Análisis exploratorio de datos de las pruebas Aprender – 2018

Ciencia de Datos – ClusterAi

Trabajo Práctico Final – Grupo 7

Integrantes del grupo:

Mauro Corigliano

Leg. 150.376-5

Santiago Manuel Sosa

Leg. 153.070-7

Matías Grosso Basto

Leg. 153.241-8

Objetivos

El objetivo es evaluar las incidencias de distintas variables culturales, económicas y socioambientales en el desempeño en la asignatura matemática de alumnos de sexto grado de los colegios de Argentina.

Introducción

A partir de la encuesta realizada en el 2018 por el gobierno de la nación denominada “Aprender”, que consistía en evaluar el nivel académico en Matemática y Lengua, y, en paralelo, una serie de preguntas que daban dimensión del entorno en el cual desarrollaban sus estudios, se planteó mediante este análisis la búsqueda de una posible relación entre variables propias del entorno del encuestado, que influyen en el respectivo rendimiento en matemáticas. Según datos oficiales, participaron de esta encuesta el 94% de las escuelas primarias del país en 19.600 establecimientos educativos.[1][1]

Descripción del Dataset

El dataset elegido posee aproximadamente 580.000 filas y 124 columnas. Las filas representan las distintas muestras (samples:alumnos) que realizaron dichas encuestas, y las columnas muestran los resultados obtenidos en los puntos a evaluar para cada una de las etiquetas (labels). De las 124, columnas para este estudio, se consideraron como las punto de interés las siguientes:

Nacionalidad	Provincia	Edad	Sexo
Nota Matemática	Nota Lengua	Ámbito (Púb o priv)	Sector (Rural o urb)
Asistirá al Secundario	Tiene celular	Repitió	Educación de la madre
Educación del padre	Cantidad de libros que hay en la casa	Tiempo de viaje	Viaje Secundario

1

https://www.argentina.gob.ar/noticias/resultados-aprender-2018-accede-al-reporte-nacional-y-los-24-provinciales

Análisis exploratorio de datos

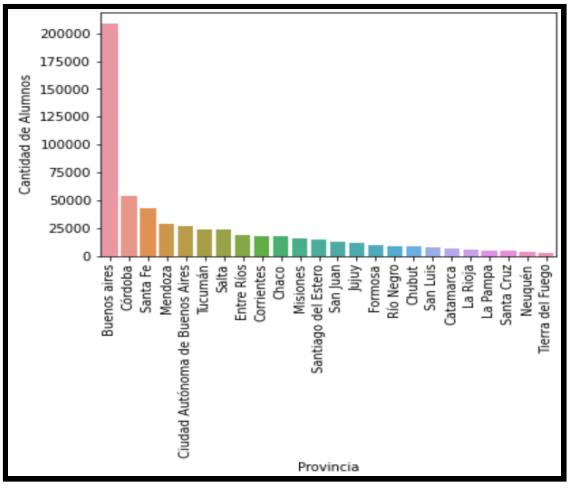
Como introducción al análisis exploratorio de datos, se realizó un reemplazo de los nombres de las columnas que interesan analizar con el fin de que sean más descriptivas. Paso siguiente, se generó un diccionario para cada una de las variables numéricas reemplazandolas por categorías para su mejor comprensión y visualización en los gráficos y tablas.

Uno de los primeros puntos para el análisis, consiste en determinar cuál es el grado de correlación entre la nota de matemática y lengua

	nota_mate	nota_lengua
nota_mate	1.00000	0.63181
nota_lengua	0.63181	1.00000

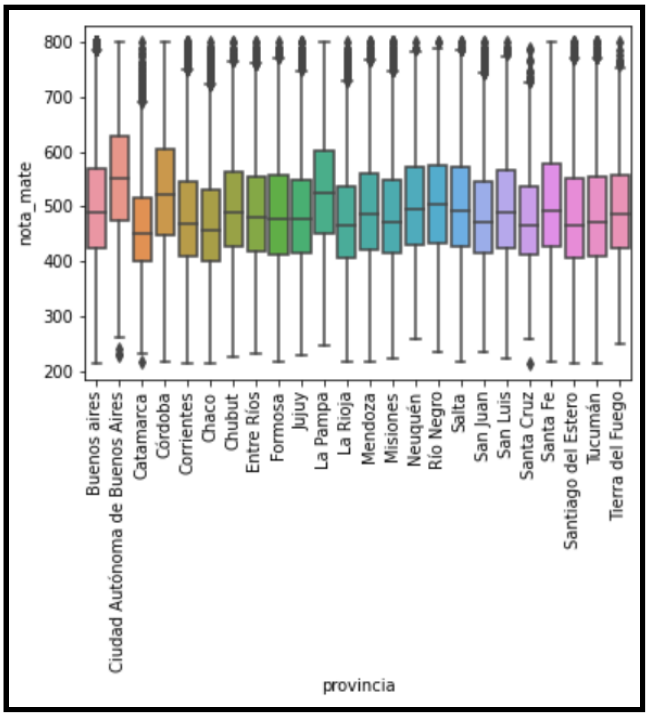
Podemos comprobar que existe un 63% de correlación entre matemática y lengua.

