optimizada a través del método de gradiente adaptativo, y el error cuadrático medio (RMSE) para los 120 análisis resultados. Podemos ver en la Figura 15, que ambos casos convergen rápidamente después de la 4$^{ta}$ iteración.

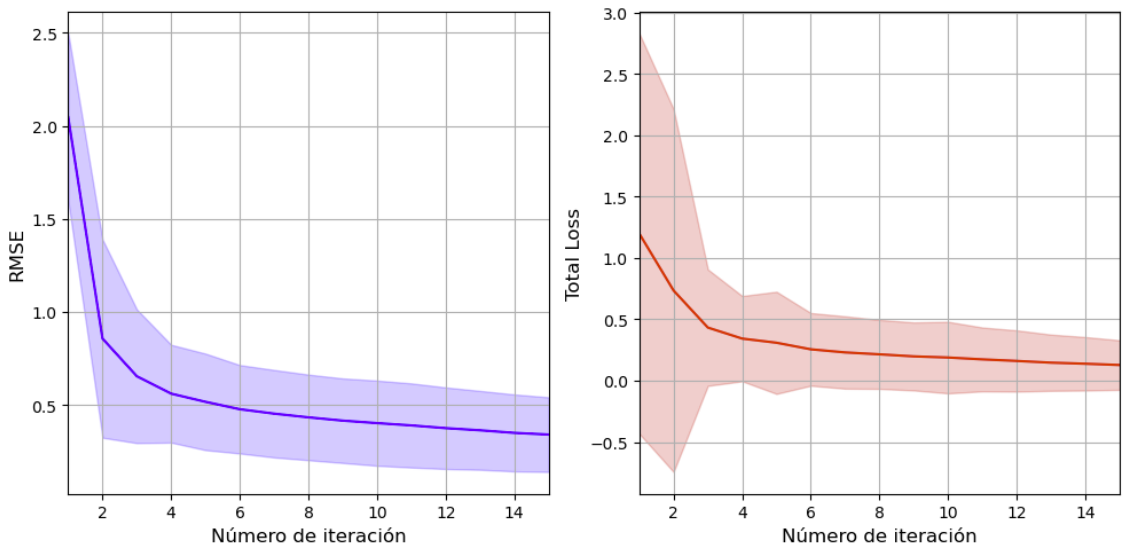



**Figura 15:** Grafico de la convergencia del error cuadrático medio (RMSE) y función de pérdida del algoritmo planteado, entre las 120 postulaciones seleccionadas, la zona sombreada corresponde a la desviación estándar y la línea solida corresponde al promedio.

**Visualización de los resultados**

Para poder visualizar los resultados obtenidos de forma adecuada, se programó en PHP, un mensaje con características relevantes de los colegios recomendados como; el tipo de colegio, valor mensual, orientación religiosa, puntaje del colegio (ecuación 4) y score asignado por el procedimiento de factorización matricial. Realizaremos 3 ejemplos

de selección de colegios; a) selección de colegios particulares pagados, b) selección de colegios municipales y c) selección de una combinación de colegios municipales con particular subvencionado.

*Para la selección de colegios particulares pagados*: se utilizaron las siguientes características:

- Dirección: Rosario sur 600, Santiago, Santiago
- Radio de búsqueda: 3 kilómetros
- Nivel alumno: 1 (primero básico)
- Genero alumno: 1 (hombre)
- Colegio tipo: 3 (Particular pagado)
- Religión: [ ] (Indiferente)
- Pago mensual: [ ] (Indiferente)

El resultado de la convergencia del algoritmo se muestra en la siguiente figura.

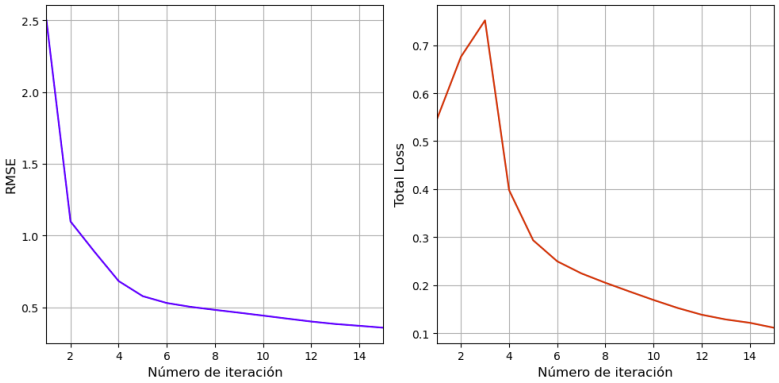
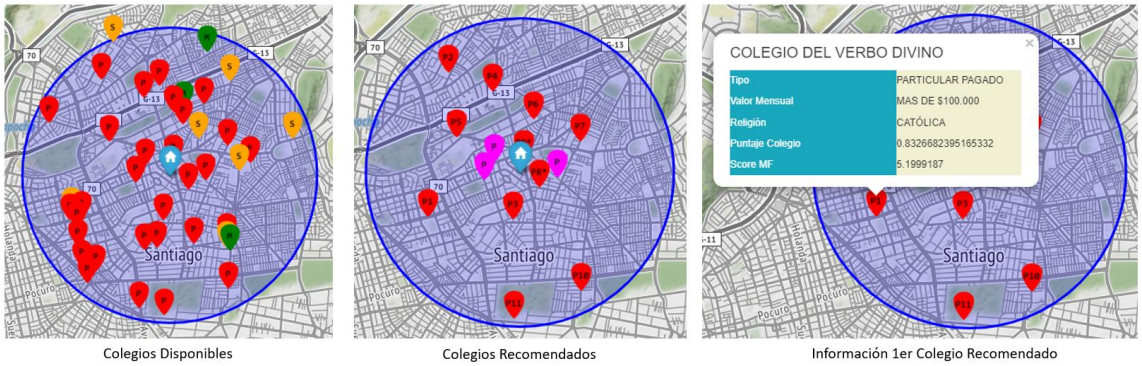



