Hackatón Data Science 2302. **Octubre 2023**

**GRUPO 3**

Dominguez Paula Lucía

Bonino Mauricio Edmundo

Gabirondo Lucas Martín

Lope de Barrios Nicolás

Rinaldi Juan Francisco

**Objetivo General:**

***Establecer la posible relación entre determinados factores de riesgo y la propensión a la hipertensión arterial, en personas de la República Argentina.***

**Objetivos Específicos:**

* *Explorar la base de datos publicada*
* *Seleccionar variables a analizar*
* *Caracterizar las variables seleccionadas*
* *Analizar de forma bivariada los factores de riesgo con la prevalencia de HTA*
* *Determinar y comprender la relación que se establece entre las variables analizadas y la HTA*

**¿Por qué elegimos la HTA?**

Existe amplia evidencia científica para considerar la hipertensión arterial (HTA) como el factor de riesgo de mayor peso en el desarrollo de enfermedades cardiovasculares, ya que está fuertemente asociada a la ocurrencia de diversas patologías como la enfermedad cerebrovascular, la enfermedad coronaria, la insuficiencia cardíaca, la fibrilación auricular, las enfermedades de la aorta y arterias periféricas, la demencia, la insuficiencia renal y la retinopatía.

Su prevalencia impacta negativamente en la mortalidad total, en la discapacidad y en los costos sociosanitarios. En términos de muertes atribuibles, **la HTA es responsable del 19% de todas las defunciones a nivel mundial.**

**Fuente** **de los datos**

El dataset que utilizamos para esta actividad surge de la Encuesta Nacional de Factores de Riesgo (en adelante, ENFR) que confecciona el Instituto Nacional de Estadística y Censos (INDEC), conjuntamente con el Ministerio de Salud de la Nación.

La ENFR releva en las 24 jurisdicciones del país en aglomerados urbanos de más de 5000 habitantes, los principales factores de riesgo de las Enfermedades no Transmisibles (ENT), estas a su vez se clasifican en Enfermedades Crónicas no Transmisibles (ECNT) y las consideradas Lesiones de Causas Externas.

La última ENFR se llevó a cabo en 2018 y se publicó en 2019 (la cuarta), y, en la actualidad, se están definiendo los lineamientos para la realización de la quinta. Su objetivo principal es proporcionar información confiable sobre las Enfermedades No Transmisibles (en adelante, ENT), como obesidad, diabetes e hipertensión entre otras.

La Encuesta se divide en dos grandes bloques: **Bloque Hogar y Bloque Individual.** Con respecto al primer Bloque, se relevaron características habitacionales y familiares de los hogares de la muestra a través de un cuestionario de tipo presencial.

En tanto, el Bloque Individual, obtiene características de los individuos de la muestra, las que fueron subdivididas en dos grandes grupos: situación laboral (ocupado, desocupado en búsqueda activa o no, y subocupado) y presencia o no de factores de riesgo de las ENT (realización de actividad física frecuente, consumo de frutas y verduras, Índice de Masa Corporal, etcétera).

El dataset cuenta con 29.224 observaciones y 287 características, de forma que posee un tamaño de 8.387.338 datos en total, con un porcentaje total de nulos de un 42% aproximadamente.

**Su objetivo es proporcionar información válida, confiable y oportuna sobre los factores de riesgo y prevalencias de las principales enfermedades no transmisibles en la población de 18 años y más de la República Argentina.**

Las principales enfermedades que se analizan en la ENFR son:

* enfermedades cardio y cerebrovasculares,
* el cáncer,
* la diabetes y
* las enfermedades respiratorias crónicas.

Las mismas equivalen al **71%** de las muertes que se producen en el mundo.

Las ENT comparten los mismos factores de riesgo, los cuatro más importantes son

* el consumo de tabaco y la exposición al humo de tabaco ajeno,
* la alimentación inadecuada,
* la inactividad física y
* el uso nocivo de alcohol.

Todos ellos responden a comportamientos personales y sociales principalmente adquiridos en la infancia y en la adolescencia e influenciados por los entornos.

**Metodología**

1. Preparación de Datos y Limpieza de datos

Esta etapa consistió en preparar los datos para su posterior análisis.

