Estudio del rendimiento estudiantil en Chile

CC5205-3
Grupo 3 - otoño 2023

Profesora: Jazmine Maldonado
Auxiliar: Cinthia Sánchez Macías
Integrantes: Allen Arroyo, Benjamín Angulo , José Badilla , Bárbara Aguayo , Benjamín Llancao

Proyecto

Determinar patrones y perfilamientos sobre el rendimiento estudiantil en Chile en enseñanza básica y media.

Motivación

- Desigualdad social/económica en la educación, Pandemia COVID-19, Género y Políticas públicas.
- Información obtenida del centro de estudios del MINEDUC, Datos Abiertos.

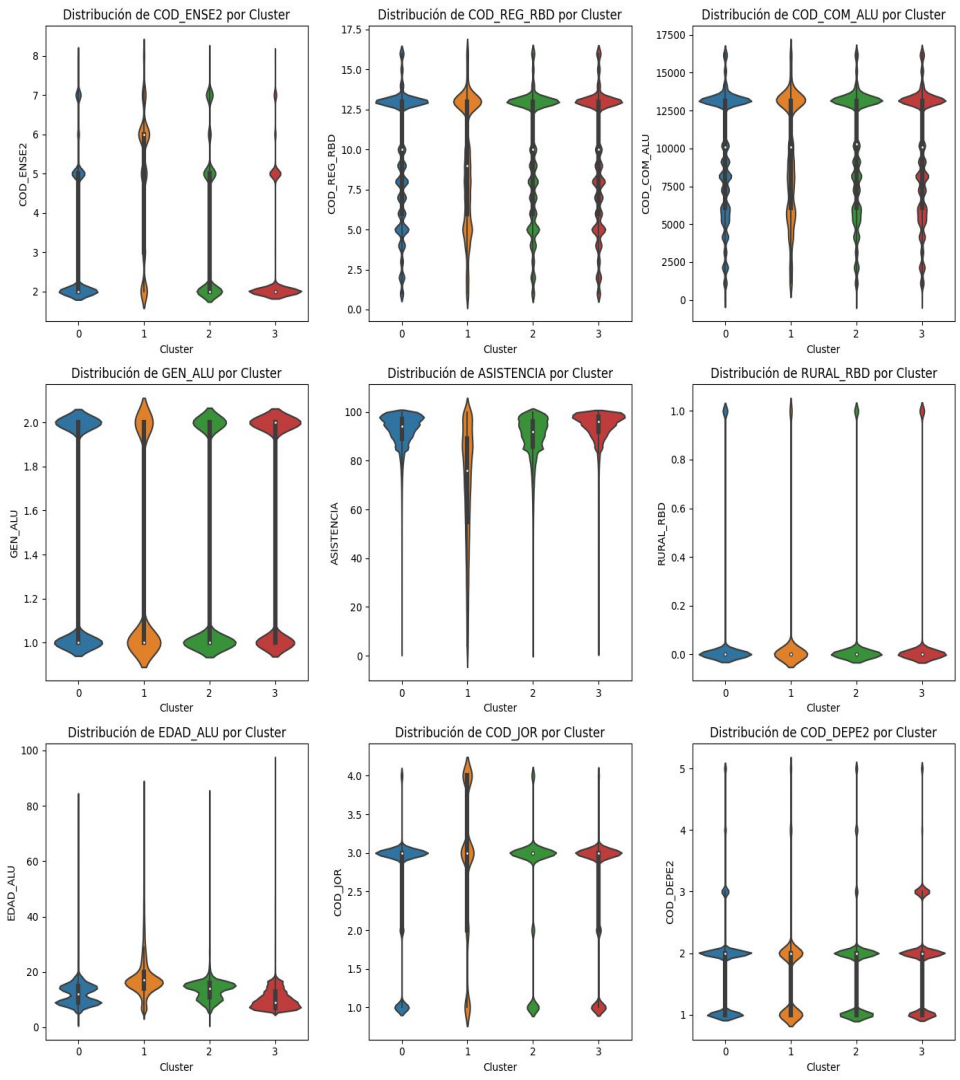
Preguntas

- ¿Existen factores que influyen más en el promedio general del estudiante?¿Es la región una de ellas?
- ¿Es posible predecir qué estudiante tendrá un mejor rendimiento a lo largo de los años? ¿Cuáles son las variables más relevantes?
- ¿Los colegios con mayor equidad en distribución entre hombres y mujeres tiene mejores rendimiento?

Ideas iniciales

Pregunta 1 : ¿Existen factores que influyen más en el promedio general del estudiante?¿Es la región una de ellas?

