En la Primera Parte de este trabajo vamos a realizar una preprocesamiento de la base Aprender 2018 para lengua y matemática en 6° grado de la primaria. Luego en la Segunda Parte vamos a predecir problemas en el aprendizaje de matemática en utilizando los métodos Ridge y LASSO.

PRIMERA PARTE

La prueba Aprender consiste en un relevamiento del desempeño del sistema educativo del país a través de diversas pruebas estandarizadas. Esta se realiza a alumnos de primario y secundario de todo el país que busca identificar falencias en el aprendizaje de las materias centrales, así como también identificar problemas sociales en las escuelas y en los hogares de los niños.

En la página https://www.argentina.gob.ar/educacion/aprender se puede encontrar la descripción de la prueba, así también como los microdatos de la misma.

La base abarca más de 28 mil escuelas de todo el país con una participación de más de 950 mil estudiantes, llegando a casi la totalidad de escuelas del país.

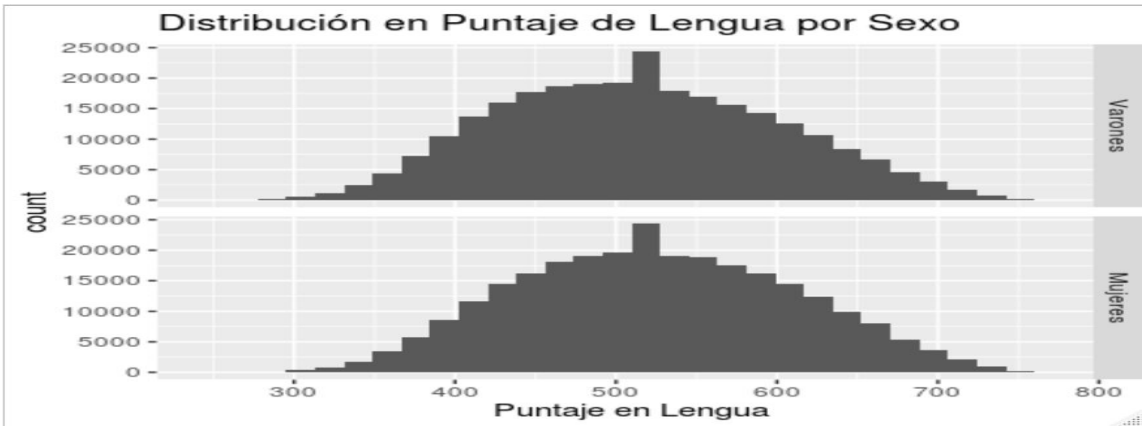
Respecto del tratamiento que vamos a hacer con los missings values:

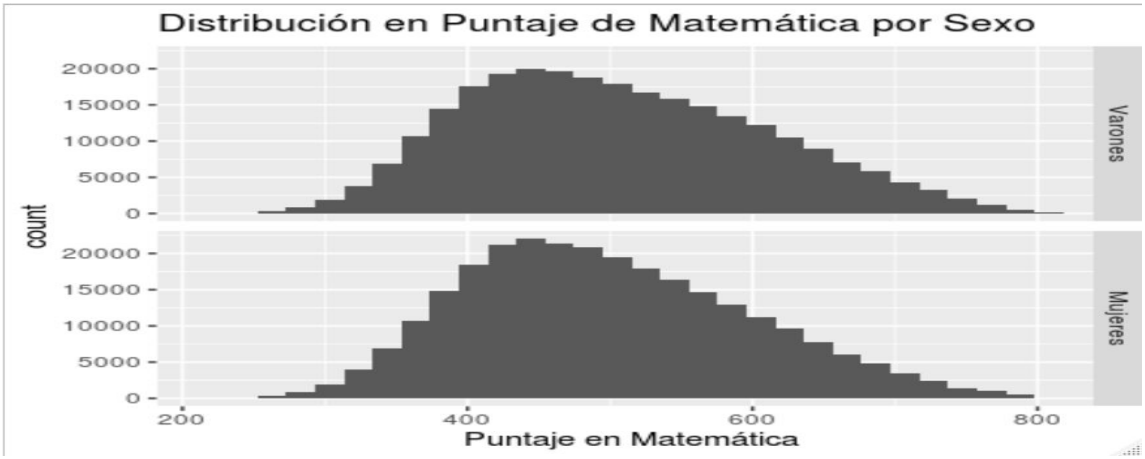
1- Vemos en la base que cuando la respuesta es **“En Blanco”**, **“Multimarca”** o **“No corresponde”**, se asigna valores -6, -9 y -1 respectivamente. Estos también los vamos a considerar missings .