**Figura 16:** Grafico de la convergencia del error cuadrático medio (RMSE) y función de pérdida del algoritmo planteado, para la selección de colegios particulares pagados.

Los colegios disponibles (M = municipal, S = Subvencionado, P = Particular pagado), los colegios recomendados y el panel desplegable para la primera recomendación se muestran en la figura 17. Donde se aprecia, que la primera recomendación corresponde al colegio Verbo Divino, tiene un puntaje de 0.83, es un colegio particular pagado, con orientación católica y corresponde al score más alto entregado por el modelo de factorización matricial de 5.19.




Colegios Disponibles | Colegios Recomendados | Información 1er Colegio Recomendado

**Figura 17:** Recomendación de colegios particulares pagados. En la figura izquierda se observan todos los colegios disponibles en el rango buscado de 3 kilómetros. En la figura central se aprecian los resultados de los colegios recomendados en color rojo. Los colegios destacados con magenta en esa figura corresponden a la primera selección en base a distancia que se realiza de forma automática para poder recomendar colegios. En la figura derecha se muestra la información resumida para el primer colegio recomendado.

Si observamos el DataFrame para las primeras 10 recomendaciones, podemos ver que todos los colegios recomendados tienen un SCORE entregado por el modelo de factorización matricial cercano a 5, con un puntaje del colegio (variable SUM, ecuación 4) mayor que 0.75, las distancias no superan los 3 kilómetros correspondiente al límite consultado.

| | RBD | NOM_RBD | DISTANCIA | SUM | COD_DEPE2 | ORI_RELIGIOSA | PAGO_MATRICULA | PAGO_MENSUAL | ID_OS | SCORE | ID_NIVEL | MRUN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 8888 | COLEGIO DEL VERBO DIVINO | 1.998916 | 0.832668 | 3 | 2 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.199919 | 1 | 1 |
| 1 | 8923 | LYCEE ANTOINE DE SAINT EXUPERY | 2.360426 | 0.784957 | 3 | 1 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.103971 | 1 | 1 |
| 2 | 8868 | COLEGIO LA GIROUETTE | 0.950799 | 0.821320 | 3 | 1 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.093509 | 1 | 1 |
| 3 | 8919 | COLEGIO SAN PEDRO NOLASCO | 1.629121 | 0.816574 | 3 | 2 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.057034 | 1 | 1 |
| 4 | 24884 | COLEGIO ALCAZAR DE LAS CONDES | 1.385599 | 0.780875 | 3 | 6 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.034893 | 1 | 1 |
| 5 | 8893 | COLEGIO INTERNACIONAL SEK CHILE | 1.038169 | 0.787187 | 3 | 1 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.006482 | 1 | 1 |
| 6 | 8900 | COLEGIO SEMINARIO PONTIFICIO MENOR | 1.312292 | 0.797487 | 3 | 2 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.006111 | 1 | 1 |
| 7 | 8892 | COLEGIO SAN JUAN EVANGELISTA | 0.503896 | 0.798202 | 3 | 2 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.004812 | 1 | 1 |
| 8 | 8870 | COLEGIO COMPANIA DE MARIA APOQUINDO | 0.257917 | 0.804560 | 3 | 2 | MAS DE $100.000 | MAS DE $100.000 | 2 | 5.003054 | 1 | 1 |
| 9 | 9054 | COLEGIO SAINT JOHN S VILLA ACADEMY | 2.628312 | 0.822094 | 3 | 2 | MAS DE $100.000 | MAS DE $100.000 | 2 | 4.984198 | 1 | 1 |
| 10 | 9046 | COLEGIO THE GRANGE SCHOOL | 2.883127 | 0.790249 | 3 | 1 | MAS DE $100.000 | MAS DE $100.000 | 2 | 4.636018 | 1 | 1 |