Como criterio de selección, se estableció que haya por lo menos una variable cada bloque vinculado a los factores de riesgo, y para su selección se realizó una investigación preliminar acerca de la temática abordada.

Se omitieron las categorías que tenían mayoría de datos incompletos (>80%), como colesterol y glucemia elevados.

**Variables predictoras a utilizar:**

1. nivel\_instruccion (Nivel de instrucción)
2. bhch03 (Sexo biológico)
3. rango\_edad (Rango de edad)
4. bisl06 (¿Cuántas horas semanales trabaja habitualmente en todos sus empleos/ ocupaciones?)
5. nivel\_actividad\_fisica (Nivel de actividad física)
6. consumo\_tabaco\_100 (Condición de fumador)
7. imc\_categorias (Índice de masa corporal agrupado)
8. bial02 (Habitualmente, ¿le agrega sal a los alimentos una vez que están cocidos o al sentarse a la mesa…)
9. consumo\_regular\_riesgo (Consumo regular de alcohol de riesgo)
10. ta\_elevada (Presión arterial elevada)

**Variable predecida**

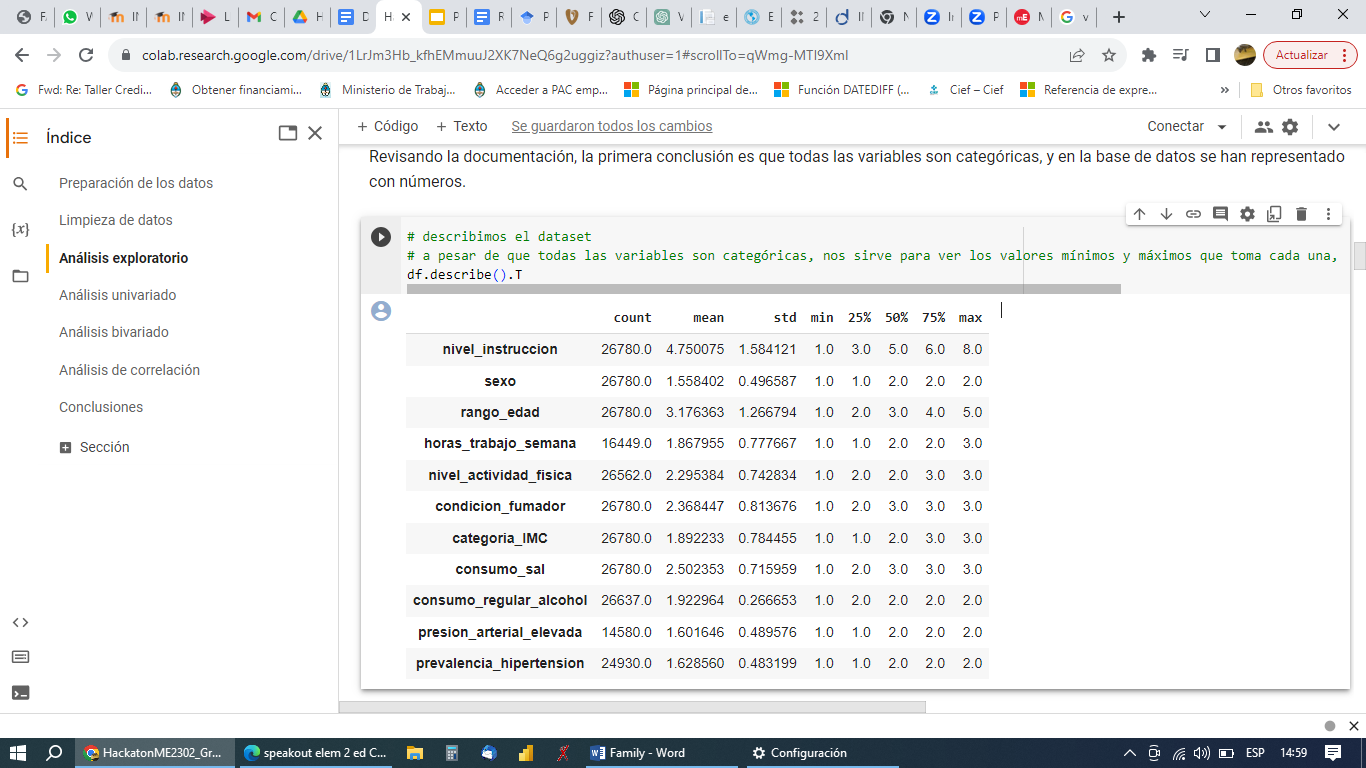
1. prevalencia\_hipertension (Prevalencia de presión arterial elevada)

