INFORME FINAL

Relación Notas de EM - Puntaje PSU/PDT Construyendo un Modelo de Regresión

Claudio Rojas
Mario Vilches

Introducción

El presente informe busca dar cuenta de nuestra propuesta de investigación para el curso Fundamento en Ciencias de Datos.
La indagación nace en un contexto de un colegio particular pagado, que tiene como desafío y exigencia la mejora de sus resultados en pruebas estandarizadas, que han ido disminuyendo en los últimos años.
Si bien, la explicación de este problema es multicausal, optamos por explorar una respuesta desde los resultados académicos de los estudiantes. Esto además, cobra relevancia en un contexto, en el cual, hay un escaso uso de datos para el análisis y la toma de decisiones.
Esta indagación se realizará a través del Mapa Metodológico trabajado en clases. Por lo cual, a partir de una pregunta, se buscará predecir los resultados futuros, basado en la información que entrega una base con datos que son numéricos y estáticos, y sobre los que se aplicará una Regresión.

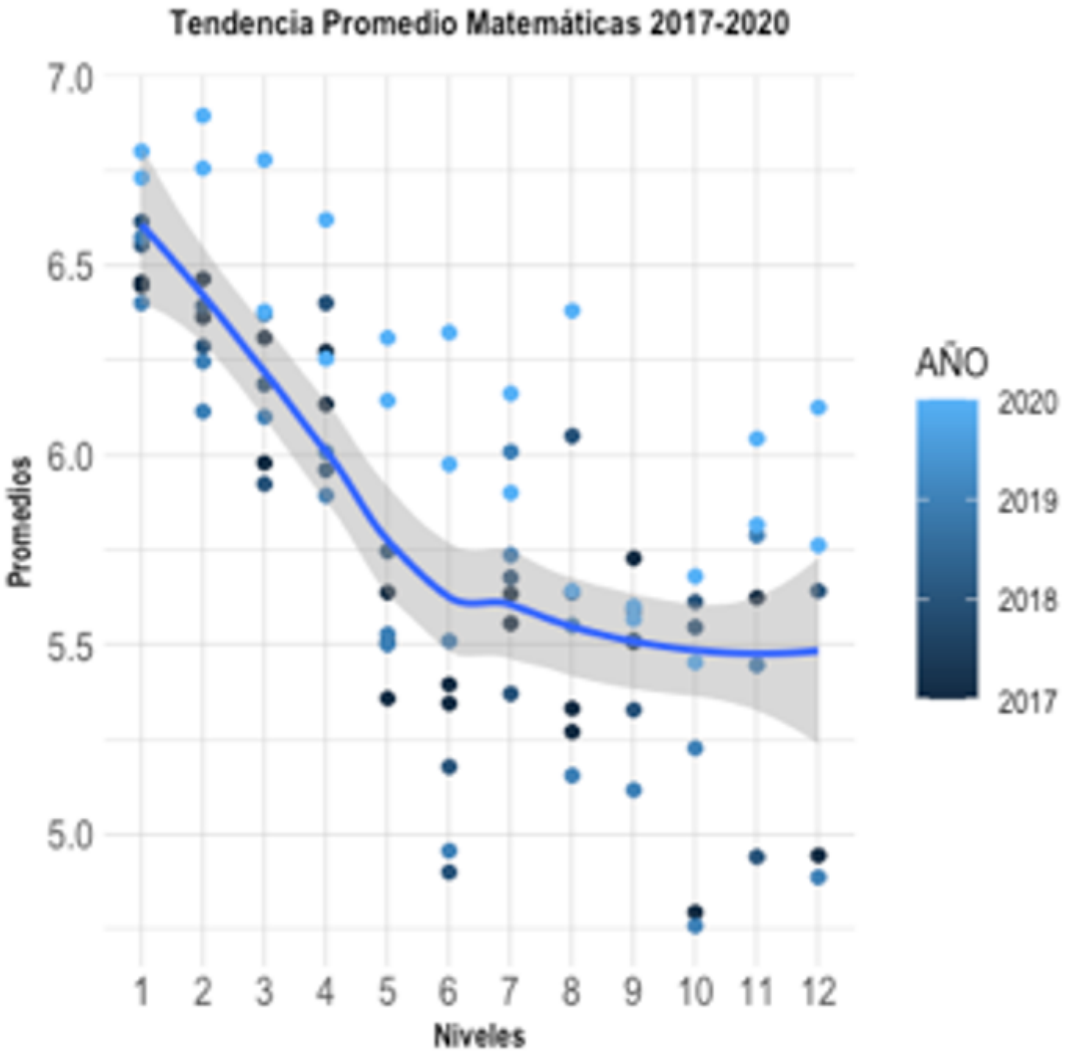
Contexto Académico del Colegio

Algunas características del contexto escolar:

a) Disminución de los resultados académicos de los estudiantes en la medida que avanza en su trayectoria académica y que tiende a una meseta en la medida que se avanza a cuarto medio.

Figura 1:

Tendencia Promedio de Matemáticas 2017-2020



Fuente: Información Interna del Colegio, 2020.

b) En la prueba PSU/PDT de matemáticas el colegio, obtiene los mejores promedios de generación e individual todos los años a diferencia de las otras pruebas (Lenguaje, Historia y Ciencias) que han tenido mayor variabilidad en los resultados obtenidos en la PSU/PDT. Si bien son puntajes altos, son puntajes que han ido disminuyendo a lo largo de los años.