**Figura 18:** DataFrame con las 10 mejores recomendaciones de colegios particulares pagados.

*Para la selección de colegios municipales*: se utilizaron las siguientes características:

- Dirección: Lira 245, Santiago, Santiago
- Radio de búsqueda: 4 kilómetros
- Nivel alumno: 3 (tercero básico)
- Genero alumno: 1 (hombre)
- Colegio tipo: 1 (Municipal)
- Religión: [ ] (Indiferente)
- Pago mensual: [ ] (Indiferente)

El resultado de la convergencia del algoritmo se muestra en la siguiente figura.

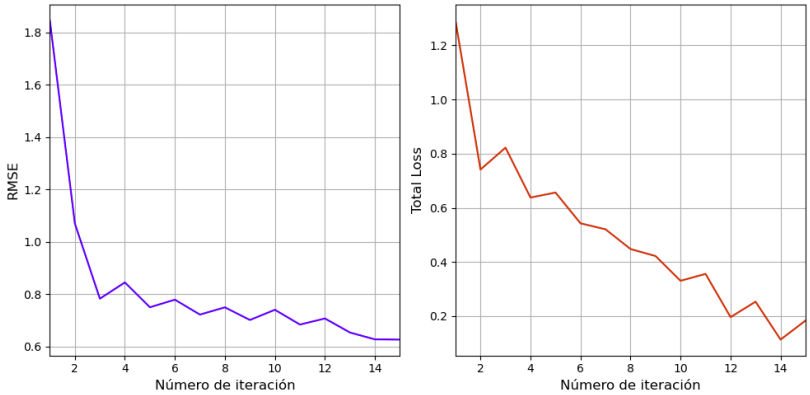



**Figura 19:** Grafico de la convergencia del error cuadrático medio (RMSE) y función de pérdida del algoritmo planteado, para la selección de colegios municipales.

Los colegios disponibles (M = municipal, S = Subvencionado, P = Particular pagado), los colegios recomendados y el panel desplegable para la primera recomendación se muestran en la figura 20. Donde se aprecia, que la primera recomendación corresponde al Liceo Republica de Brasil, tiene un puntaje de 0.70, es un colegio municipal, con orientación laica y corresponde al score más alto entregado por el modelo de factorización matricial de 4.64.

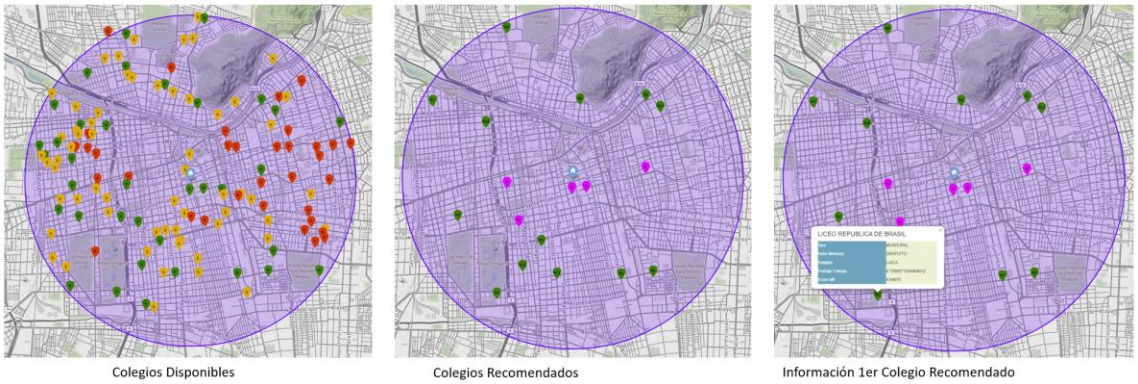



**Figura 20:** Recomendación de colegios municipales. En la figura izquierda se observan todos los colegios disponibles en el rango buscado de 4 kilómetros. En la figura central se aprecian los resultados de los colegios recomendados en color verde. Los colegios destacados con magenta en esa figura corresponden a la primera selección en base a distancia que se realiza de forma automática para poder recomendar colegios. En la figura derecha se muestra la información resumida para el primer colegio recomendado.