Población que autorreporta hipertensión. Es decir que se había realizado mediciones de presión arterial en el lapso de 2 semanas previas o menos a ser encuestados, dando un valor mayor de 140/90.

Para completar la limpieza de los datos con los que vamos a trabajar, procedimos a reemplazar los nombres de las variables que no resultaban intuitivas; con una función recursiva eliminamos las filas con valores “No sabe/No contesta” (99).

1. Análisis exploratorio

Análisis preliminar de datos para descubrir relaciones entre medidas en los datos y para obtener una idea de las tendencias, patrones y relaciones entre las entidades presentes en el conjunto de datos, con la ayuda de estadísticas y herramientas de visualización.

Revisando la documentación, la primera conclusión es que todas las variables son categóricas, y en la base de datos se han representado con números.

Identificamos la cantidad de valores faltantes en cada columna.

Se observó que las variables que se encuentran con su totalidad de datos son: Nivel de Instruccion, Sexo, Edad, Condición de Fumandor; Categoria de Indice de Masa Corporal y Consumo de Sal. Mientras que las variables que más faltantes presentan son Presión Arterial Elevada, Horas de Trabajo por Semana, Prevalencia Hipertensión, Nivel de Actividad física y Consumo Regular de Alcohol de manera decreciente.

1. Análisis univariado

A continuación, categorizamos cada variable, convirtiendo los valores en la categoría que representan, y las incorporaremos al dataframe. Esto facilitará la lectura e interpretación de los datos, paso seguido validamos el mapeo de los datos y graficamos las variables, obteniendo los siguientes resultados:

1. Poco más de **25% de la población tiene hasta primario completo,** las muestras tomada por esta encuentra representa que se encuestador más **mujeres (55%) que varones (44%)**,
2. El rango de edad de las personas encuestadas determina que **27%** pertenecen al estrato de **35-49** y el **41%** representa a personas más de **50** años,
3. En relación a la carga horaria de trabajo un **46,6% determina que trabaja menos de 45** horas y solo un 15% de los encuestados declara trabajar más de 45 horas semanales.
4. En cuanto a la actividad física, **46%** manifiesta desarrollar **poca o nula actividad física**, un **41% fuma actualmente o es ex fumador.**
5. En cuanto al **ICM un 63%** es considerado con índice de masa corporal igual o mayor de 25, de esta población un 41% presenta un índice mayor a 30.
6. **Un gran número de personas no agrega sal** a la comida una vez cocida o servida en la mesa 63% y poco más de 7.5% manifiesta consumir alcohol de manera regular.
7. En relación a la prevalencia que elegimos trabajar, **un 21,7% manifiesta tener presión elevada,** esta cifra asciende a 34.6% cuando establecemos la prevalencia de hipertensión arterial.
8. Análisis bivariado

A continuación se procedió a confeccionar las tablas de contingencia, donde se relacionan la variable explicada del modelo **(recordemos que es prevalencia de hipertensión),** con cada una de las explicativas. Para mayor claridad interpretativa, las cifras están expresadas como porcentaje del total. **Es importante aclarar que los valores totales de las categorías no suman el 100% por la presencia de un 7% de datos faltantes en la variable prevalencia de hipertensión.**

1. Conclusiones de las matrices de contingencia

Las tablas de contingencia nos permiten conocer a priori las relaciones que existen entre las variables explicada y explicativa. A continuación, resumimos las conclusiones fundamentales que pueden extraerse de estas:

1. La Tabla 0 relaciona la prevalencia de hipertensión con el nivel de instrucción del encuestado. En la misma, el **19% de los encuestados con prevalencia de hipertensión** **tienen como máximo nivel educativo el secundario incompleto**, frente al 16% restante con mayores niveles educativos (de secundario completo a universitario completo).