Tabla 1:

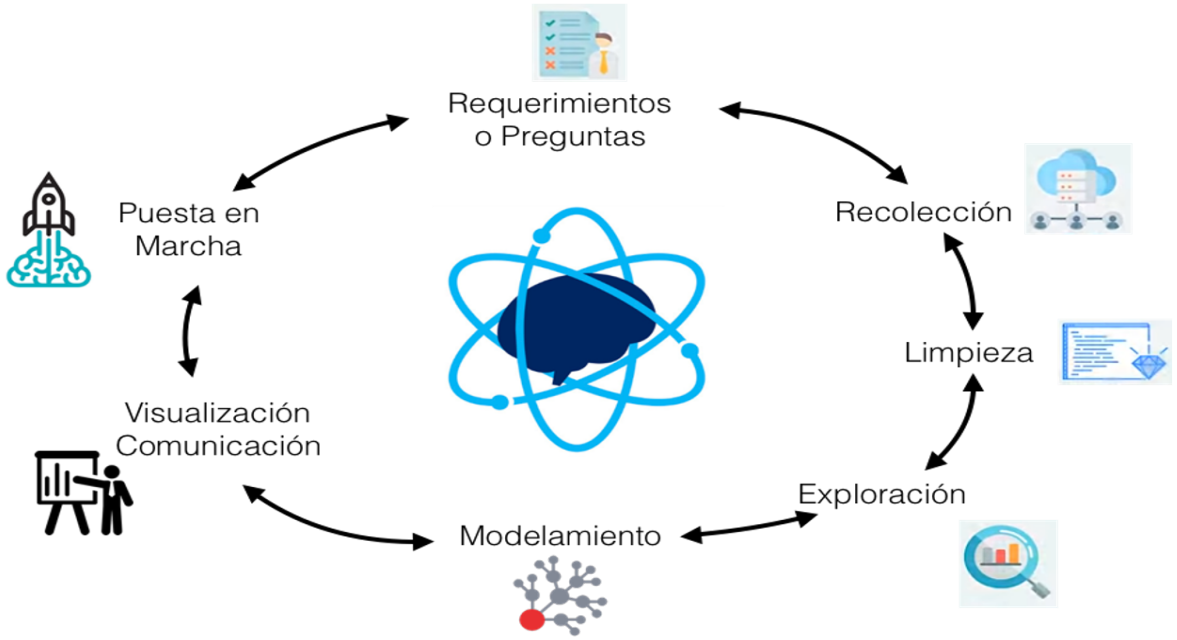
Tabla de resultados PSU/PDT (2015 - 2020)

	Lenguaje	Matemáticas	X PSU	Historia	Ciencias
2015	624	650	637	663	640
2016	630	636	633	619	633
2017	621	628	625	644	636
2018	593	618	605	598	594
2019	605,6	622	613,9		617,9
2020	595	621	608	583	610

Información Interna del Colegio, 2021

Mapa Metodológico

Para abordar esta investigación, utilizaremos el mapa metodológico aprendido en clases, que tiene las siguientes características:



a) **Requerimientos o Preguntas**

Si bien, la forma de abordar las preguntas implica considerar varios aspectos de los que ocurren en un centro educativo, se ha decidido utilizar el análisis de datos- en este caso las notas de los estudiantes y puntajes PSU/PDT- para identificar las relaciones existentes y con ello dar respuestas a las inquietudes planteadas.

Las preguntas que guían esta investigación son:

¿Los promedios obtenidos en la asignatura de Matemáticas entre primero y cuarto medio son predictores del puntaje que un estudiante logra en la PSU/PDT? En el caso de que no se cumpla lo anterior: ¿Qué promedios de asignatura o cursos son predictores del puntaje que obtiene en la PDT/PSU de Matemáticas?

b) Recolección

Para responder las preguntas planteadas, se elaboró una base de datos con las notas de enseñanza media en cada una de las

asignaturas que cursaron en enseñanza media las últimas 5 generaciones de estudiantes (2016-2020) del colegio en cuestión.
La característica principal de estos datos es que son numéricos.

Esta base de datos contiene la siguiente información:
- Nombre del Estudiante.
- Sexo (considerando que sólo la generación 2020 es mixta)
- Promedio General de cada uno de los cuatro años de Enseñanza media.
- Porcentaje de asistencia final de cada uno de los cursos de Enseñanza media.
- Promedios finales de cada una de las asignaturas que cursaron los estudiantes durante cada uno de los cuatro años de enseñanza media.
- Puntaje PSU/PDT en Lenguaje, Matemáticas y en la prueba específica que realizaron.

Todos los estudiantes que se encuentran en la base de datos cursaron en el establecimiento, al menos, desde Iº a IVº medio y además, rindieron la PSU/PDT y entregaron su puntaje al colegio.

c) Limpieza

Luego de construida la base de datos se procedió a realizar una cura de esta que consistió en:
- Eliminar los estudiantes que no hayan cursado los cuatro años en el colegio por falta de datos en algunos de los niveles.
- Eliminar los estudiantes que no tenían puntaje PDT, ya sea porque no rindieron la prueba o porque hacen uso de su derecho de no permitir que el colegio acceda a sus puntajes. Esto se ha vuelto frecuente desde el 2016 en adelante.
- Unificar el formato de los datos: *float*, *int64*, *object*.
- Establecer algunos diccionarios para algunos resultados.