Si observamos el DataFrame para las primeras 10 recomendaciones, podemos ver que los primeros 3 colegios recomendados tienen un SCORE entregado por el modelo de factorización matricial mayor que 4, con un puntaje del colegio (variable SUM, ecuación 4) mayor que 0.5, las distancias no superan los 4 kilómetros correspondientes al límite consultado.

| | RBD | NOM_RBD | DISTANCIA | SUM | COD_DEPE2 | ORI_RELIGIOSA | PAGO_MATRICULA | PAGO_MENSUAL | ID_OS | SCORE | ID_NIVEL | MRUN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 8535 | LICEO REPUBLICA DE BRASIL | 3.357574 | 0.705807 | 1 | 1 | GRATUITO | GRATUITO | 2 | 4.640750 | 3 | 1 |
| 1 | 8928 | LICEO JOSE VICTORINO LASTARRIA | 2.482367 | 0.770000 | 1 | 1 | GRATUITO | GRATUITO | 2 | 4.628519 | 3 | 1 |
| 2 | 8498 | LICEO DARIO SALAS | 2.884591 | 0.520384 | 1 | 1 | GRATUITO | GRATUITO | 2 | 4.299942 | 3 | 1 |
| 3 | 8562 | CIUDAD SANTIAGO DE CHILE | 1.770889 | 0.589124 | 1 | 1 | GRATUITO | GRATUITO | 2 | 3.811263 | 3 | 1 |
| 4 | 8576 | ESCUELA BASICA REYES CATOLICOS | 2.646265 | 0.588117 | 1 | 1 | GRATUITO | GRATUITO | 2 | 3.616225 | 3 | 1 |
| 5 | 20440 | LICEO DOCTOR JUAN VERDAGUER PLANAS | 1.720478 | 0.574204 | 1 | 6 | SIN INFORMACION | SIN INFORMACION | 2 | 2.425461 | 3 | 1 |
| 6 | 8561 | ESCUELA REPUBLICA DE ALEMANIA | 3.684811 | 0.561616 | 1 | 1 | GRATUITO | GRATUITO | 2 | 2.401151 | 3 | 1 |
| 7 | 8497 | LICEO GABRIELA MISTRAL | 3.718322 | 0.663745 | 1 | 1 | GRATUITO | GRATUITO | 2 | 2.397297 | 3 | 1 |
| 8 | 24959 | COMPL EDUCACIONAL BRIGIDA WALKER ANEXO | 2.994867 | 0.568392 | 1 | 2 | GRATUITO | GRATUITO | 2 | 2.247493 | 3 | 1 |
| 9 | 8532 | ESCUELA BASICA LIBERTADORES DE CHILE | 2.347939 | 0.579872 | 1 | 1 | GRATUITO | GRATUITO | 2 | 2.092812 | 3 | 1 |
| 10 | 8531 | ESCUELA BASICA IRENE FREI DE CID | 2.400851 | 0.579347 | 1 | 1 | GRATUITO | GRATUITO | 2 | 2.063806 | 3 | 1 |

**Figura 21:** DataFrame con las 10 mejores recomendaciones de colegios municipales.

*Para la selección de colegios municipales y particulares subvencionados*: se utilizaron las siguientes características:

- Dirección: Lira 245, Santiago, Santiago
- Radio de búsqueda: 3,5 kilómetros
- Nivel alumno: 5 (quinto básico)
- Genero alumno: 2 (mujer)
- Colegio tipo: [1, 2] (Municipal, Particular Subvencionado)
- Religión: [ ] (Indiferente)
- Pago mensual: [ ] (Indiferente)

El resultado de la convergencia del algoritmo se muestra en la siguiente figura.

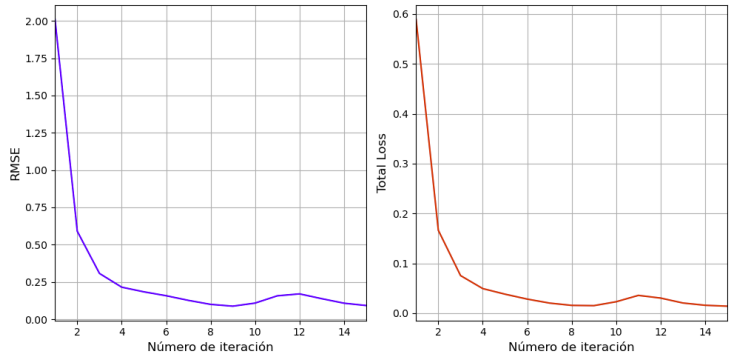
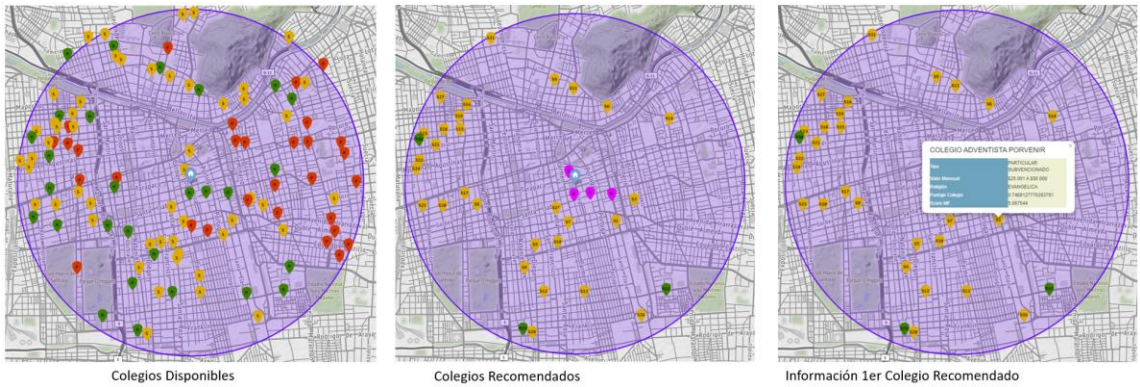