Esto confirma lo que sugiere la bibliografía consultada (Viego y Temporelli, 2016) donde se concluye que a mayor nivel de instrucción, mayores cuidados y, por ende, menos posibilidades de padecer una ENT.

1. La Tabla 1 relaciona la prevalencia con el sexo biológico del encuestado. Claramente, se ve que la misma **es superior en las mujeres (20% del total de casos positivos)**, frente al 14% de los varones, lo que también se condice con la bibliografía.
2. La Tabla 2 relaciona la prevalencia de la hipertensión con el rango de edades. También es fuerte el grado de relación entre esta última y la posibilidad de que el encuestado sea hipertenso, ya que el 22% de los casos positivos se concentra en las edades que van de los 50 a los 65 y más años. **Apenas el 14% de los menores de 50 años es hipertenso.**
3. La Tabla 3 relaciona la prevalencia de la hipertensión con el número semanal de horas trabajadas. En este punto, no parece existir una relación nítida, ya que apenas el 4% de los individuos hipertensos trabajan más de 45 horas semanales frente al 7% que lo hace por menos de 35 horas. **Cabe aclarar que el dato de prevalencia de hipertensión estaba presente para el 57% de los casos analizados**, de modo que no es posible establecer una relación positiva entre horas de trabajo e hipertensión.
4. La Tabla 4 alude acerca de la importancia de la actividad física continua como factor para reducir el riesgo de hipertensión arterial, ya que apenas el 4% de las personas hipertensas hace actividad física en forma intensa, así también podemos observar que **entre quienes no presentan esta patología un 11% actividad física de manera recurrente.**
5. La Tabla 5 arroja conclusiones que van en contra de lo que especifica la bibliografía, ya que, en todas las categorías se demuestra que no hay una relación entre esta condición y la variable dependiente.
6. La Tabla 6 relaciona el Índice de Masa Corporal (IMC) y sus diferentes niveles con la prevalencia de la hipertensión. Como se ve y en consonancia con lo que surge de la literatura médica, esta enfermedad está muy presente entre personas con problemas de obesidad, ya que se concentra entre el 13% de las personas con IMC Alto y el 13% de las personas con IMC Medio, **afectando apenas al 4% de los que poseen un IMC Bajo.**
7. La Tabla 7 une los hábitos alimenticios de los encuestados con la hipertensión. Como se ve, la misma afectaría más frecuentemente (24% de los casos) a quienes nunca le echan sal a la comida una vez que la sirven en el plato, frente a un 7% de quienes lo hacen raras veces y un 4% de quienes lo hacen casi siempre. Se trata de una conclusión que contradice a la bibliografía y que puede obedecer a cuestiones ya mencionadas.
8. La Tabla 9 relaciona el consumo de alcohol frecuente (otro de los factores de riesgo) con la hipertensión. Apenas afecta al 2% de los encuestados que reportó tomar alcohol de manera frecuente, algo que también contradice la bibliografía.
9. Finalmente, la Tabla 10 relaciona los estudios objetivos de hipertensión –en concreto, mediciones físicas– con la prevalencia. Como se ve a pesar de que es una variable que presenta pocos datos, un 8% reportaba no haberse realizado medición previa y en el momento en que fue encuestado este valor le dio elevado. También resulta interesante el dato de las personas cuyos estudios médicos habían arrojado un resultado elevado, pero que con las mediciones objetivas en el encuadre de esta encuesta, resultan no serlo (10%).
10. Análisis de correlación

Aplicamos la **prueba de chi-cuadrado de independencia**, que se utiliza para comprobar la independencia de dos variables categóricas. El objetivo es analizar si los valores característicos de la primera variable están influidos por los valores característicos de la segunda variable y viceversa. La prueba se basa en la comparación de las frecuencias observadas y esperadas en una tabla de contingencia. Si los valores observados son significativamente diferentes de los valores esperados, se puede concluir que existe una relación entre las dos variables (Barreto, 2013).