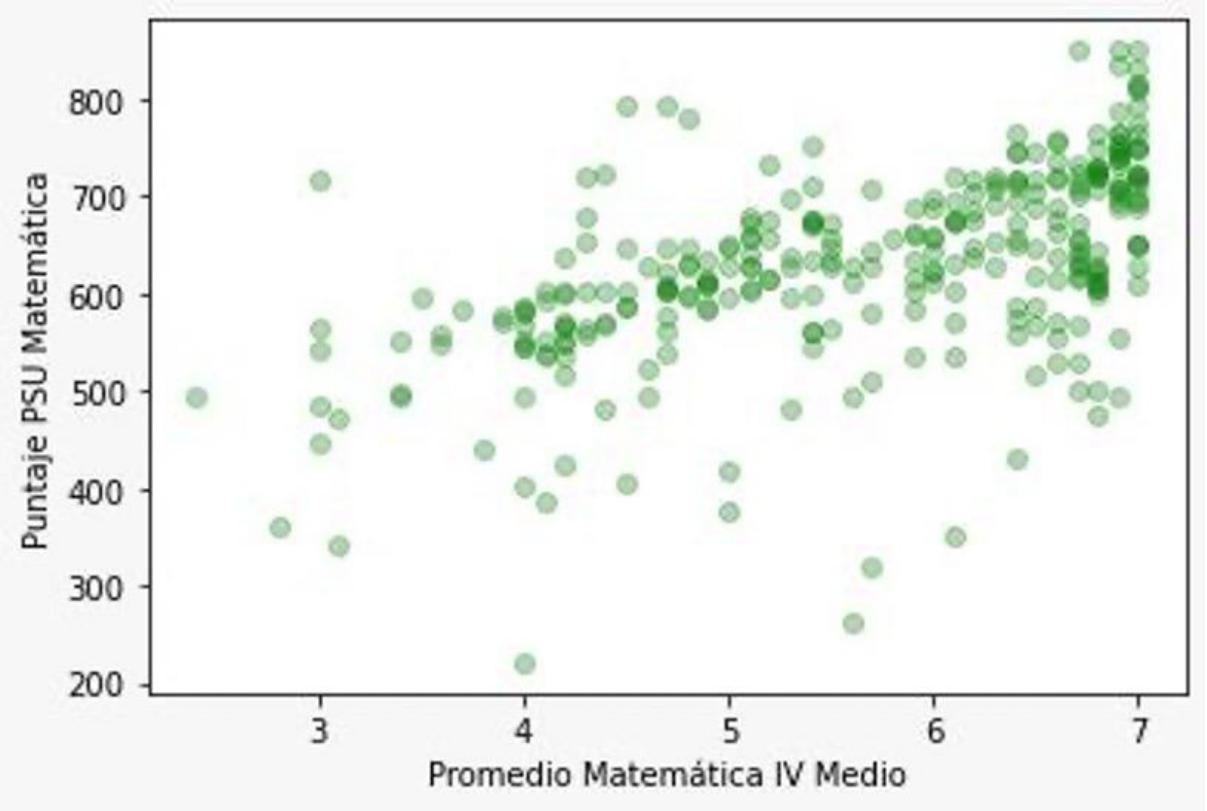
d) Exploración

Para efecto de este informe, se procedió a explorar los datos a través de la realización de algunos gráficos que permitieran acercarnos a la investigación que estamos realizando y establecer las primeras conclusiones de esta.

- *Exploración 1:* Se graficó el promedio de matemáticas de IV medio (eje x) con el puntaje PSU/PDT obtenidos por los estudiantes (eje y), que se aprecia a continuación (Figura 2). Al observar la distribución de los datos en el gráfico, se aprecia:
  - Un número importante de datos que se correlacionan.
  - Un número de datos que podrían considerarse outliers.
  - Estudiantes con promedios sobre 6,5 o más no logran superar los 700 puntos en la prueba.

Figura 2:

Correlación entre Promedio de IV medio en la asignatura de Matemáticas con Puntajes PDT Matemáticas (2016-2020)



- Exploración 2: En este caso, se procedió a correlacionar el promedio de IV° medio de la asignatura de Lenguaje y Comunicación (eje x) con el Puntaje de PDT en Matemáticas (eje y).

