

PROYECTO DATA SCIENCE

"Efectos COVID en la rendición de prueba de selección
universitaria y el impacto en las postulaciones a carreras de
Educación Superior"

Diciembre 2022



Análisis de la prueba de selección universitaria 2017 a 2022:

Contexto:	<p>Se pretende estudiar los efectos que podría haber provocado la pandemia Covid-19 en el proceso de acceso a la educación superior, específicamente a través de:</p> <ul style="list-style-type: none">• Los puntajes obtenidos en la prueba de selección universitaria.• Antecedentes educacionales de los y las estudiantes.• La elección de carreras según área de conocimiento.• Entre otras variables. <p>Origen de los datos(2017-2022): DEMRE y SIES</p>
Problema:	Efectos COVID en la rendición de prueba de selección universitaria y el impacto en las postulaciones a carreras de Educación Superior.
Preguntas a resolver:	<ol style="list-style-type: none">1. Determinar si el COVID es un factor relevante en el rendimiento de los puntajes de las pruebas de selección universitaria, considerando el resultado promedio de lenguaje y matemática.2. Debido a la pandemia, ¿cambió la demanda, o distribución, de carreras universitarias elegidas por los y las estudiantes al realizar la postulación a la educación superior?

Bases a analizar

Base Antecedentes Educativos

VARIABLE	TIPO DE DATO	TIPO DE VARIABLE
Puntaje promedio Lenguaje - Matemática	Númérico	Respuesta
Año del proceso de admisión	Factor	Predictora (Asociada al COVID)
Tipo de colegio de procedencia (Particular Pagado, Subvencionado y Municipal)	Factor	Predictora
Año de egreso de enseñanza media	Factor	Predictora
Puntaje Nem	Númérico	Predictora
Puntaje Historia y Ciencias	Númérico	Predictora
Puntaje Ciencias	Númérico	Predictora
Rama Educacional (HC y TP)	Factor	Predictora

- Base de datos contiene 16 variables en total
- Está compuesta por 1.487.121 registros

Base de Postulaciones (1era preferencia)

VARIABLE	TIPO DE DATO	TIPO DE VARIABLE
Año del proceso de postulación	Factor	Predictora
Región de sede de la carrera de postulación	Factor	Predictora
Nombre sede de la carrera de postulación	Factor	Predictora
Nombre de comuna de la carrera de postulación	Factor	Predictora
Área del conocimiento de la carrera de postulación	Factor	Predictora
Tipo de Institución en la que se imparte la carrera de postulación	Factor	Predictora

- Base de datos contiene 13 variables en total
- Está compuesta por 825.999 registros

Elección de Modelo

Pregunta 1: Modelo Supervisado Base Antecedentes Educativos

- Ajustaremos un Random Forest de regresión a base de "Antecedentes Educativos", teniendo en cuenta que nuestra variable target, Puntaje promedio de lenguaje y matemática, es de tipo numérica, continua.
- El modelo nos permitirá conocer la importancia relativa de cada variable.
- Determinaremos si la variable asociada al Covid (Año del proceso de Admisión) es significativa.



Pregunta 2: Modelo No supervisado Base Postulaciones

- Ajustaremos un K-Modes a la base de "Postulaciones (1era preferencia)", teniendo en cuenta que no existe una variable respuesta.
- Todas las variables de la base se pasan a factor.
- Esperamos visualizar si la variable asociada a la pandemia (Año del proceso de postulación), tuvo algún efecto en la agrupación de carreras, según áreas del conocimiento, tipo de institución y distribución geográfica.

Modelo Random Forest

Base Antecedentes Educativas

Supuestos necesarios para el modelo

Independencia

Con el fin de buscar la independencia de las variables, se eliminarán aquellas que presenten alto grado de correlación.

Normalización

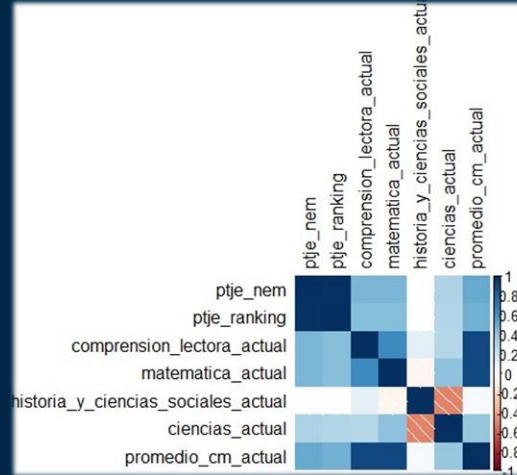
Considerando que las variables numéricas del modelo Random Forest seleccionado, se encuentran en la misma escala, éstas no se normalizarán.