La siguiente instancia de análisis de datos consistió en evaluar cuántos alumnos por provincia incluía nuestro dataset y comprender de esta forma cuál era la distribución de las muestras a nivel territorio.



Podemos ver que existe una marcada preponderancia de muestras de Buenos Aires por sobre el resto de las provincias.

Una vez obtenido este resultado,procedimos a evaluar el desempeño en matemática de las diferentes provincias en base a las muestras obtenidas :



De esta forma, mediante un Boxplot[2] pudimos concluir que la Ciudad de Buenos Aires , Córdoba, La Pampa y Santa Fe poseen una mediana mayor al resto en su desempeño en la asignatura. Es importante destacar que en el gráfico están representadas la totalidad de las muestras, y que la línea interna de la caja representa la mediana. Los límites de la misma representan el primer (25%) y el tercer cuartil (75%) de cada uno de ellos y los puntos que están por fuera representan los Outliers (anomalías).[3]

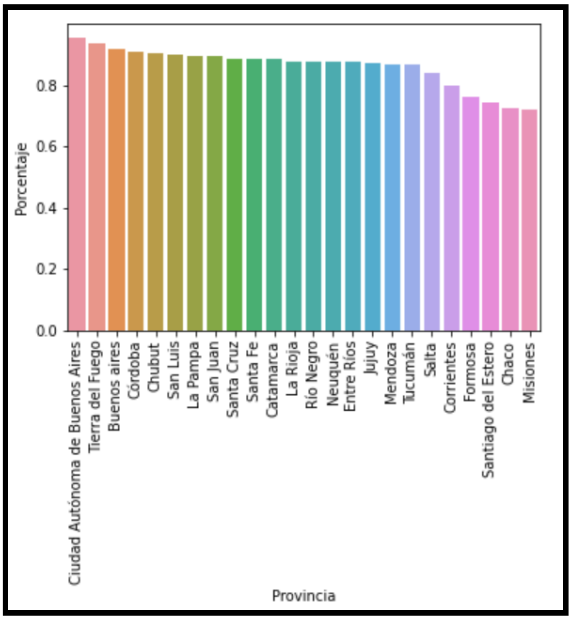
2 https://stackoverflow.com/questions/41667397/interactive-boxplot-with-pandas-and-jupyter-notebook

3 http://proceedings.mlr.press/v64/olson_tpot_2016.pdf

El análisis prosiguió con la búsqueda de variables incidentes en el desempeño de los encuestados,pero se descartaron muchas de ellas debido a que presentaban un sesgo muy importante o datos poco definido. Entre estas variables analizadas, podemos destacar el acceso a telefonía móvil y el tiempo de viaje a las escuelas secundarias,por ejemplo.

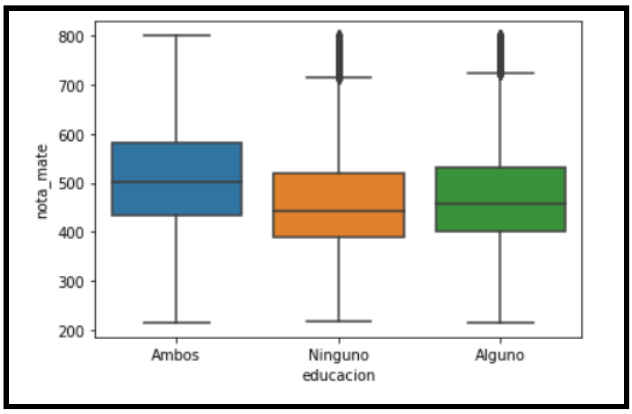
Como cierre de este análisis exploratorio de datos ,evaluamos la formación educativa de los padres contemplando diferentes grados de formación y su distribución en nuestro país.

Comenzamos por la formación primaria, es decir, evaluamos qué porcentaje de padres completaron al menos el primario. En el gráfico vemos que la variable supera el 80% en casi todas las provincias, y en aquellas donde no lo hace, se encuentra cercano al 70%.