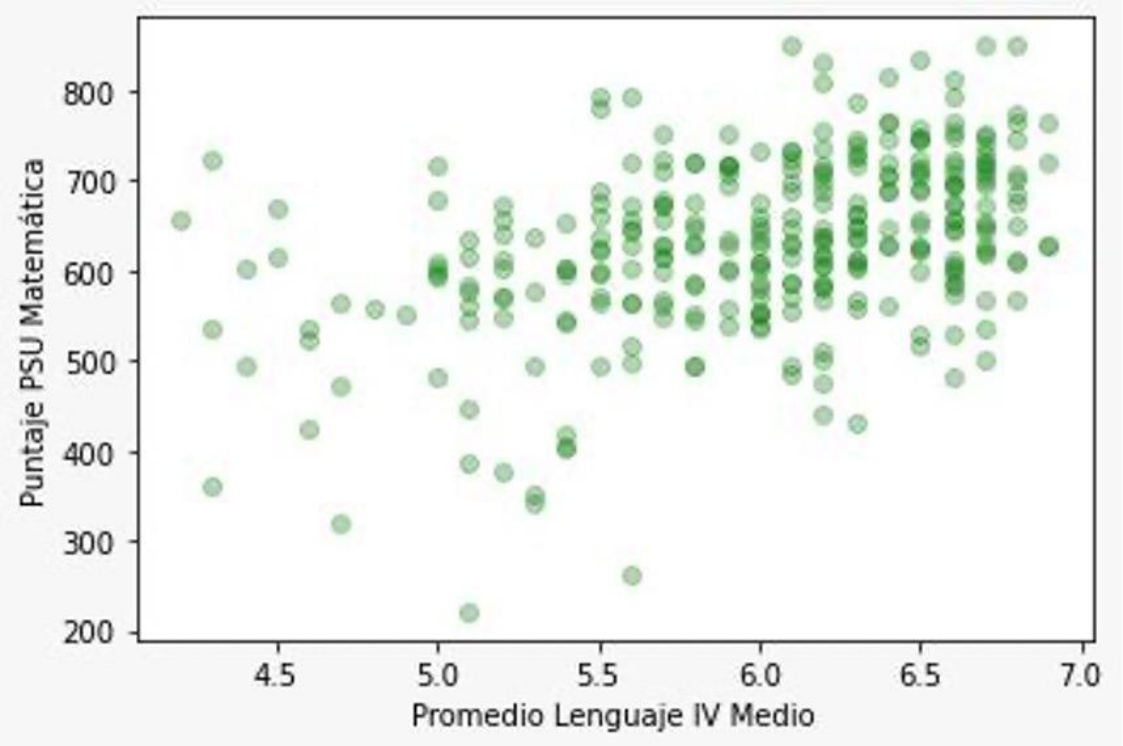
Se consideró los resultados de la asignatura porque como señala el estudio de Susan Goldin-Meadow (2011) sugiere que el lenguaje juega un importante rol en la adquisición del número.

Lo que se aprecia en la Figura 3 es que:

- Una dispersión de los datos.
- Estudiantes con promedios sobre 6,5 o más no logran superar los 700 puntos en la prueba. predomina la dispersión de datos. Los estudiantes, que obtienen un promedio final igual o superior a 6,0 en la asignatura obtienen en su mayoría puntajes en la PDT de matemáticas menores a 700 puntos.

Figura 3:

Correlación entre Promedio de IV medio en la asignatura de Lenguaje con Puntajes PDT Matemáticas (2017-2020)



- Exploración 3:

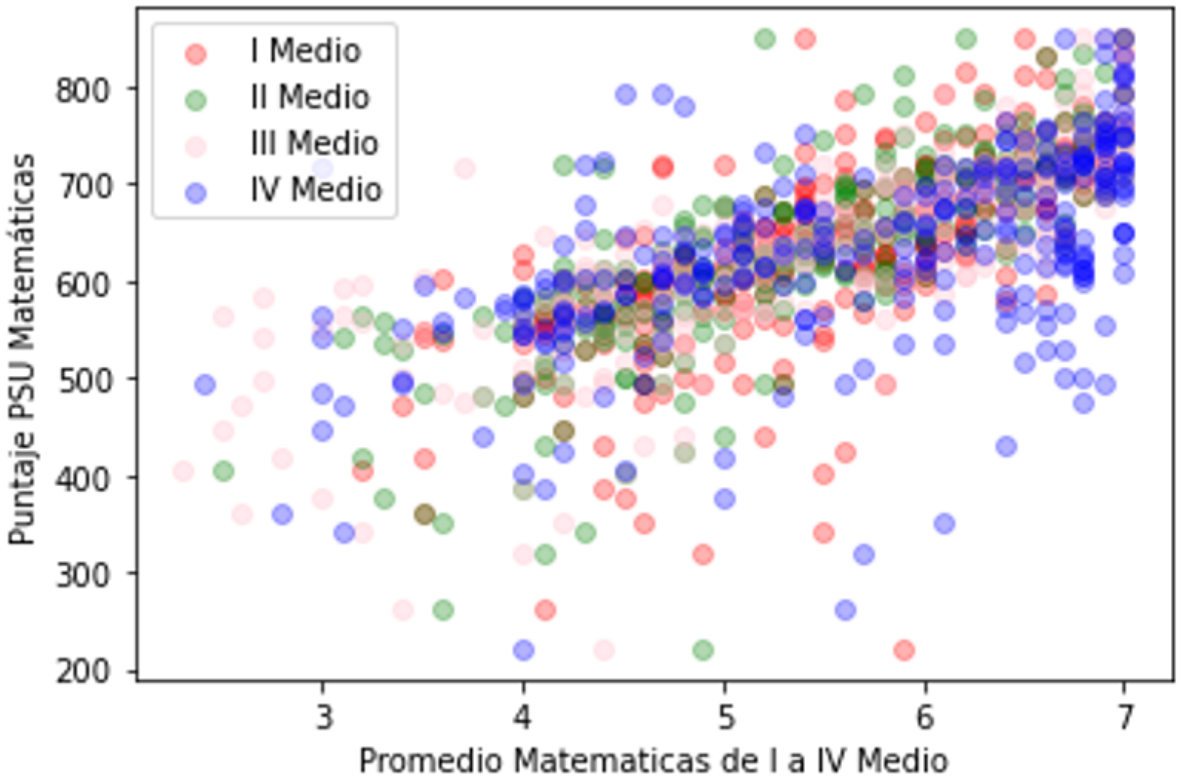
En este caso, lo que realizamos fue correlacionar (Figura 4) las notas de la asignatura de Matemáticas en cada una de los cursos de enseñanza media (eje x) con el puntaje PSU/PDT de la misma asignatura (eje y).