¿Qué va a responder este modelo?

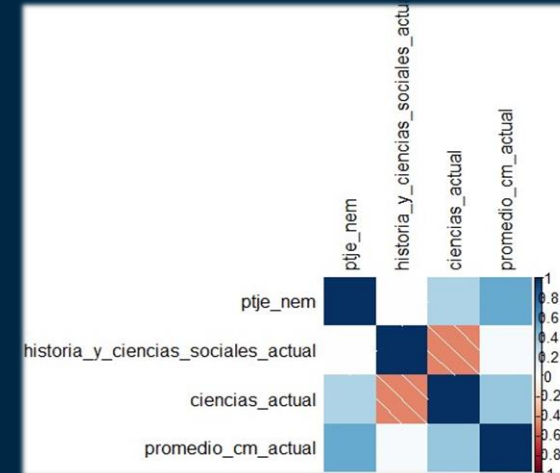
Efecto Covid

Si el efecto Covid tuvo impacto en los Puntajes promedio de lenguaje y matemática, a través de la importancia de cada variable.

Eliminación de variables numéricas



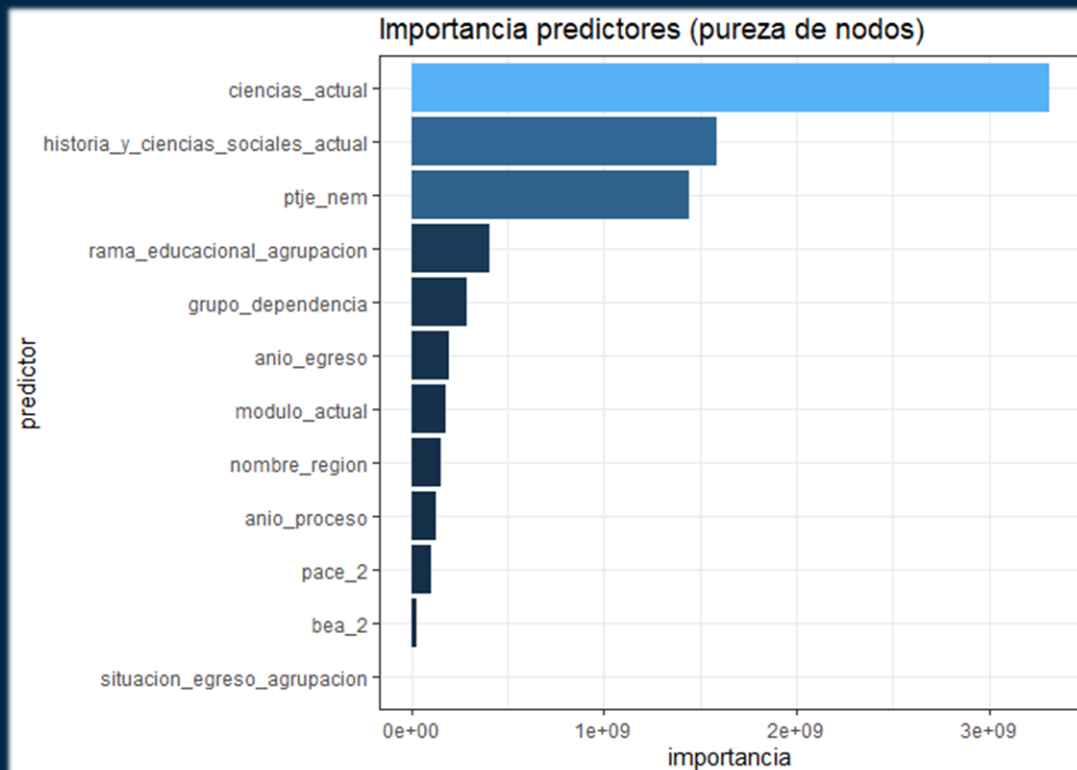
Elección de variables numéricas



Ajuste Modelo Random Forest

Base Antecedentes Educativas

Type	Regression
Number of trees	500
Simple size	1.040.984
Number of independent variables	12
Mtry	4
Target node size	30
Splitrule	Variance
R squared (OOB)	0,717964
OOB prediction error (MSE)	2.713,497
Best rmse	52,1



Interpretaciones

Se verificó que la variable año del proceso de admisión no influyó significativamente en el puntaje promedio de las pruebas de selección universitaria de lenguaje y matemática.