Es importante destacar que el 88% de los alumnos tiene ambos padres con al menos la educación primaria completa

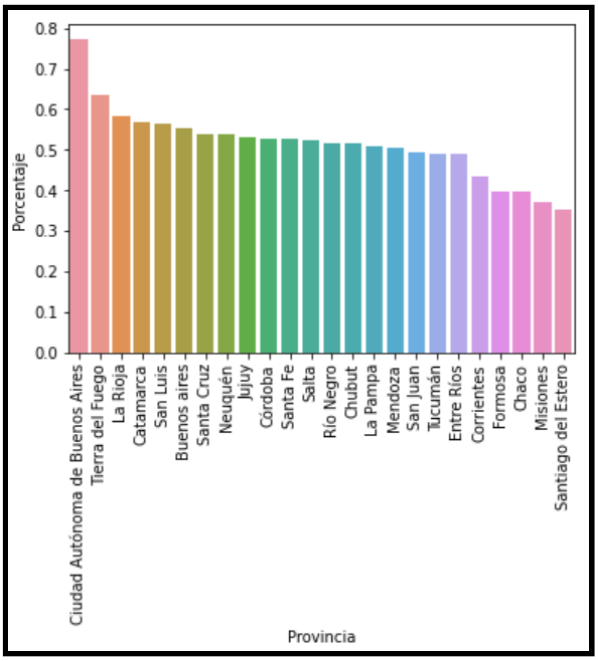
Y finalmente evaluamos la variación del desempeño en matemáticas en base a la formación primaria de los padres de los encuestados :



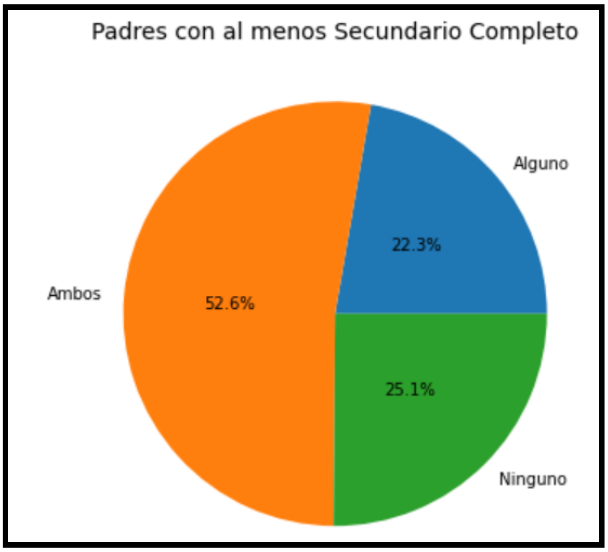
De esta forma podemos ver que hay una marcada diferencia de desempeño en favor de los encuestados que señalaron tener ambos padres con estudios primarios frente a los que no tenían educación o tan solo alguno de sus padres en tal situación.La diferencia en la media del desempeño llega a ser de 50 puntos superiores en el primer caso.

Continuando el análisis del índice de finalización de estudios de los padres, en este caso, del secundario completo, a diferencia de los resultados previamente observados, existe una marcada disminución en los mismos. La que mejor porcentaje presenta es la Ciudad Autónoma de Buenos Aires que promediando un 80%. El resto de las provincias se encuentran por debajo del 60%. Es decir, en CABA un 80% de

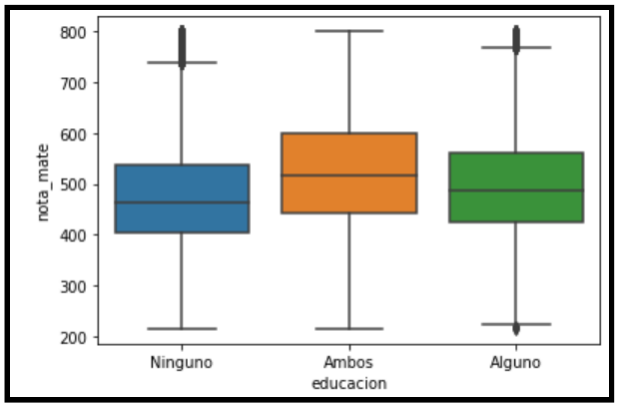
los padres terminaron el secundario y en Santiago del Estero un 40% aproximadamente.



Tal como se muestra a continuación, el 52,6% de los padres tienen ambos padres con el secundario completo, y el 25,1% secundario incompleto.