Lo que se aprecia es que:

- La existencia de una correlación desde la nota 4 en adelante.
- La presencia de algunos datos que podrían denominarse como outliers.
- Los estudiantes con promedios igual o superior a 6,5 mayoritariamente no logran superar los 700 puntos.

Figura 4:

Correlación entre Promedio de I a IVº medio la asignatura de Matemáticas con Puntajes PDT Matemáticas (2017-2020)

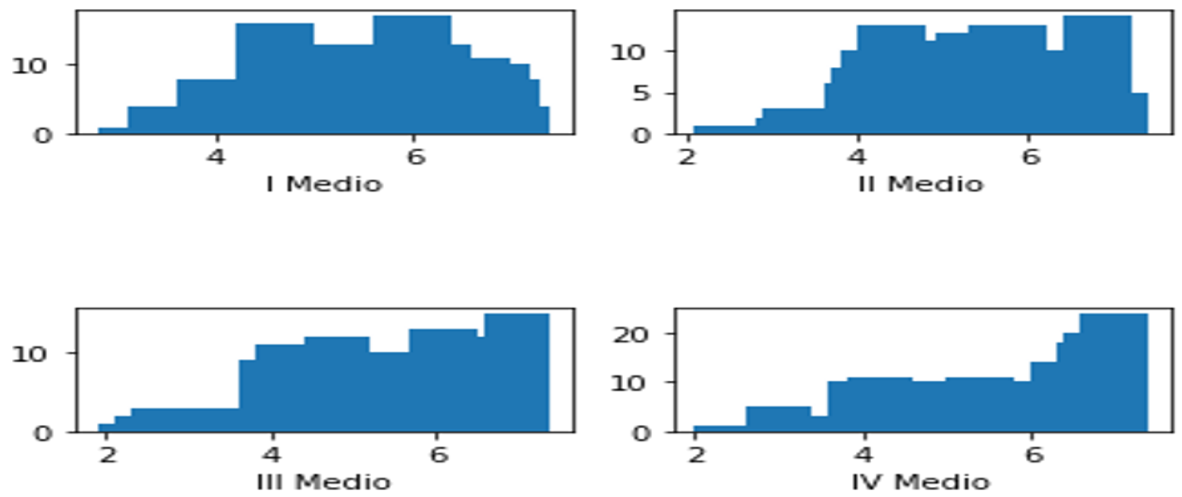


- Exploración 4

Por último, también nos interesó ver la distribución de notas de matemáticas en cada nivel (Figura 5).

Figura 5:

Histogramas promedio de Matemáticas I a IVº medio(2017-2020)



Como podemos observar, las distribuciones transitan desde una doble normal con medias en las notas 5,0 y 6,0 hacia una distribución cargada hacia la derecha, es decir, hacia las mejores notas, por tanto, creemos que los cursos que mejor representan el rendimiento de los alumnos son II y III medio, ya que I medio es un curso de transición a la enseñanza media y IV medio es un curso donde el énfasis está en la graduación del colegio, y por tanto, más “relajado” académicamente.

Este descubrimiento es análogo con lo ya presentado por el gráfico de la exploración anterior, lo que nos da muy buenos indicios de que podremos encontrar una relación que pueda predecir los puntajes de la prueba de matemáticas para el ingreso a la universidad con los resultados de los alumnos en II y III medio.

Por lo tanto, nuestra conclusión - de este primer informe- es que los estudiantes que obtienen promedios sobre 6,0 en las

asignaturas de lenguaje o matemáticas tienden a no tener correlación con el puntaje PDT en matemáticas, pero así también es posible ver que los promedios en la asignatura de matemáticas de II y III Medio pueden explicar mejor los resultados de la PDT de matemáticas, lo que creemos que se produce, pues es en estos cursos donde se encuentran las materias que permiten responder de manera correcta una mayor proporción de la PDT.

e) Modelado: Construcción y visualización.

Paso 1: Definición

Decidimos diseñar un modelo de regresión lineal que permita predecir el puntaje de PSU/PDT en matemáticas a partir de los promedios finales de las distintas asignaturas cursadas en cada uno de los años de la enseñanza media.

Para poder realizar lo anteriores, volvimos a revisar nuestra base de datos y decidimos descartar:

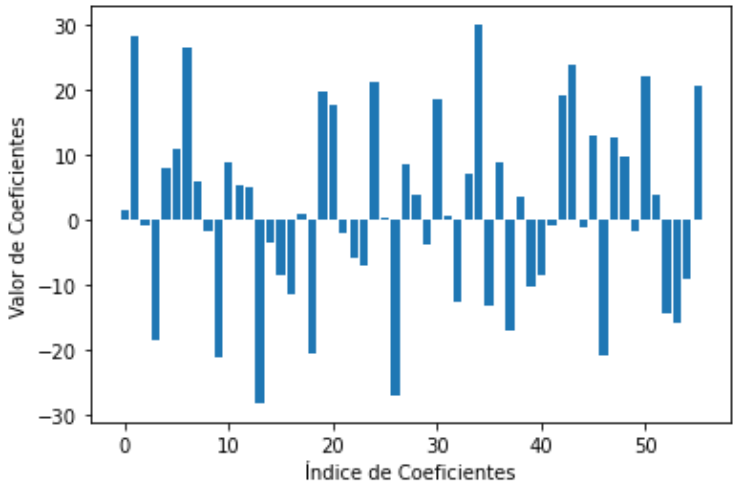
- Atributo sexo porque que sólo la generación 2020 contaba con ellas.
- El promedio PSU/PDT porque se compone en un 50% por el puntaje obtenido en Matemáticas.
- El nombre del alumno porque no es una variable numérica.
- El puntaje PSU/PDT se obtiene al mismo momento que la variable a explicar.
- Los puntajes que estuvieran fuera de +- 2 desviaciones estándar de la media (18 de 298 observaciones).