- Búsqueda de formas de agrupación de estudiantes
- Clusterización de los grupos de estudiantes a través de 3 métodos.
  - **Kmeans**
  - DBScan
  - Clustering Aglomerativo
- Análisis de los clusters y estudio de su variación según factores de estudio



¿Qué nos quieren decir estos clusters?

Resultados

- Se generan 4 clusters con el método del codo
- Clusterización de los grupos de estudiantes
- No se observa una diferencia tangible con la data obtenida

Promedios generales para 4 clusters:

1. 5.761254
2. 3.388550
3. 4.992294
4. 6.481823

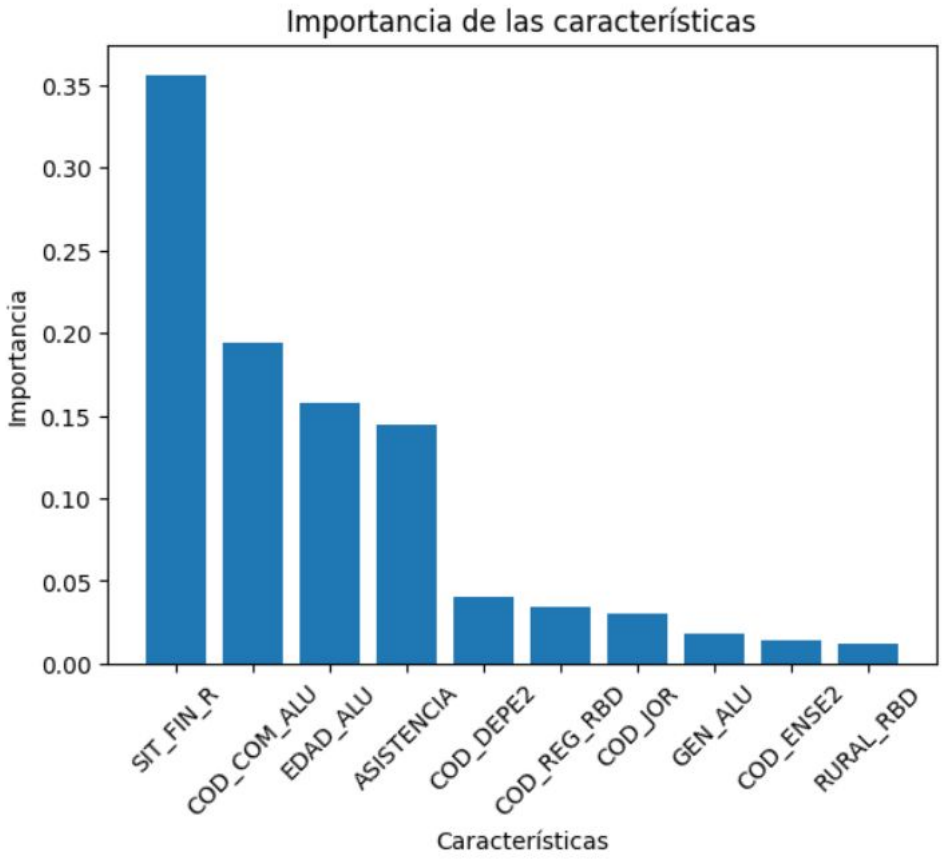
- La data observable no es lo suficientemente significativa para determinar si el promedio general de un estudiante depende de los factores de estudio.
- Es necesario obtener más fuentes de información para poder obtener una respuesta o otros métodos.

K=4

SSE: 181485.23652405717
Índice de Davies-Bouldin: 127.04330961736466
Índice de Calinski-Harabasz: 445.737955012606

Experimento P. 1

- Se realiza modelo de regresión.
- Modelo Random Forest.
- Sampling a ¼ de los datos.
- MSE = 0.36



Experimento P. 2

- La clusterización no dio resultados
- Debemos incorporar nuevos datos
  - Evaluación Docente para el año 2018

Tipo de establecimiento	Cantidad
Municipal	19259
Particular	1812
Subvencionado	974

	RBD	OC_GENERO	NIVEL	AE_PJE	EP_PJE	IRT_PJE	PF_PJE	PF_ESC	INSTR_PJE	CCE_ESC
0	1	1.0	7.0	4.0	3.0	3.0	2.23	B	2.64	C
1	1	1.0	6.0	4.0	3.0	4.0	2.40	B	2.84	C
2	1	2.0	4.0	4.0	3.0	3.0	2.58	C	2.85	D
3	1	1.0	7.0	4.0	3.0	2.0	2.52	C	2.71	C
4	1	1.0	4.0	4.0	2.0	3.0	1.80	I	2.18	B
...	...	...	...	...	...	...	...	...	...	...
22040	40429	2.0	5.0	3.0	2.0	3.0	2.05	B	2.23	B
22041	40429	1.0	6.0	4.0	3.0	4.0	1.75	I	2.45	B
22042	40429	2.0	7.0	4.0	2.0	3.0	2.09	B	2.35	B
22043	40429	1.0	6.0	3.0	2.0	3.0	1.85	I	2.11	B
22044	40429	1.0	6.0	4.0	2.0	3.0	2.02	B	2.31	B