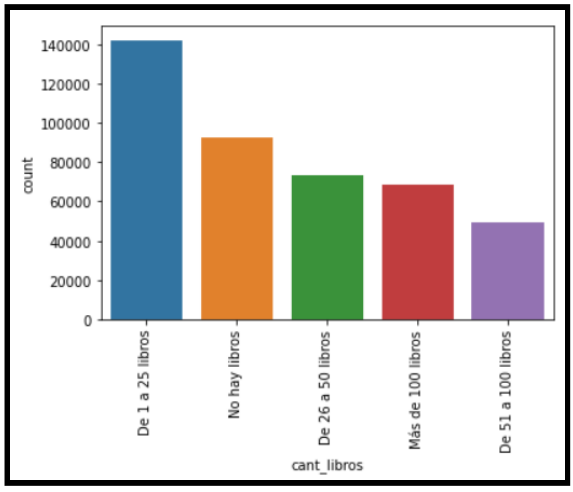


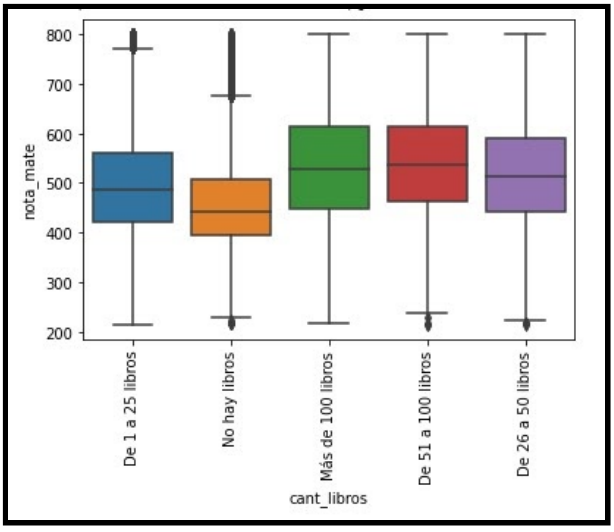
Como se realizó previamente, se representa mediante un gráfico “Boxplot” la variabilidad que presentan en las notas teniendo en cuenta los estudios secundarios de los padres:



Podemos concluir que en ambos casos, primaria y secundaria completa de los padres, tiene una incidencia directa sobre la nota de matemática. La diferencia en la media llega a ser de 50 puntos aproximadamente en los diferentes escenarios.

Finalmente,analizamos la incidencia de la disponibilidad y acceso a libros por parte de los encuestados en su rendimiento en matemáticas. Para esto,se analizó la cantidad de libros con los que cuentan los encuestados en sus hogares :





De esta forma, vemos que la media del desempeño en matemática es considerablemente menor para los alumnos que no tienen libros en sus casas y un poco mayor en los alumnos que tienen hasta 25 libros. Es decir, que la disponibilidad de libros en los hogares, incide en el desempeño de los estudiantes.

Materiales y métodos (algoritmos utilizados),

Se realizó un análisis de componentes principales (PCA),el cual tenía por objeto reducir las dimensiones de análisis y la generación de nuevas features ,las cuales poseerán la mayor cantidad de información concentrada.Por otro lado,habíamos evidenciado que existía mucho ruido en los datos disponibles.Sin embargo, a partir de un explain.ratio ,comprobamos que las 10 componentes generadas por el PCA no explicaban una varianza significativa,es decir, entregaban un explained variance bajo. Teniendo en cuenta esto, optamos por utilizar un encoder para convertir variables categóricas a binarias.Para ello,utilizamos onehot encoder y generamos las dummies para volver a utilizar el PCA obteniendo el mismo resultado,con una explained variance muy baja.

Es decir que nuestro objetivo era determinar a qué “cluster” de rendimiento en matemáticas pertenecía un alumno según su entorno social,económico y cultural.

Resultados

Como hemos podido sustentar en el presente informe,los resultados indican que existe una tendencia a obtener mejores notas en matemática cuando se forma parte de un hogar con padres con estudios iniciales concluidos y además,el apoyo con libros que puedan encontrarse en los ámbitos donde crecen los encuestados,incide en su desempeño de forma notable.

Discusión y Conclusiones

Como conclusión encontramos que este análisis de datos podría continuar con un modelo de predicción de desempeño en matemáticas en Machine Learning a partir de ciertos inputs,o un modelo no supervisado de clustering que nos permita separar variables automáticamente y agrupar según similaridades.Ambos casos fueron evaluados por el equipo,pero el score obtenido en el entrenamiento no fue satisfactorio,por lo cual no se incluyeron dichas conclusiones. Tal como mencionamos en el apartado de materiales y métodos,el dataset no posee información suficiente (features relevantes) para poder predecir el desempeño en matemática de un alumno (variable categórica) mediante variables de su entorno social (por ejemplo,educación de los padres,tiempo de viaje a la escuela,etc).

Referencias

1. *Randal S. Olson olsonran@upenn.edu and Jason H. Moore jhmoore@upenn.edu Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA*
2. *https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html*
3. *https://stackoverflow.com/questions/41667397/interactive-boxplot-with-pandas-and-jupyter-notebook*