Paso 2: Primera Regresión y Correlación de Variables

Realizamos una primera regresión lineal con los 56 atributos que teníamos a disposición para tratar de explicar el modelo y graficamos los coeficientes de esta regresión.

Figura 6:

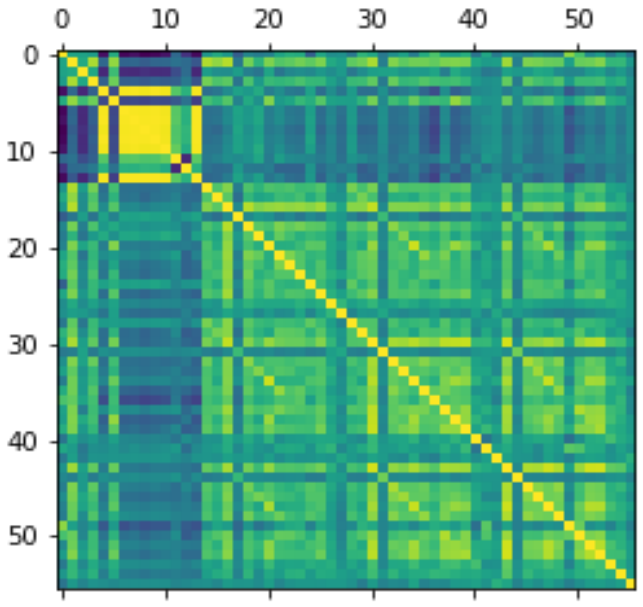
Coeficientes de Regresión y su peso



A partir de lo anterior, decidimos investigar la correlación entre las variables, para lo cual, graficamos este estadístico entre cada uno de los 56 atributos.

Figura 7:

Correlación de variables



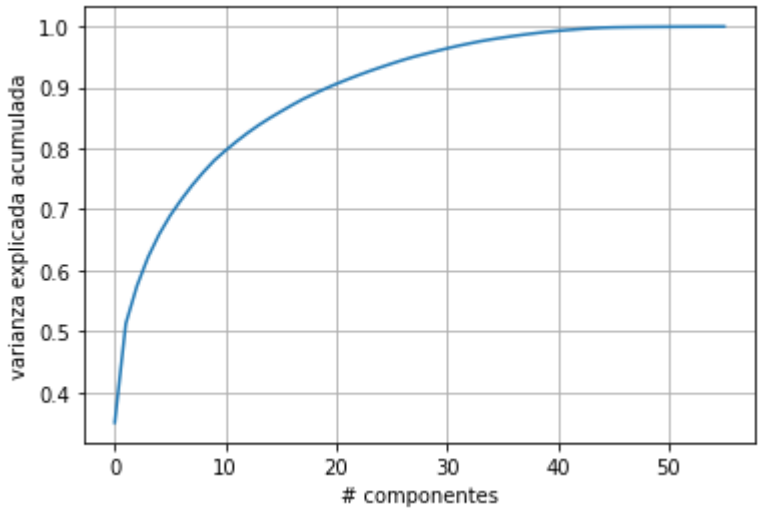
- **Paso 3: Aplicación de PCA**

A partir de lo anterior, decidimos realizar un Análisis de Componentes Principales (PCA por sus siglas en inglés),pero antes optamos por normalizar nuestros atributos.

Calculamos los atributos para quedarnos con al menos el 80% de la varianza del modelo y a la vez con el menor error cuadrático medio. Es decir, usamos los principales componentes como hiperparámetro.

Figura 8

Varianza acumulada explicada por el número de componentes



Finalmente, nos quedamos con 17 componentes principales que explican el 86% de la varianza. Esto debido al análisis que se desprende de la siguiente tabla.

Tabla 2:

Varianza Explicada	Componentes	Error Cuadrático Medio en raiz	R2	Error cuadrático medio Validación Cruzada
80%	12	45,55	0,66	44,980 +- 5,764
81%	12	45,55	0,66	44,980 +- 5,764
82%	13	45,25	0,67	45,251 +-5,447
83%	14	45,22	0,67	45,930 +- 5,343
84%	15	44,96	0,67	45,846 +- 5,096
85%	16	45,2	0,67	46,236 +- 5,141
86%	16	45,2	0,67	46,236 +- 5,141
87%	17	45,07	0,67	46,474 +- 4,893
88%	18	45,11	0,67	47,149 +- 4,816
89%	20	46,15	0,65	48,502 +- 5,080
90%	21	46,71	0,65	48,786 +- 5,093
91%	22	46,35	0,65	48,797 +- 5,635
92%	23	46	0,66	48,191 +- 5,799
93%	25	45,18	0,67	49,922 +- 5,769
94%	27	45,09	0,67	51,004 +- 5,635
95%	28	45,04	0,67	51,380 +- 5,861
96%	31	43,67	0,69	52,186 +- 7,046
97%	33	42,6	0,71	53,434 +- 7,072
98%	36	41,4	0,72	54,439 +- 8,935
99%	40	40,2	0,74	55,771 +- 9,788

Paso 4: Validación Cruzada

Realizamos Validación Cruzada, que la hicimos con el error cuadrático medio en raíz negativo, con el objetivo de no perder la generalización del modelo.