**Figura 22:** Grafico de la convergencia del error cuadrático medio (RMSE) y función de pérdida del algoritmo planteado, para la selección de colegios municipales y particulares subvencionados.

Los colegios disponibles (M = municipal, S = Subvencionado, P = Particular pagado), los colegios recomendados y el panel desplegable para la primera recomendación se muestran en la figura 23. Donde se aprecia, que la primera recomendación corresponde al Colegio Adventista Porvenir, tiene un puntaje de 0.74, es un colegio particular subvencionado, con orientación evangélica y corresponde al score más alto entregado por el modelo de factorización matricial de 5.06.




Colegios Disponibles | Colegios Recomendados | Información 1er Colegio Recomendado

**Figura 23:** Recomendación de colegios municipales. En la figura izquierda se observan todos los colegios disponibles en el rango buscado de 3.5 kilómetros. En la figura central se aprecian los resultados de los colegios recomendados en color amarillo (Particulares Subvencionados) y color verde (Municipales). Los colegios destacados con magenta en esa figura corresponden a la primera selección en base a distancia que se realiza de forma automática para poder recomendar colegios. En la figura derecha se muestra la información resumida para el primer colegio recomendado.

Si observamos el DataFrame para las primeras 10 recomendaciones, podemos ver que todos los colegios recomendados tienen un SCORE entregado por el modelo de factorización matricial mayor que 5, con un puntaje del colegio (variable SUM, ecuación 4) mayor que 0.7, las distancias no superan los 3.5 kilómetros correspondientes al límite consultado. Y adicionalmente podemos ver que los primeros 10 colegios recomendados corresponden a colegios particulares subvencionados (COD_DEPE2 = 2).

| | RBD | NOM_RBD | DISTANCIA | SUM | COD_DEPE2 | ORI_RELIGIOSA | PAGO_MATRICULA | PAGO_MENSUAL | ID_OS | SCORE | ID_NIVEL | MRUN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 8680 | COLEGIO ADVENTISTA PORVENIR | 1.252846 | 0.746913 | 2 | 3 | 1.000A10.000 | 25.001A50.000 | 2 | 5.067544 | 5 | 1 |
| 1 | 8627 | COLEGIO SANTA MARIA DE SANTIAGO | 0.761623 | 0.707908 | 2 | 1 | GRATUITO | 50.001A100.000 | 2 | 5.066140 | 5 | 1 |
| 2 | 9151 | COLEGIO SANTA MARGARITA DE ESCOCIA | 1.314361 | 0.698098 | 2 | 1 | 1.000A10.000 | 25.001A50.000 | 2 | 5.059122 | 5 | 1 |
| 3 | 8604 | COLEGIO SANTA CRUZ | 2.151968 | 0.766707 | 2 | 2 | 1.000A10.000 | 25.001A50.000 | 2 | 5.058455 | 5 | 1 |
| 4 | 8658 | COLEGIO FRANCISCO ARRIARAN | 1.634472 | 0.734635 | 2 | 1 | GRATUITO | GRATUITO | 2 | 5.055610 | 5 | 1 |
| 5 | 8954 | COLEGIO MARIA INMACULADA | 1.573501 | 0.785185 | 2 | 2 | 1.000A10.000 | MAS DE $100.000 | 2 | 5.055320 | 5 | 1 |
| 6 | 8697 | COLEGIO ORATORIO DON BOSCO | 0.964194 | 0.775010 | 2 | 2 | GRATUITO | 50.001A100.000 | 2 | 5.054286 | 5 | 1 |
| 7 | 8791 | COLEGIO FILIPENSE | 2.105011 | 0.766940 | 2 | 2 | 1.000A10.000 | 25.001A50.000 | 2 | 5.051588 | 5 | 1 |
| 8 | 8682 | COLEGIO RAFAEL SANHUEZA LIZARDI | 2.031109 | 0.761832 | 2 | 1 | GRATUITO | GRATUITO | 2 | 5.051178 | 5 | 1 |
| 9 | 25509 | COLEGIO PARTICULAR MARIA AUXILIADORA DE SANTIAGO | 1.404645 | 0.753253 | 2 | 2 | 1.000A10.000 | 50.001A100.000 | 2 | 5.050180 | 5 | 1 |
| 10 | 8636 | LICEO LEONARDO MURIALDO | 1.805726 | 0.766945 | 2 | 2 | 1.000A10.000 | 25.001A50.000 | 2 | 5.050006 | 5 | 1 |

