María García Alba
mpgaral98@gmail.com

Memoria

1. Contexto

Para finalizar el primer bloque del Bootcamp de Data Science en The Bridge se requiere la realización de un EDA (Análisis Exploratorio de Datos). En mi caso, me apasiona el mundo de la educación, pues es el tema que he elegido. Este trabajo está basado en un DataSet obtenido de la página web Kaggle, en el que se muestran los resultados académicos en matemáticas, lectura y escritura de 1000 estudiantes estadounidenses. También se muestran más variables como el género de los alumnos, el nivel de educación parental, si tienen programa de comida o no y si han realizado curso de preparación previo a los exámenes.

Con esta información pretendo analizar los resultados de los alumnos y comprobar las hipótesis planteadas sobre si interfieren algunos de los factores marcados anteriormente para mejorar los resultados en el futuro.

2. Hipótesis

Las hipótesis planteadas para este análisis surgen de la necesidad de comprobar si existen mecanismos que ayudan al rendimiento escolar. Por ello, hemos planteado tres hipótesis:

1. ¿Existen diferencias entre el género y los resultados académicos en diferentes áreas de estudio (matemáticas, lectura, escritura)?
2. ¿Es útil el curso de preparación para aprobar los exámenes?
3. ¿En qué alumnado debemos incentivar el curso de preparación?

3. Análisis

1. Análisis univariante

En el análisis univariante podemos observar que hay cuatro variables categóricas y cuatro numéricas discretas.

Por un lado, analizando las variables categóricas con una medida de tendencia central como la moda podemos observar que el género más repetido es el femenino, pero realmente viendo las frecuencias absolutas no hay tanta diferencia finalmente. El nivel de educación parental más repetido es algo de universidad aunque muy seguido de técnico superior. En la variable comedor si observamos más diferencia siendo la moda la opción estándar. Por último, para la variable

curso de preparación observamos que la moda es ninguno, lo cuál es interesante para tener en cuenta en el análisis bivariante.

Por otro lado, analizando las variables numéricas, aplicamos la media como medida de tendencia central y comparándola con la mediana sin obtener grandes diferencias, por lo que la media puede ser representativa. En las tres variables numéricas principales (resultados en matemáticas, lectura y escritura), vemos que la tendencia es un aprobado simple con un 66-68.

Mediante un diagrama de cajas observamos que la variable numérica que más outliers tiene es resultados en matemáticas, lo que nos lanza algo interesante para el análisis bivariante. Los resultados en lectura y escritura no tienen demasiados outliers, por lo que podemos razonar que las matemáticas es una asignatura más difícil en general que las otras dos asignaturas medidas. Al aplicar un histograma para visualizar los valores podemos observar que genera una campana gaussiana y por lo tanto podemos seguir trabajando con los outliers sin problema, ya que no genera gran diferencia con otros.

2. Análisis bivariante

En el análisis bivariante vamos a analizar por un lado, la relación entre el género de los estudiantes y las diferentes áreas de estudio para comprobar la primera hipótesis. Por otro, para la segunda hipótesis, analizaremos la relación con la realización del curso de preparación y la media de resultados para comprobar si el curso es interesante y proporciona resultados o no. Por último y coincidiendo con la tercera hipótesis, analizaremos la relación entre el género y el curso de preparación para comprobar el target del curso.