El resultados que nos entregó fue el siguiente:

Tabla 3:

Resultados Scores y Accuracy de la Validación Cruzada con 16, 17 y 18 componentes.

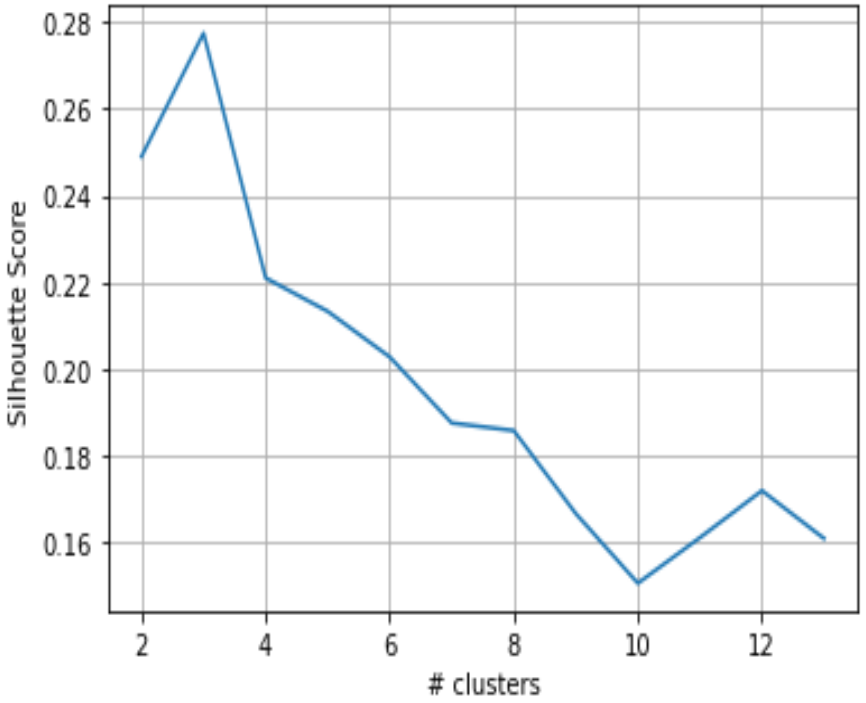
```
Componentes: 16
Error cuadrático medio en raiz: 45.20
Estadístico R_2: 0.67
Scores:  [42.56258181 53.94305098 40.70452989 47.73505224]
Raíz del Error Cuadratico Medio : 46.236303727281125 +- 5.1414238011090765
17
Componentes: 17
Error cuadrático medio en raiz: 45.07
Estadístico R_2: 0.67
Scores:  [42.56625819 53.94046253 41.65802044 47.73330389]
Raíz del Error Cuadratico Medio : 46.47451126332085 +- 4.893822971777579
18
Componentes: 18
Error cuadrático medio en raiz: 45.11
Estadístico R_2: 0.67
Scores:  [43.84594493 54.33700096 41.8412929  48.57440377]
Raíz del Error Cuadratico Medio : 47.14966064034577 +- 4.8161462661165455
```

Paso 5: Cluster

Decidimos revisar si podíamos agrupar, para poder identificar grupos de rendimientos y para ello usamos K-means. El resultado de ésto fue que no logramos identificar clusters dado el bajo Score de Silhouette y que decrece con el número de clusters como se aprecia en la siguiente figura.

Figura 9:

Score de Silhouette versus el número de clusters

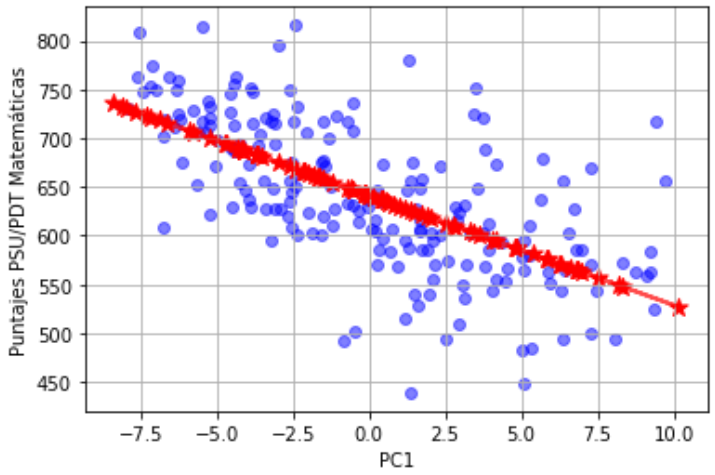


Téngase presente, que no descartamos una hipótesis de los grupos de estudios porque sabemos que este es un proceso iterativo.

