Limpieza y Preparación de Datos Para Realizar Pruebas de Independencia Entre Ocupación y Enfermedades Crónicas

Heriberto Espino Montelongo, *175199*

***Abstract*— Este documento describe el proceso de preparación de datos, incluyendo su limpieza y transformación.**

I. INTRODUCCIÓN

En este trabajo buscamos imputar los valores faltantes de la base de datos de ENSANUT (2023) con el fin de hacer pruebas de independencia entre Ocupación y Enfermedades Crónicas en la población de México.

Se seleccionaron dos encuestas para el estudio: la Encuesta a Adultos y la Encuesta a Hogar.

Del cuestionario de hogar se recolectaron las siguientes variables de características demográficas:

TABLA I
VARIABLES DE LA ENCUESTA A HOGAR

Variable	Descripción	Tipo
H0303	Edad	Discreta
H0305	Parentesco con el jefe del hogar	Categórica
H0306	Lugar de nacimiento	Categórica
H0307	Vive con su madre	Categórica
H0308	Vive con su padre	Categórica
H0311	Habla algúna lengua indígena	Dicotómica
H0312	Habla español	Dicotómica
H0317A	Ultimo nivel de escolaridad	Ordinal
H0317G	Ultimo año de escolaridad	Discreta
H0318	Alfabetización	Dicotómica
H0321	Condición de actividad	Dicotómica
H0322	Verificación de la actividad	Categórica
H0323	Busqueda de trabajo y actividad	Categórica
H0324	Posición en la ocupación	Categórica
H0327	Ingresos en el hogar	Categórica

Se seleccionaron estas variables porque pertenecen a la misma sección del cuestionario de hogar y se considera que podrían estar más relacionadas con la Posición en la Ocupación, variable que es nuestro objetivo lleran.

Para la Encuesta a Adultos, se eligieron las variables asociadas a enfermedades crónicas. Estas fueron respondidas por todos los entrevistados mediante opciones dicotómicas: “Sí padezco la enfermedad” o “No padezco la enfermedad”, las preguntas categóricas eran dirigidas también a personas embarazadas, estas respuestas fueron excluidas del análisis.

Variables seleccionadas de la Encuesta a Adultos:

TABLA II
VARIABLES DE LA ENCUESTA A ADULTOS

Variable	Descripción	Tipo
A0301	Diabetes	Categórica
A0401	Hipertensión	Categórica
A0502A	Infartos	Dicotómica
A0502B	Angina de pecho	Dicotómica
A0502C	Insuficiencia cardiaca	Dicotómica
A0502D	Embolia	Dicotómica
A0601A	Infección en vías urinarias	Dicotómica
A0601B	Cálculos renales	Dicotómica
A0601C	Insuficiencia renal	Dicotómica

II. LIMPIEZA DE LOS DATOS PARA ENCUESTA DE HOGAR

El objetivo de las variables seleccionadas del cuestionario de hogar es para conocer la ocupación “H0324”. Los datos faltantes se ven así:

```
h0303         0
h0305         0
h0306         0
h0307         0
h0308         0
h0312     23212
h0317a      854
h0317g     2060
h0318     16050
h0327         0
h0321      4185
h0322     15156
h0323     15613
h0324     12471
h0327         0
```

Donde 0 representa que no hay datos faltantes.

A. Valores Faltantes por Estructura de la Encuesta

Hay variables que no son contestadas por la misma estructura de la encuesta, como la variable de alfabetización, que no es contestada por las personas que recibieron educación primaria o mayor.

Los que contestaron que no hablaban alguna lengua indígena no contestaron si hablaban español, por lo que si no hablaban una lengua indígena se puso que hablaban español.

Las variables H0321, H0322, H0323 y H0324 están interconectadas. La primera variable determina si la persona trabajó al menos una hora; en caso de responder 'no', se procede a la segunda pregunta.

La segunda pregunta indaga sobre cualquier actividad económica realizada durante la semana pasada. Si la respuesta indica que la persona ayudó en alguna actividad económica, se pasa directamente a la última pregunta. En cambio, si la respuesta es negativa, se avanza a la tercera pregunta.

La tercera pregunta cuestiona qué hizo la persona durante la semana pasada, omitiendo así la última pregunta.

Tras analizar los valores de cada pregunta, se concluye que los valores desconocidos en estas variables corresponden únicamente a los casos desconocidos en la primera pregunta.