**Figura 24:** DataFrame con las 10 mejores recomendaciones de colegios municipales.

# 6. Conclusiones

A partir de los análisis descritos en los resultados de este trabajo, se pudo validar la hipótesis propuesta, obteniendo para todos los niveles de enseñanza (1ro básico a 4to medio) mejores recomendaciones de colegios que las seleccionadas por los postulantes de la base de datos SAE, con una distancia máxima de 4 kilómetros, entre la dirección del alumno y el colegio recomendado.

Como trabajos futuros, se espera que este desarrollo pueda transferirse a una aplicación (p.ej, usando Streamlit), para que el usuario que no esta familiarizado con la programación pueda utilizar este sistema de recomendación de forma fácil y amigable. Adicionalmente se espera incluir el resto de las regiones del país, para que no solo los usuarios de la región metropolitana puedan solicitar recomendaciones de colegios, sino que cualquier persona del país pueda consultar.

Este trabajo no esta libre de limitaciones. Primero el modelo se entrena cada ves que el usuario hace una consulta, creando inicialmente la base de datos ficticia de postulaciones, y posteriormente ejecutando el modelo basado en factorización matricial. Lo cual puede verse enlentecido en computadores que cuenten con muy pocos recursos. Segundo esta metodología solo se probo en el sistema operativo Windows, no se asegura que funcione de la misma forma en macOS o Linux. Tercero, el uso de una base de datos ficticia de ranquin de colegios puede sesgar en parte los resultados, por lo que seria provechoso contar con un ranquin real, entregado por los usuarios a partir de información objetivas de los colegios (p.ej. uno que sea generado a partir de encuestas).

Finalmente, el sistema de recomendación hibrido basado en factorización matricial, desarrollado en este proyecto de grado, permitió entregar una nueva alternativa de colegio, más objetiva, con un mayor puntaje que los colegios seleccionados por los apoderados en el SAE. Utilizando información cuantitativa, geográfica y estadística de los establecimientos. Esta herramienta será de gran ayuda para los padres que estén en búsqueda de un colegio para sus hijos, con el fin de seleccionar una alternativa de colegios basada en parámetros objetivos, facilitando la toma de decisiones, y evitando el estrés que esto pueda generar.

# Bibliografía

[1] Peraita L. Cómo reducir la ansiedad de los padres cuando tienen que elegir el colegio de sus hijos. 2022. Extraído de https://www.abc.es/familia/educacion/abci-como-reducir-ansiedad-padres-cuando-tienen-elegir-colegio-hijos-202202021700_noticia.html

[2] Priyadarshani J. Parent decision-making when selecting schools: The case of Nepal. Prospects 2014; 44:411–428

[3] Hofflinger A., Gelber D., Tellez-Cañas S. School choice and parents' preferences for school attributes in Chile. Economics of Education Review. 2020; 74:101946.

[4] Gómez D., Chumacero RA., Paredes RD. School choice and information. Estudios de Economía. 2012;39(2):143-157

[5] Ricci, F., Rokach, L., Shapira, B. Recommender Systems: Techniques, Applications, and Challenges. In: Ricci, F., Rokach, L., Shapira, B. (eds) Recommender Systems Handbook 2022. Springer, New York, NY. DOI: https://doi.org/10.1007/978-1-0716-2197-4_1

[6] Cora Urdaneta-Ponte M., Mendez-Zorrilla A., Oleagordia-Ruiz I. Recommendation Systems for Education: Systematic Review. Electronics 2021;10(14):1611

[7] Beregovskaya I., Koroteev M. Review of Clustering-Based Recommender Systems. arXiv. 2021. DOI: 10.48550/ARXIV.2109.12839

[8] Icaran R. Worthwhile or not?, estimating the impacts of AI based recommendation system on Chile's School Choice system. 2020. Tesis, Magíster en Economía, IE-PUC.

https://economia.uc.cl/publicacion/worthwhile-or-not-estimating-the-impacts-of-ai-based-recommendation-systems-on-chiles-school-choice-system/

[9] Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 2022;9:59. DOI: https://doi.org/10.1186/s40537-022-00592-5

[10] Gomez-Uribe, C. A. and Hunt, N. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Transactions on Management Information Systems, 2016;6(4):13:1–13:19.