La prueba de chi-cuadrado es una herramienta útil para analizar datos categóricos y puede ser utilizada en una variedad de campos, incluyendo la medicina, la biología, la psicología y las ciencias sociales.

La función **scipy.stats.chi2\_contingency** se utiliza para realizar una prueba de chi-cuadrado de independencia en una tabla de contingencia. Las frecuencias observadas se calculan con las tablas de contingencia (pandas.crosstab) y las frecuencias esperadas son calculadas por chi2\_contingency.

El **nivel de significancia** se establece mediante el argumento **alpha**, que es un valor entre 0 y 1 que representa la probabilidad de rechazar la hipótesis nula cuando es verdadera (usualmente 0,05).

El valor devuelto **chi2** es el estadístico de prueba chi-cuadrado, que mide la discrepancia entre las frecuencias observadas y esperadas en la tabla de contingencia. Un valor alto de chi2 indica una mayor discrepancia entre las frecuencias observadas y esperadas, lo que sugiere que las variables no son independientes. El **valor p** devuelto por la función indica la probabilidad de obtener un estadístico de prueba chi-cuadrado al menos tan extremo como el observado, asumiendo que la hipótesis nula es verdadera. Si el **valor p es menor que el nivel de significancia establecido**, se rechaza la hipótesis nula y se concluye que **hay una relación significativa entre las dos variables categóricas.**

De esta manera, la aplicamos para analizar las correlaciones sobre las variables predictoras y la variable predecida **Prevalencia de hipertensión.**

Luego de realizar las pruebas de independencia, se obtiene evidencia acerca de una muy fuerte relación entre las mismas, con excepción de **Consumo regular de alcohol en exceso**, la cual arroja un p-valor de 0.0546 en la prueba de independencia de Chi-cuadrado.

**Conclusión**

Concluímos que las variables seleccionadas presentan una fuerte relación con la HTA, a excepción del consumo regular de alcohol en exceso.

Se podrá extender el mismo introduciendo otras variables de la base de datos no contempladas hasta ahora.

Dado que la mayoría de las variables arrojaron un p-valor muy cercano a 0 en las pruebas de independencia de Chi-cuadrado, se sugiere profundizar el presente trabajo a fin de analizar esta situación.

Trabajo futuro

La regresión lineal es un modelo que se utiliza para analizar la relación entre una variable dependiente y una o más variables independientes continuas. Por lo tanto, no es apropiado aplicar un modelo de regresión lineal sólo con variables categóricas, ya que estas no son continuas.

Sin embargo, existen modelos de regresión que se pueden utilizar para analizar variables categóricas. Uno de estos modelos es la regresión logística, que se utiliza para predecir la probabilidad de un evento binario (sí/no) en función de una o más variables independientes categóricas o continuas. Otro modelo es la regresión de Poisson, que se utiliza para analizar datos de conteo y predecir la frecuencia de un evento en función de una o más variables independientes categóricas o continuas.

**Bibliografía**

1. 4° Encuesta Nacional de Factores de Riesgo Resultados definitivos <https://www.indec.gob.ar/ftp/cuadros/publicaciones/enfr_2018_resultados_definitivos.pdf>
2. Características epidemiológicas de la hipertensión arterial en pacientes que concurren al Hospital Británico de Buenos Aires. Campaña 2019 de Hipertensión Arterial <http://adm.meducatium.com.ar/contenido/articulos/22501690172_1651/pdf/22501690172.pdf>
3. Determinantes socioeconómicos y ambientales de la hipertensión arterial en población adulta de Argentina- Revista Cubana de Salud Pública, Vol. 42 N° 1. 2016 Dr. Valentina Natividad Viego, Dr. Karina Luján <https://revsaludpublica.sld.cu/index.php/spu/rt/printerFriendly/586/476>
4. Prueba de Independencia chi-cuadrado. Riesgo Relativo. Mg. Carmen Barreto, agosto 2013. Curso de Bioestadística de la Universidad Los Ángeles de Chimbote. <http://files.uladech.edu.pe/docente/32765808/BIOESTADISTICA/SESION%2011/S11BIOEFO.pdf>