2- Luego vamos a borrar todas las filas donde hay missings de **“Puntaje en matemática”** ya que en la PARTE 2 va a ser nuestra variable target.

3- Posteriormente borramos todas las columnas que tienen más de 25% de missings, que son en total 26.

3) Cómo se trata de una variable cualitativa, habíamos reemplazado los “-9” y “-6” por “-1”. Entonces los eliminamos.



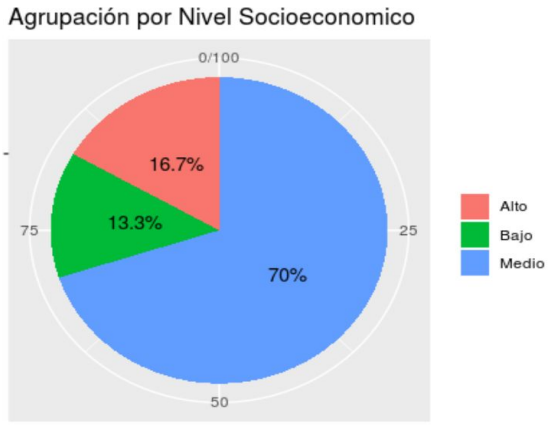


Nótese que en el histograma de **Puntaje en Matemática** no hay una concentración en la mediana como sí lo hay en **Puntaje en Lengua**. Esto es así por que anteriormente borramos los missings de **Puntaje en Matemática**. Con lo cual no hubo que reemplazar los missing por la mediana como sí sucedió con **Puntaje en Lengua**.
En los gráficos se puede ver que el puntaje en Matemática y en Lengua tienen una distribución parecida si bien Puntaje en Matemática tiene la curva de asimetría positiva. Esto quiere decir que el rendimiento de los alumnos suele ser mejor en Lengua que en Matemática.

Podemos observar que la mediana y el valor acumulado a la derecha de ella para los puntajes de Matemática es más elevado para varones que mujeres. En el caso de las evaluaciones de lengua ambas distribuciones parecieran estar más parejas.

Asimismo la correlación entre Puntaje en Lengua y Puntaje en Matemática es de 0.62. Con esto podemos afirmar que el desempeño en una asignatura es un buen predictor del desempeño en la otra.

En el siguiente gráfico mostramos la proporción de alumnos en cada nivel del índice socioeconómico del alumno



Por último elaboramos una tabla que reporta por provincia la proporción de alumnos que cae en cada nivel del índice socioeconómico.

	NIVEL SOCIOECONOMICO		
PROVINCIA	1	2	3
CABA	2,4	59,7	37,8
BS AS	8,9	73,6	17,4
CATAMARCA	14,6	71,8	13,5
CORDOBA	12,3	69,1	18,5
CORRIENTES	23,4	65,5	11,1
CHACO	25,8	62,7	11,5
CHUBUT	9,3	77,2	13,5
ENTRE RIOS	14,1	70,6	15,3
FORMOSA	27,5	64	8,5
JUJUY	16,8	72,2	11
LA PAMPA	7,2	75,7	17,1
LA RIOJA	12,6	71	16,4
MENDOZA	16,2	68	15,8
MISIONES	28,5	60,6	10,9
NEUQUEN	10,6	68,9	20,5
RIO NEGRO	11,3	72,6	16,1
SALTA	18,5	69,7	11,8
SAN JUAN	15,6	71,8	12,6
SAN LUIS	10	74,3	15,8
SANTA CRUZ	6,5	80	13,5
SANTA FE	12,7	68,8	18,4
STGO DEL ESTERO	32	59,6	8,4
TUCUMAN	17,4	69,2	13,4
T. DEL FUEGO	4,3	80,5	15,3

