

- **Título del proyecto.** Influencia del ingreso económico en el rendimiento en enseñanza media.
- **Contexto y motivación:** discuta la motivación y razones para su elección del proyecto, incluyendo contexto y alguna investigación que haya incluido en su decisión. Comente también sobre la audiencia objetivo de su análisis.

Considerando el contexto de pandemia en el que nos encontramos, se planteó como problema las dificultades que se han presentado en el desempeño de los estudiantes en el país, y a partir de esto surgió la idea de analizar distintas variables para saber si existe algún tipo de relación. Se encontraron databases del Ministerio de Educación, y al leer de qué se trataban los datos resumidos del rendimiento académico en los últimos años se tomó la decisión de llevar esta investigación a cabo. Debido a que el tema escogido habla particularmente de ellos, la audiencia a la que se quiere llegar es mayoritariamente estudiantes de la enseñanza media, a partir de los 15 años en adelante.

- **Objetivos:** describa los objetivos científicos y de inferencia de su proyecto. ¿Qué pregunta espera resolver o responder, y para qué? ¿Cuáles son los beneficios o accionables que podría lograr mediante este proyecto? ¿Cuál es la audiencia objetivo de su análisis?

El objetivo principal es encontrar un vínculo entre el desempeño estudiantil y variables como el ingreso anual por región y las conexiones a redes de internet, para determinar los posibles efectos que estas tienen en el rendimiento de los estudiantes en la enseñanza media. La pregunta que esperamos resolver es ¿qué factores influyen en los estudiantes, en el contexto actual, en su desempeño académico? La respuesta a esta pregunta podría ayudar a entender las necesidades que hoy en día están teniendo los jóvenes estudiantes y tomar medidas según estas para poder evitar un posible incremento de desigualdad en el ámbito de la educación.

- **Datos:** describa los datos a utilizar, sus características (variables, tipo, formato, volumen), origen y forma de recolección. Los datos pueden provenir de fuentes privadas (siempre que no tengan restricciones de uso y publicación), de fuentes públicas (gobierno, plataformas de datos abiertos, Kaggle, etc.) o pueden ser recolectados mediante algún método propuesto por su equipo (Ej. web scrapping, Twitter, RRSS, etc.).

Se ocuparán datasets del Gobierno de Chile en cuanto al ingreso total por región y el desempeño estudiantil, de los sitios de datos municipales y datos del Mineduc respectivamente. La tercera base de datos es de la Subsecretaría de Comunicaciones, que muestra la cantidad de redes de internet fijas en el país. Todas se encuentran en formato .csv y .xlsx, por lo que se expresan como dataframe.

El database de Mineduc (.xlsx) tiene variables como el código de dependencia de los establecimientos, la comuna y región donde se encuentran, el promedio general anual y todo otro tipo de información tanto del establecimiento como de los estudiantes, pero no todos los años presentan las mismas características. Por lo general son códigos que representan las variables, de tipo numérico. Debido a la cantidad de estudiantes por año en el país, los archivos son extensos, de millones de líneas.

El sitio de datos municipales (.xls) tiene los ingresos anuales de cada municipio y el ingreso de los hogares categorizados por porcentajes. Los primeros datos, a pesar de tener variables que representan de dónde viene cada ingreso, serán utilizados según la suma por región como variable general, siendo esta de tipo numérico. La segunda variable a considerar también será trabajada de esta forma, dejando toda la información por región. Es una base de datos con mucha información, pero al ser trabajada de forma distinta su extensión decrecerá considerablemente. Aun así, de ser posible se intentará conseguir una base de datos distinta que pueda ser más cómoda de trabajar, con los mismos datos.

Por último, los datos de redes a internet fijas (.csv) se encuentran clasificados por año y región, además de tener una última variable que representa el total. Todas son de tipo numérico, ya que son una cantidad. El volumen no es tan extenso como el desempeño estudiantil, teniendo menos de 200 líneas.

- **Preguntas de investigación:** proponga al menos 5 preguntas específicas que buscarán responder en base a los datos presentados. Estas preguntas pueden ser relativas a relaciones entre variables, capacidad de predicción de una variable, clasificación, identificación de clusters o categorías, etc.

¿Cuáles son las comunas de mayor desempeño académico y cuáles son las diferencias con las que tienen peor rendimiento? ¿Coincide lo anterior con el ingreso que estas presentan?

¿Se puede decir que hay una relación lineal entre ambas variables?

¿Se puede afirmar que el acceso a internet y los ingresos económicos influyen en los estudiantes?

¿Existe alguna conexión además entre el acceso a internet y los ingresos económicos?

¿Se puede predecir el desempeño del próximo año a partir de lo observado? Y si es así, ¿cómo sería?

- **Diseño tentativo:** explique brevemente los métodos computacionales y estadísticos que espera usar en su análisis.

Principalmente se trabajará con el análisis de dataframes y gráficos de tipo relplot e histogramas para múltiples variables como los datos de varias regiones distintas, además de gráficos de barra cuando sea necesario analizar relaciones más simples. Se utilizarán además fórmulas de correlación como la de Pearson, para comprender el vínculo entre variables de bases de datos distintos.

#### **Consideraciones del repositorio:**

En la carpeta "Propuesta" se encuentran los archivos en bruto de redes de internet fijas e ingresos de los hogares municipales, además de capturas de pantalla de muestra de estos últimos y los archivos estudiantiles, debido a su formato y peso. Además se encuentra un archivo en formato pdf que describe las variables del database del Mineduc, para facilitar su comprensión.