[11] Björklund, G., Bohlin, M., Olander, E., Jansson, J., Walter, C.E., Au-Yong-Oliveira, M. An Exploratory Study on the Spotify Recommender System. In: Rocha, A., Adeli, H., Dzemyda, G., Moreira, F. (eds) Information Systems and Technologies. WorldCIST 2022. Lecture Notes in Networks and Systems, 2022, vol 469. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-04819-7_36

[12] Linden, G., Smith, B., and Com, J. Y. A. Industry report: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Distributed Systems Online. 2003.

[13] Gallego, Francisco A. and Hernando, Andres, School Choice in Chile: Looking at the Demand Side (September 1, 2009). Pontificia Universidad Catolica de Chile Documento de Trabajo No. 356, 2010. DOI: http://dx.doi.org/10.2139/ssrn.1725911

[14] Farias, M. School Choice and Inequality in Educational Decisions. Multidisciplinary Journal of Educational Research, 2014; 4(1), 1–34. DOI: https://doi.org/10.4471/remie.2014.01

[15] Valentine, Daria N. How Do Parents Engage in School-Choice Decisions?. The George Washington University ProQuest Dissertations Publishing, 2016. 10146922.

[16] Limanto, S., Asmawati, E., and Putra, Y. W. K., “School finder, intelligent recommendation system for elementary school selection”, in American Institute of Physics Conference Series, 2022, vol. 2470, no. 1. DOI:10.1063/5.0080461.

[17] Chen D. Recommender System - Matrix Factorization. 2020. Extraído de https://towardsdatascience.com/recommendation-system-matrix-factorization-d61978660b4b

[18] https://www.tensorflow.org/recommenders

# 7. ANEXOS

Tabla 1: Descripción de la base de datos de directorio utilizada en este proyecto.

| **Nombre** | **Tipo** | **Descripción** | **Valores** |
|---|---|---|---|
| RBD | Numérico | Rol base de datos del establecimiento | |
| DGV_RBD | Numérico | Dígito verificador del RBD | |
| NOM_RBD | Cadena | Nombre del Establecimiento | |
| COD_REG_RBD | Numérico | Código de región en que se ubica el establecimiento | |
| COD_DEPE2 | Numérico | Código de dependencia del establecimiento (agrupado) | 1: Municipal<br>2: Particular Subvencionado<br>3: Particular Pagado (o no subvencionado) |
| RURAL_RBD | Numérico | Índice de ruralidad del establecimiento | 0: Urbano<br>1: Rural |
| LATITUD | Numérico | Coordenada de Latitud para el establecimiento | |
| LONGITUD | Numérico | Coordenada de Longitud para el establecimiento | |
| ENS_01 | Numérico | Código de enseñanza 01 | 10. Educación Parvularia<br>110. Enseñanza Básica<br>310. Enseñanza Media H-C niños y jóvenes<br>410. Enseñanza Media T-P Comercial Niños y Jóvenes<br>510. Enseñanza Media T-P Industrial Niños y Jóvenes<br>610. Enseñanza Media T-P Técnica Niños y Jóvenes<br>710. Enseñanza Media T-P Agrícola Niños y Jóvenes<br>810. Enseñanza Media T-P Marítima Niños y Jóvenes |

| | | | |
|---|---|---|---|
| | | | 910. Enseñanza Media Artística Niños y Jóvenes |
| MATRICULA | Numérico | Establecimiento con matrícula al 30 de abril | 0: No<br>1: Si |
| ESTADO_ESTAB | Numérico | Estado del establecimiento | 1: Funcionando |
| ORI_RELIGIOSA | Numérico | Orientación religiosa del establecimiento | 1: Orientación Religiosa Laica.<br>2: Orientación Religiosa Católica.<br>3: Orientación Religiosa Evangélica.<br>5: Orientación Religiosa Judía.<br>6: Otra o Sin información. |
| PAGO_MATRICULA | Cadena | Pago de matrícula en el establecimiento | SIN INFORMACION<br>GRATUITO<br>$1.000 A $10.000<br>$10.001 A $25.000<br>$25.001 A $50.000<br>$50.001 A $100.000<br>MAS DE $100.000 |
| PAGO_MENSUAL | Cadena | Pago mensual en el establecimiento | SIN INFORMACION<br>GRATUITO<br>$1.000 A $10.000<br>$10.001 A $25.000<br>$25.001 A $50.000<br>$50.001 A $100.000<br>MAS DE $100.000 |
| PROM | Numérico | Promedio general por curso, entre los años 2004 al 2020 | |
| ASIS | Numérico | Porcentaje de asistencia por curso, entre los años 2004 al 2020 | |
| APRB | Numérico | Porcentaje de aprobados respecto al total (aprobados + reprobados) por curso, entre los años 2004 al 2020 | |
| NOND | Numérico | Porcentaje de alumnos que no fueron retirados o trasladados con respecto al total de | |