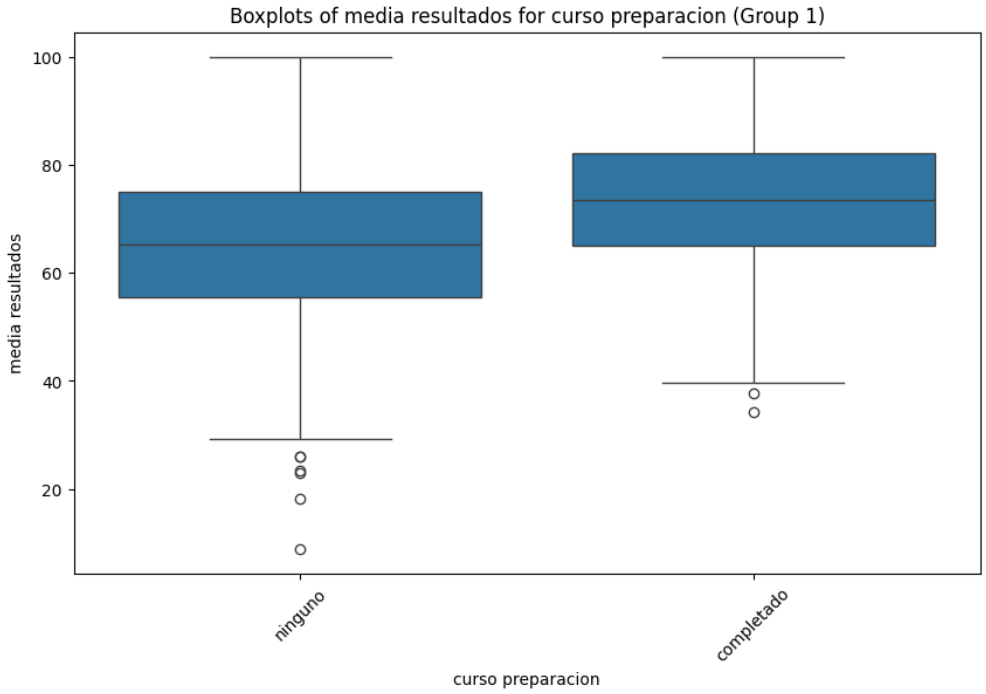
En primer lugar, se analiza la variable género con las diferentes variables numéricas que contienen los resultados académicos. Primero mediante un diagrama de barras que no nos proporciona información muy relevante, porque sí, podemos observar que los chicos obtienen mejores notas en matemáticas que las chicas mientras que las chicas obtienen mejores resultados en escritura y lectura que los chicos. Sin embargo, mediante un diagrama de cajas podemos observar que en matemáticas si es relevante el diagrama de barras, pero en lectura y escritura no, pues en los resultados de escritura existen múltiples outliers que nos indican que un número considerable de chicas no es tan buena en lectura mientras que los chicos tienen menos outliers y los resultados están más concentrados dentro del aprobado. El mismo resultado obtenemos con las calificaciones de escritura, incluso con más outliers en las chicas.

En segundo lugar, se analiza la variable curso de preparación con la media de los resultados, y nos ocurre lo mismo que en el caso anterior: con el diagrama de barras no obtenemos diferencias significativas pero con el diagrama de cajas si podemos observar anomalías. Hay más alumnos que no realizaron curso previo a los exámenes y sacaron malas notas mientras que la variable de los alumnos que sí realizaron el curso tiene menos outliers y más resultados dentro o cerca del aprobado.

En tercer y último lugar, se analizan las variables categóricas género y curso de preparación sin obtener resultados relevantes, pues en ambos géneros la opción más elegida es que no realizaron el curso de preparación. Solo podemos ver que el porcentaje de chicas que hicieron el curso es superior al de los chicos.

3. Visualizaciones relevantes

Figura relación entre si realizaron o no el curso de preparación antes de los exámenes y los resultados académicos:



Mediante esta visualización podemos observar que los alumnos que realizaron el curso de preparación previo al examen obtuvieron mejores resultados que aquellos alumnos que no realizaron ningún curso.

María García Alba
mpgaral98@gmail.com

4. Conclusiones

Mediante el análisis hemos podido comprobar nuestras hipótesis. Por norma general, el género femenino tiende a tener mejores resultados que el masculino y que una de las causas o ayudas puede ser la realización de un curso de preparación previo al examen. Esta hipótesis no es definitiva, pues hemos observado que aunque los alumnos aprueban sus asignaturas con más frecuencia si realizan el curso de preparación, el número de alumnos que realizan este curso es mucho más bajo que el de los que no lo realizan. De cualquier modo, de forma lógica el repaso previo a un examen siempre es interesante, por lo que quizá se debería incentivar más el curso de preparación a todos los estudiantes y hacerlo más atractivo para ellos.

5. Recomendaciones

Queremos dedicar esta parte para hablar sobre estudios futuros que pueden ser interesantes con este DataSet. Podríamos plantear otras dos hipótesis más, como por ejemplo: ¿influye el servcio de comedor en los resultados académicos? Por un lado, para comprobar si este servicio afecta o no a los resultados de los alumnos y por lo tanto si es rentable para los padres pagarlo o no.

La segunda hipótesis podría ser: ¿afecta el nivel de estudio de los padres a los resultados académicos de los hijos? Con esta hipótesis podríamos comprobar a nivel demográfico si el conocimiento académico de los familiares puede influenciar o no en los hijos.