Concluyendo este proceso nuestras variables quedaron de la siguiente manera:

```
h0303        0
h0305        0
h0306        0
h0307        0
h0308        0
h0312        0
h0317a     854
h0317g    2060
h0318        0
h0327        0
h0324     4185
```

B. Nivel de Escolaridad

Para imputar los 854 valores faltantes del grado de estudios, "H0317A", se construyó un árbol de decisión utilizando como variables predictoras aquellas que están completas: "H0303", "H0305", "H0306", "H0307", "H0308", "H0312", "H0318" y "H0327". En un principio nuestros datos se ves así:

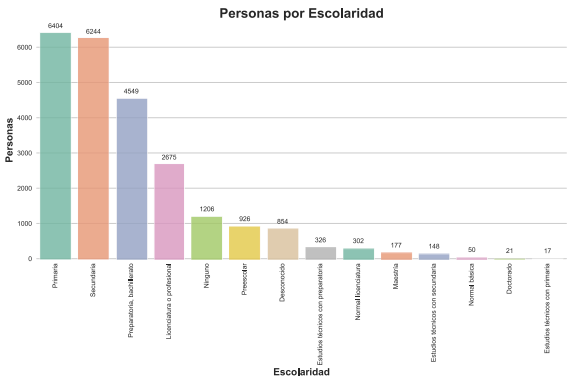

Fig. 1. Gráfico de barras de escolaridad, original.

Estos son los parámetros del árbol de decisión:

- Se utilizó el criterio de Gini, `criterion = 'gini'`
- El árbol escoge la mejor división posible, `splitter = 'best'`
- El árbol crece hasta que todos los nodos sean puros o no haya más muestras, `max_depth = None`
- El número mínimo de muestras necesarias para dividir un nodo es 2, `min_samples_split = 2`
- El número mínimo de muestras necesarias en una hoja es 1, `min_samples_leaf = 1`
- No se exige ninguna fracción mínima del peso de las muestras en las hojas, `min_weight_fraction_leaf = 0.0`
- Se consideran todas las características disponibles para dividir los nodos, `max_features = None`
- No se ha fijado un valor para controlar la aleatoriedad, `random_state = None`
- No se ajustan los pesos de las clases, `class_weight = None`
- No hay límite en el número de nodos hoja que el árbol puede generar, `max_leaf_nodes = None`
- No se realiza poda en el árbol después del entrenamiento, `ccp_alpha = 0.0`

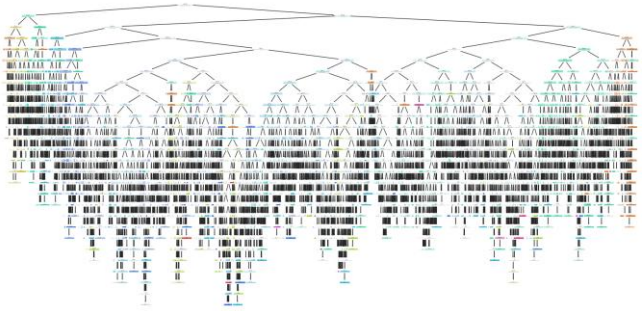

Fig. 2. Dendograma de árbol de decisión.

Luego de imputar los valores faltantes nuestro nivel de escolaridad quedo de la siguiente manera:

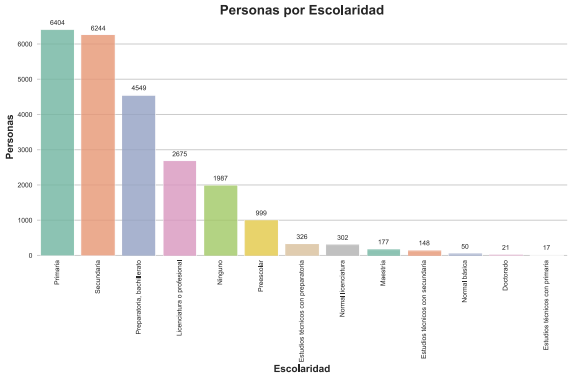

Fig. 3. Gráfico de barras de escolaridad, luego de imputar con árbol de decisión.

Concluyendo con estos valores faltantes:

```
h0303        0
h0305        0
h0306        0
h0307        0
h0308        0
h0312        0
h0317a       0
h0317g    2060
h0318        0
h0327        0
h0324     4185
```

C. Último Año de Escolaridad

Los valores faltantes en el último año de escolaridad se deben a que la persona entrevistada respondió que otro miembro del hogar estudió, por ejemplo, la primaria, pero no especificó hasta qué año.