Se comprobó que la variable puntaje ciencias influye con mayor relevancia en el puntaje promedio de lenguaje y matemática.



El promedio de los puntajes de lenguaje y matemática es explicado en un 72% por las variables predictoras y el valor rmse más bajo obtenido fue de 52,1 puntos.

Considerando solo las variables de tipo factor, se interpretó que la variable rama educativa (TP y HC) es la más relevante a la hora de explicar la variable respuesta.

Ajuste Modelo K-Modes

Base Postulaciones primera preferencia

El ajuste al modelo K-Modes se realizó considerando 825.999 registros, y las variables categóricas;

- Año del proceso de postulación
- Región
- Área del conocimiento
- Tipo de institución

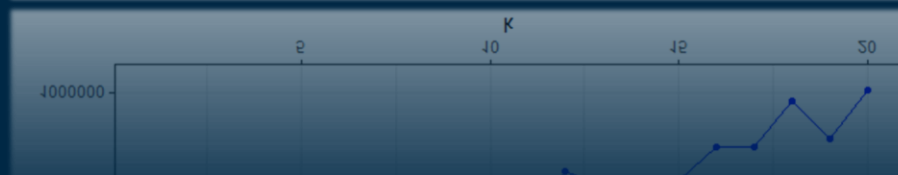
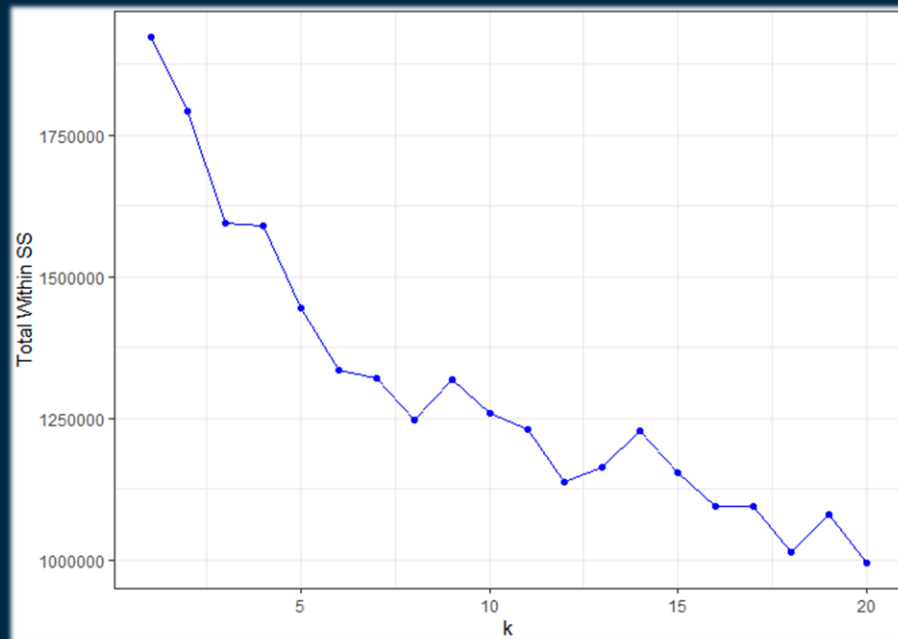
Se atendió el criterio del “Gráfico del Codo” para determinar la cantidad de grupos, siendo en este caso $K = 2$

Como resultado del ajuste se obtuvieron dos clústeres de tamaños:

- Cluster_1: **571.120**
- Cluster_2: **254.879**

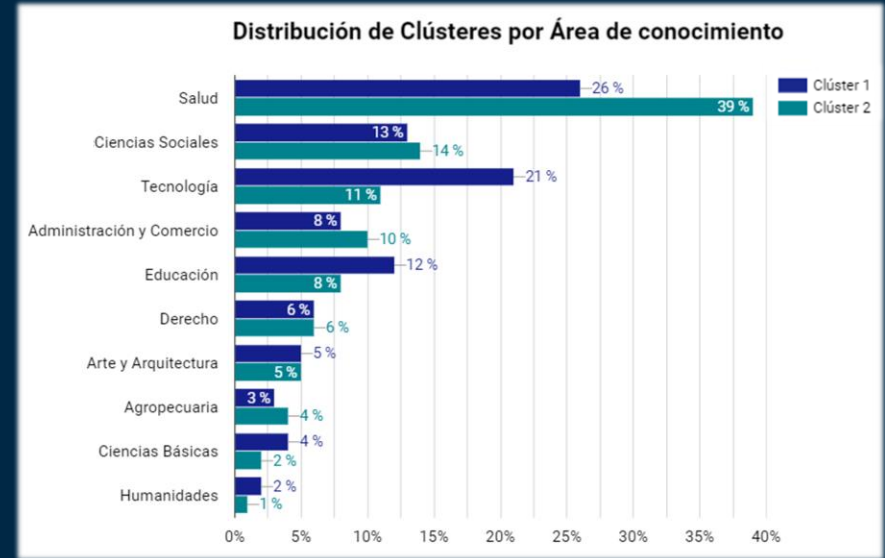
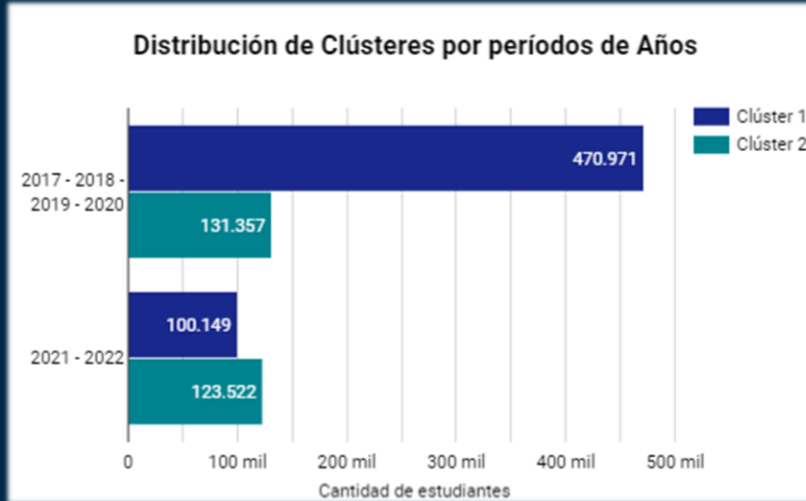
Siendo las modas del ajuste:

Variable	Cluster_1	Cluster_2
Año del proceso de postulación	2020	2022
Región	R. Metropolitana	R. Metropolitana
Área del conocimiento	Salud	Salud
Tipo de institución	Univ. CRUCH	Otras universidades



Ajuste Modelo K-Modes

Análisis por demanda



Se definieron los años de los procesos de selección previos a la pandemia; 2017, 2018, 2019 y 2020 como años pre pandemia y los años 2021 y 2022 como años pandemia.

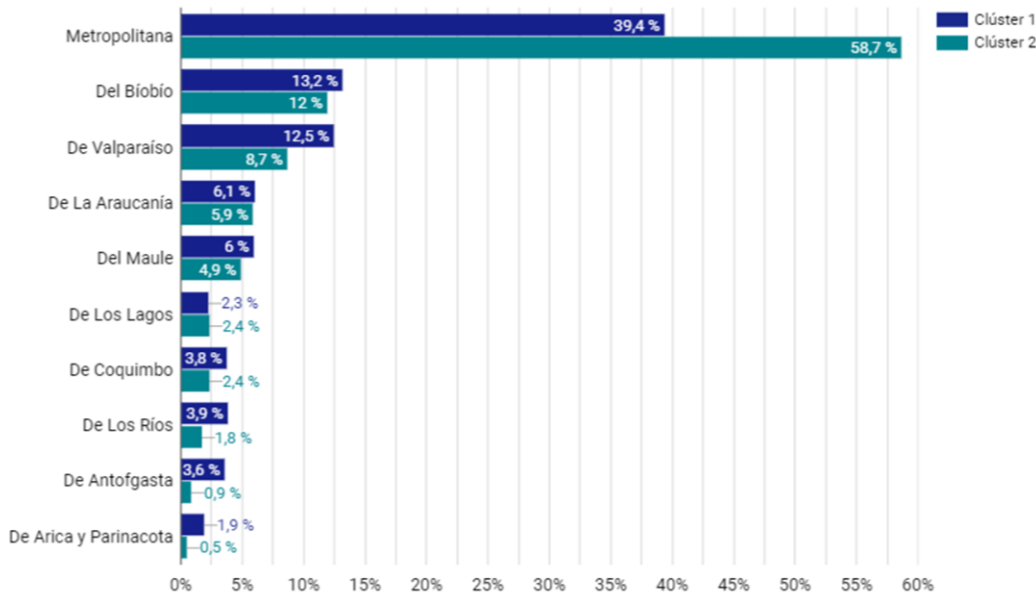
Luego del ajuste se denominan los clústeres encontrados como:

- Clúster Pre-Pandemia (Clúster 1)
- Clúster Pandemia (Clúster 2)

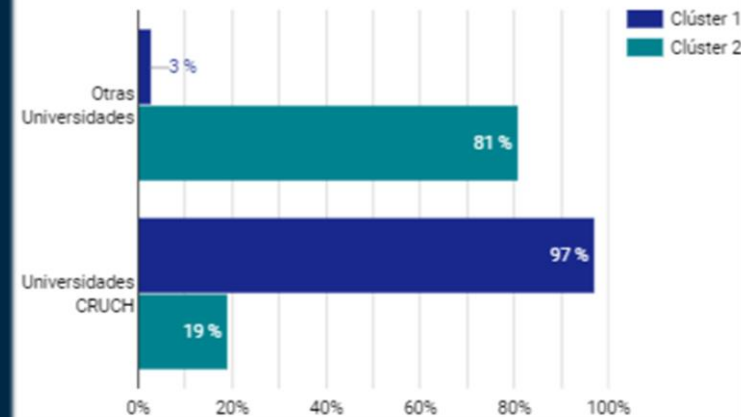
Ajuste Modelo K-Modes

Análisis por distribución

Distribución de Clústeres por Región



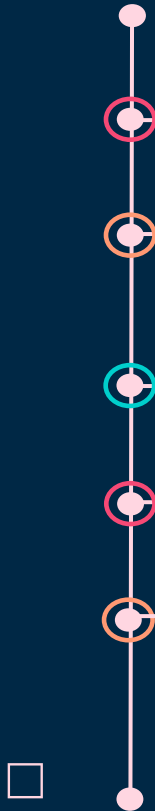
Distribución de Clústeres por Tipo de Institución



Caracterización de los clústeres

Clúster Pre-Pandemia	Clúster Pandemia
<p>Las postulaciones en los años previos a la pandemia se concentran en mayor proporción. 470.971 postulaciones v/s 131.357 en el Clúster Pandemia.</p>	<p>Las postulaciones en los años correspondientes a la pandemia se concentran en mayor proporción. 123.522 postulaciones v/s 100.149 en el Clúster Pre-Pandemia.</p>
<p>El postulante opta preferentemente por carreras del área de la salud y tecnología. 47% de las postulaciones.</p>	<p>El postulante opta preferentemente por carreras del área de la salud. En el área de la salud las postulaciones suben de 26% en el Clúster Pre-Pandemia a 39% en el Clúster Pandemia. En tecnología bajan de 21% a 11%.</p>
<p>Tiene mayor inclinación por estudiar en universidades CRUCH. 97% de las postulaciones.</p>	<p>Tiene mayor inclinación por estudiar en otras universidades que no pertenecen al CRUCH. 81% de las postulaciones.</p>
<p>Las postulaciones se concentran mayoritariamente en la Región Metropolitana. 39% de las postulaciones.</p>	<p>Las postulaciones se concentran mayoritariamente en la Región Metropolitana, siendo esta probabilidad mayor que en el Clúster Pre-Pandemia. 59% de las postulaciones.</p>

CONCLUSIONES

- 
- La pandemia no generó gran impacto en los puntajes promedio de lenguaje y matemática en comparación con los períodos pre pandemia.
 - Se verificó un cambio en el comportamiento de las postulaciones de primera preferencia. Específicamente en el área de la salud se evidencia un aumento en la demanda de estas carreras en comparación con las otras áreas en tiempos de pandemia.
 - En relación a la distribución por región se evidencia un cambio en las postulaciones, destacando en ambos grupos que la postulación se concentra mayoritariamente en la Región Metropolitana.
 - En cuanto a la distribución por tipo de institución, se observó diferencias en los periodos analizados, sobresaliendo la preferencia por universidades que no pertenecen al CRUCH en periodos de pandemia.
 - Como propuesta para profundizar más en este estudio, se propone ajustar un Árbol de Clasificación al Clúster Pandemia para pronosticar si un alumno estudiará en una ciudad determinada considerando algunas de sus preferencias, esto enmarcado en un contexto de pandemia o alguna situación que determine condiciones similares (restricción de movilidad, conflictos sociales, incertidumbre generalizada, etc.)