| | | | |
|---|---|---|---|
| | | alumnos por curso, entre los años 2004 al 2020 | |
| PHOM | Numérico | El porcentaje de hombres con respecto a mujeres por curso, al 2020 | |
| CALU | Numérico | Cantidad total de alumnos por curso, al 2020 | |
| CCUR | Numérico | Cantidad de cursos por niveles, al 2020 | |
| CHOM | Numérico | Cantidad de hombres por niveles, al 2020 | |
| CMUJ | Numérico | Cantidad de mujeres por niveles, al 2020 | |
| PEND | Numérico | Pendiente de la regresión lineal del promedio general por niveles, entre los años 2004 al 2020 | |
| SIMCE-LEC | Numérico | Valor promedio del puntaje obtenido en lenguaje entre los años 2012 al 2018 | |
| SIMCE-MAT | Numérico | Valor promedio del puntaje obtenido en matemática entre los años 2012 al 2018 | |
| PEND-LEC | Numérico | Pendiente de la regresión lineal del valor promedio obtenido para lenguaje, entre los años 2012 a 2018 | |
| PEND-MAT | Numérico | Pendiente de la regresión lineal del valor promedio obtenido para matemática, entre los años 2012 a 2018 | |
| PROM-PUNT | Numérico | Puntaje promedio obtenido por los alumnos, que se matricularon en la universidad, entre los años 2005 y 2020, solo | |

| | | | |
|---|---|---|---|
| | | para los colegios que tengan registro de poseer 4to medio | |
| PROM-POST | Numérico | Número de alumnos promedio que se matriculan en la universidad, entre los años 2005 y 2020, solo para los colegios que tengan registro de poseer 4to medio | |
| DE-F-YEAR | Numérico | Porcentaje de deserción al primer año, esto quiere decir el alumno que no se salieron de la universidad, para ingresar al año siguiente a otra universidad | |
| PORC-MATR | Numérico | Porcentaje de matriculados con respecto a la cantidad de alumnos por curso en cuarto medio, entre los años 2005 y 2020 | |
| GEN_H | Numérico | Identificador si el nivel del establecimiento posee alumnos hombres. | |
| GEN_M | Numérico | Identificador si el nivel del establecimiento posee alumnos mujeres. | |
| COD_NIVEL | Numérico | Identificador del nivel de enseñanza que se imparte en el colegio. | 1 = 1ro Básico<br>2 = 2do Básico<br>3 = 3ro Básico<br>4 = 4to Básico<br>5 = 5to Básico<br>6 = 6to Básico<br>7 = 7mo Básico<br>8 = 8vo Básico<br>9 = 1ro Medio<br>10 = 2do Medio<br>11 = 3ro Medio<br>12 = 4to Medio |

Tabla 2: Descripción de la base de datos SAE utilizada para validación.

| **Nombre** | **Tipo** | **Descripción** | **Valores** |
|---|---|---|---|
| MRUN | Numérico | Máscara del RUN del postulante | |
| COD_NIVEL | Numérico | Identificador del nivel de enseñanza que se imparte en el colegio. | 1 = 1ro Básico<br>2 = 2do Básico<br>3 = 3ro Básico<br>4 = 4to Básico<br>5 = 5to Básico<br>6 = 6to Básico<br>7 = 7mo Básico<br>8 = 8vo Básico<br>9 = 1ro Medio<br>10 = 2do Medio<br>11 = 3ro Medio<br>12 = 4to Medio |
| RBD | Numérico | Rol base de datos del establecimiento | |
| PREFERENCIA_POSTULANTE | Numérico | Lugar de preferencia dentro del ranking declarado en la plataforma de postulación | |
| PRIORIDAD_HERMANO | Numérico | Prioridad por tener un hermano matriculado en el establecimiento | 0: No<br>1: Sí |
| PRIORIDAD_HIJO_FUNCIONARIO | Numérico | Prioridad por tener padre y/o madre trabajando en el establecimiento | 0: No<br>1: Sí |
| PRIORIDAD_EXALUMNO | Numérico | Prioridad por ser exalumno del establecimiento | 0: No<br>1: Sí |
| ES_MUJER | Numérico | Indicador si el postulante es mujer | 0: No<br>1: Sí |
| LAT_CON_ERROR | Numérico | Coordenada geográfica del hogar del estudiante con | |

| | | | |
|---|---|---|---|
| | | error aleatorio: Latitud | |
| LON_CON_ERROR | Numérico | Coordenada geográfica del hogar del estudiante con error aleatorio: Longitud | |
| AGNO | Numérico | Año de la postulación del 2019-2021. | |
| LATITUD | Numérico | Coordenada de Latitud para el establecimiento. | |
| LONGITUD | Numérico | Coordenada de Longitud para el establecimiento. | |
| DISTANCIA | Numérico | Distancia en kilómetros entre las coordenadas reportadas por el postulante y el establecimiento elegido. | |
| NOM_RBD | Cadena | Nombre del Establecimiento | |