Para imputar estos valores se utilizó una imputación basada en la mediana dentro de cada nivel de escolaridad, es decir, las observaciones se agruparon según su nivel educativo y, con base en ello, se calculó la mediana correspondiente a cada grupo para imputar los valores.

La elección de la mediana en lugar de la media se debe a que es menos sensible a valores atípicos, lo que evita que personas con escolaridad inusualmente alta o baja dentro de un nivel sesguen la imputación; preserva mejor la distribución de los datos, ya que la cantidad de años estudiados dentro de un mismo nivel educativo puede no seguir una distribución simétrica y refleja de manera más precisa el típico número de años estudiados dentro de cada nivel educativo, sin ser afectada por casos extremos, realizando así una imputación más representativa que mantiene la coherencia en los datos.

A continuación, se presenta la mediana de cada nivel de escolaridad:

```
0     0.0
1     2.0
2     5.0
3     3.0
4     3.0
5     3.0
6     3.0
7     3.0
8     3.0
9     3.0
10    4.0
11    2.0
12    3.0
```

D, Años de Estudio

Se creo una nueva variable llamada “H0317”, que es la suma de la edad esperada en la que se empieza cada escolaridad más su último año de estudios, consiguiendo el tiempo en años estudiado.

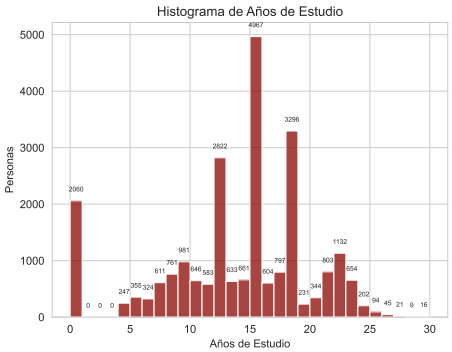

Fig. 4. Histograma de años de estudio.

Mostrando picos que muestran la edad promedio en las personas que no estudiaron, en los estudiantes que terminan la primaria, secundaria y preparatoria.

También se muestra como se ve con kde con un kernel gaussiano y una longitud de barra de 3 años para una mejor visualización:

Fig. 5. Histograma de años de estudio de KDE.

No se utilizarán las variables H0317A ni H0317G. Para analizar la escolaridad, optaremos por la variable que creamos para representar el total de años de estudio, ya que, al ser una variable cuantitativa, la podemos llevar a un espacio métrico de manera coherente, para luego aplicar métodos de imputación basados en cálculos de distancia, como el algoritmo K-Nearest Neighbors (KNN). A continuación, se presentan las variables que presentan valores faltantes:

```
h0303        0
h0305        0
h0306        0
h0307        0
h0308        0
h0312        0
h0318        0
h0327        0
h0324     4185
h0317        0
```

E. Ingresos en el Hogar

La variable de ingresos en el hogar no muestra valores nulos, pero existe la opción de preferir no contestar o no saber el ingreso del hogar, que son la 8 y la 9 respectivamente, por lo que serán tratados como valores faltantes que serán imputados con KNN, estos son los datos de la variable:

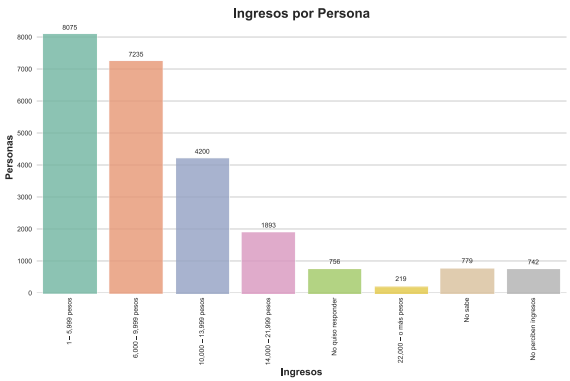

Fig. 6. Gráfico de barras de ingresos, original.

Antes de aplicar KNN tenemos que transformar las variables. Antes de transformar la variable de edad, es importante saber cómo se distribuyen los datos para luego transformarlos, entonces, así se distribuyen las edades en la encuesta:

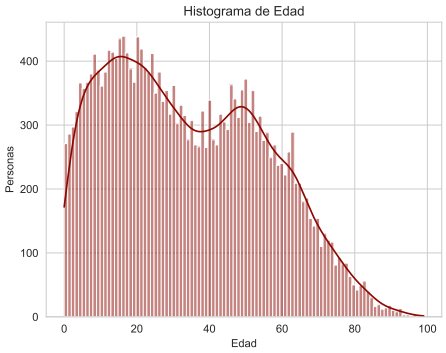

Fig. 7. Histograma de edad.