Aquí observamos que en CABA, Sta Cruz, Tierra del Fuego y Bs As hay un porcentaje inferior al 9 acumulado para el sector socioeconómico bajo. En provincias como Santiago Del Estero, Chaco, Misiones y Corrientes encontramos una distribución significativamente más elevada para el mismo sector socioeconómico con valores desde 23 hasta 32

En todos los casos el sector medio acumula un valor más elevado que la suma de los otros grupos sociales. Encontramos la acumulacion de CABA para el grupo socioeconómico elevado particularmente alta con respecto de las demás provincias.

SEGUNDA PARTE

Para realizar la parte de modelización optamos por convertir las variables categóricas a dummies dado que vamos a trabajar con regresiones y creemos que es un sistema más

acorde. Sin embargo eso multiplica la cantidad de vectores y hace más difícil el procesamiento. Por esta razón tomamos una muestra del 20% de los datos a los fines que sea procesable en cualquier computadora.

Después de armar los modelos nos queda que:

Para Ridge seleccionó lambda= 6,35
Para Lasso seleccionó lambda= 0,04

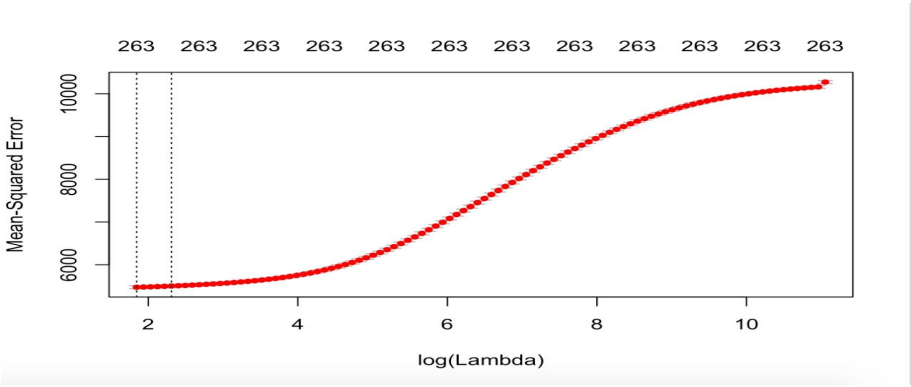
Con Lasso fueron descartadas 88 variables cualitativas y 2 numericas. Las numéricas descartadas fueron “¿Cuál es el máximo nivel educativo de tu papá?” y “Factor de expansión”.

La teoría afirma que a mayor valor de lambda, mayor es la tendencia a 0 que aparece como valor para los predictores. Por esto, el modelo presenta un mayor sesgo a la vez que una menor varianza. La disminución en la varianza, sin embargo, nos resulta particularmente útil para predecir.

El lambda óptimo para cada modelo, lo elegimos a partir de 10-fold Cross Validation. Según el caso de lo que analizamos encontraremos que no siempre un modelo es superior al otro.

Analizamos el ECM para cada modelo, en este caso, y encontramos que Lasso proporciona un valor de 73.72447, mientras que Ridge 73.9873. El ECM de LASSI es ligeramente menor, casi idéntico al de Ridge. Ante este punto de indiferencia, elegimos favorecer a LASSO ya que contiene menor cantidad de variables y por ende el riesgo de overfitting que acarrea es menor.

En lo que respecta a l2norm, hallamos esto como un criterio útil solamente para comparar entre dos modelos del mismo tipo, y no entre modelos distintos, pero solo para aquellos con distintos parámetros de ajustes. De todos modos, el valor de l2 para ridge es 260.1, y el de lasso es 211.9.



Sirviendonos del scatter plot podemos observar que hay un elevado grado de precisión, en función de que los puntos asemejan bastante una proyección lineal.