Paso 6: Nuevo PCA para graficar resultados

Volvimos a reducir los datos a través de PCA a un componente y obtuvimos el siguiente gráfico:

Figura 10:

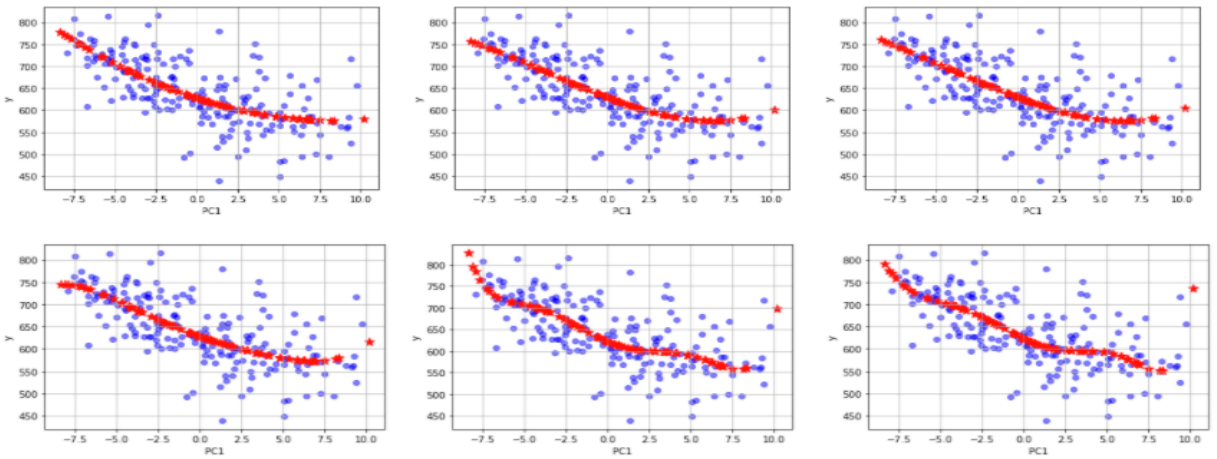



```
Componentes: 17
Error cuadrático medio en raiz: 45.07
Estadístico R_2: 0.67
```

Luego de reducir, nos preguntamos ¿qué pasa si la relación no era lineal? Ante lo cual, decidimos intentar con más grados, pero que finalmente no fue la mejor solución, ya que el error cuadrático medio de los modelos aumentaba a la vez que iba disminuyendo la significancia, incluso volviéndose negativa.

Figura 11:

Representación gráfica de regresiones con polinomios de grado 2 a 7 proyectados sobre la Componente Principal de mayor varianza.

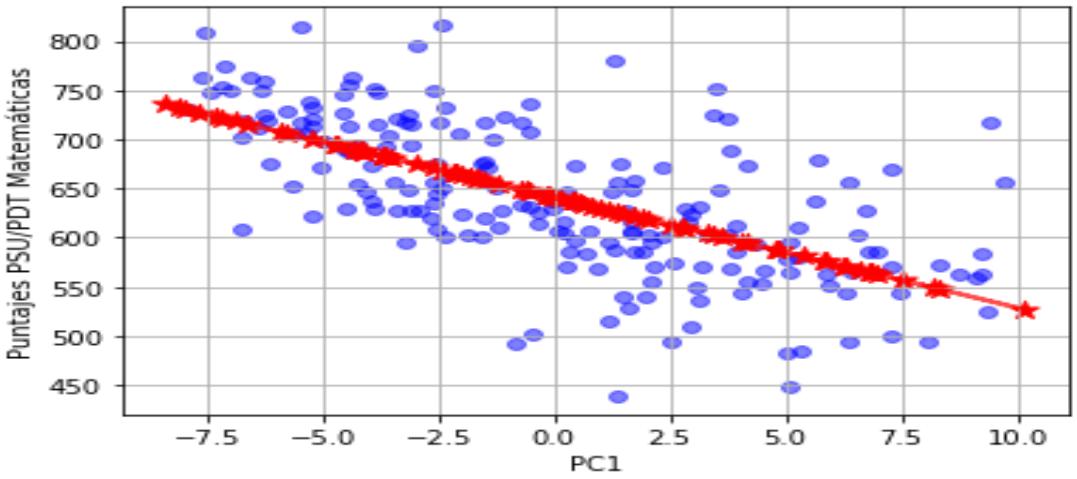


Conclusiones

1. Se establece un modelo que predice con solo las notas de enseñanza media el rendimiento de un estudiante del colegio estudiado con un error razonable.Las notas son causales de una parte del rendimiento en las pruebas de acceso a la universidad, pero no explican todo

Figura 12:

Representación gráfica de nuestro modelo con 17 componentes ortonormalizado al componente de mayor varianza para visualización.



2. Se puede lograr predecir, con un margen de error razonable, el rendimiento de un estudiante con su puntaje en pruebas de selección universitaria: notas de matemáticas y promedios generales.

```
Componentes: 17
Error cuadrático medio en raiz: 45.07
Estadístico R_2: 0.67
```

3. La relación es lineal, por tanto, al incluir el resultado de la prueba y las notas en la ponderación de la entrada a la universidad, se está evaluando, en parte, dos veces el mismo indicador
4. En el camino, pudimos sacar conclusiones que son útiles para el colegio: necesidad de hacer un análisis más

complejo de los datos y por ende, incorporar otros aspectos.

5. Aún queda bastante por profundizar en conocimientos, para hacer predicciones más certeras y por ende, hay que seguir iterando e incorporando conocimiento al modelo, con el fin que pueda servir para identificar claramente la causalidad de los puntajes y permite enfocar la labor educativa con el fin de ayudar a los estudiantes a cumplir sus objetivos.

Bibliografía

[1] Spaepen, E., Coppola, M., Spelke, E., Carey, S., & Goldin-Meadow, S. (22 de Febrero de 2011). Number without a language model. *PNAS, 108*(8), 3163-3168.