22045 rows × 10 columns

Nivel	Ed. Parv	Primer Ciclo	Segundo Ciclo	Ed. Media	Ed. Especial	Ed. Adultos
Cantidad	1888	4868	7068	3546	3703	377

Experimento P. 2

Corresponden a Ed. Media
Colegio con RBD = 1

	RBD	COD_ENSE2	COD_REG_RBD	COD_COM_ALU	GEN_ALU	PROM_GRAL	ASISTENCIA	RURAL_RBD	EDAD_ALU	COD_JOR	COD_DEPE2	SIT_FIN_R
0	1	8	15	15101	2	6.4	95	0	34.0	4	1	P
1	1	8	15	15101	1	5.3	90	0	52.0	4	1	P
2	1	8	15	15101	2	6.4	97	0	41.0	4	1	P
3	1	8	15	15101	1	6.0	94	0	23.0	4	1	P
4	1	8	15	15101	2	5.8	90	0	39.0	4	1	P
...	...	...	...	...	...	...	...	...	...	...	...	...
2997210	40429	8	1	1101	2	6.2	86	0	23.0	4	1	P
2997211	40429	8	1	1101	2	5.1	49	0	20.0	4	1	P
2997212	40429	8	1	1101	2	5.8	79	0	17.0	4	1	P
2997213	40429	8	1	1101	2	5.7	90	0	28.0	4	1	P
2997214	40429	8	1	1404	2	5.0	85	0	19.0	4	1	P

342238 rows × 12 columns

RBD	AE_PJE	EP_PJE	IRT_PJE	PF_PJE	INSTR_PJE	CCE_ESC	PF_ESC
1	4	2	3	2	3	C	B

RBD	AE_PJE	EP_PJE	IRT_PJE	PF_PJE	INSTR_PJE	CCE_ESC	PF_ESC
1	4	2	3	2	3	C	B

Profesores de Ed. Media
Colegio con RBD = 1

Experimento P. 2

- Algoritmo DecisionTreeRegressor (Regresión)
- 70% Datos de entrenamiento y 30% de prueba
- Variable objetivo continúa (Promedio General)
- Con y sin evaluaciones docentes para el año 2018

Utilizamos las siguientes metricas usuales:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Median Absolute Error (MedAE)
- Explained Variance Score (EV)
- R-squared (coeficiente de determinación)

- Valores para 2018(sin atributo SIT_FIN_R sin evaluación)

```
Mean Squared Error en test set: 4.4207477693804913e-23
Mean Absolute Error: 4.891732453243697e-12
Median Absolute Error: 3.608668919241609e-12
Explained Variance Score: 1.0
R-squared: 1.0
```

- Valores para 2018(sin atributo SIT_FIN_R con evaluación)

```
Mean Squared Error en test set: 4.898482992763825e-25
Mean Absolute Error: 5.343745056799431e-13
Median Absolute Error: 5.648814749292796e-13
Explained Variance Score: 1.0
R-squared: 1.0
```

Experimento P. 3

- Cálculo de el factor de equidad por establecimiento
- Se realiza un modelo de regresión lineal por año

2022

Coeficiente del Factor de equidad
3.048e-05

R-squared
0.264

2020

Coeficiente del Factor de equidad
0.0003

R-squared
0.216

2018

Coeficiente del Factor de equidad
0.0002

R-squared
0.229

Futuras Direcciones

- **Camino a seguir:**

1. Dado el modelo de regresión lineal del experimento P.1.

Entender de qué manera exactamente influye la comuna en los promedios generales.

2. Dado el algoritmo de regresión del experimento P.2.

Realizar el mismo experimento pero no solamente en establecimientos públicos y además trabajar con más años que sólo 2018.

Probar con más algoritmos de regresión y no solo DecisionTreeRegressor.

3. Dado el modelo de regresión lineal y factor de equidad del experimento P.3.

No existen buenos resultados con el dataset inicial por ello es conveniente a futuro utilizar los datos de las evaluaciones docentes.

- **Recomendaciones:**

Buscar otros dataset con mayor información socio-económica de cada estudiante, informarse de qué métodos son los más adecuados para cantidades grandes de datos, concentrarse en enseñanza media o básica.