La variable no sigue una distribución simétrica, por lo que no se utilizó una estandarización con z-score, ya que esta técnica asume que los datos tienen una distribución aproximadamente normal y pueden ser afectados por la presencia de outliers.

Se consideró inicialmente la transformación de Box-Cox, que es útil para corregir la asimetría en distribuciones sesgadas. Sin embargo, Box-Cox requiere que todos los valores sean positivos, lo que representa un problema porque la variable de edad incluye valores de cero años.

Por esta razón, se optó por la transformación Yeo-Johnson que puede incluir valores iguales a cero corrige asimetrías, aproximando los datos a una distribución normal; y reduce la influencia de outliers. Así se homogeniza al escala y las variables contribuyen mas equitativamente para KNN.

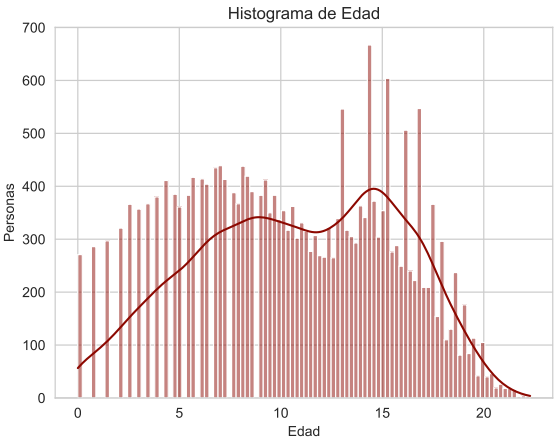

Fig. 8. Histograma de edad con transformación de Yeo-Johnson.

Después de aplicar la transformación Yeo-Johnson a la edad, se pensó hacer un escalado Min-Max (0-1), pero al final del análisis seo obtuvieron las siguientes distribuciones de edad:

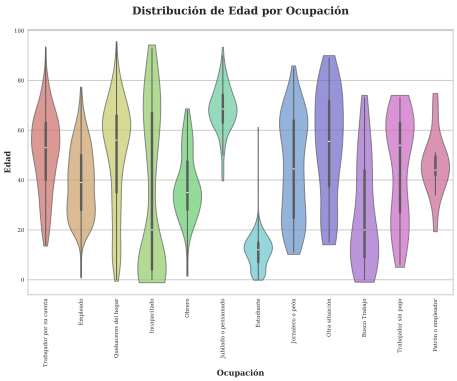

Fig. 9. Gráfico de violín para edad por ocupación con Min-Max(0,1).

Resultados que se compararon con estos, sin aplicar Min-Max(0-1):

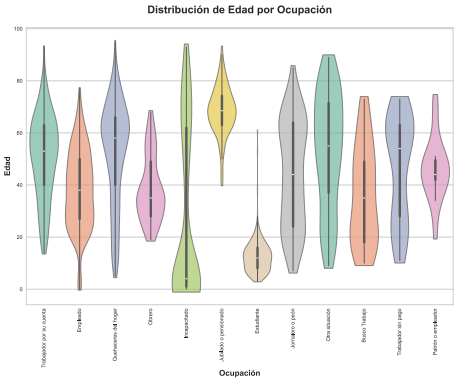

Fig. 10. Gráfico de violín para edad por ocupación sin Min-Max(0,1).

Al comparar ambas transformaciones, se observó que la

segunda opción preserva mejor la interpretación de los datos, ya que, a diferencia del primero, los estudiantes tienen al menos tres años de edad, lo cual es consistente con la realidad y los obreros tienen un mínimo de dieciocho años de edad, lo cual es razonable. En cambio, con la transformación Min-Max Scaling, los resultados eran menos coherentes, ya que aparecían estudiantes con cero años de estudio y se registraban obreros con solo tres años de edad, lo cual no tiene sentido.

Por esta razón, se decidió no aplicar Min-Max Scaling a la variable de edad y únicamente conservar la transformación Yeo-Johnson para corregir la asimetría en la distribución.

El tiempo en años de estudio también se dejó sin ninguna transformación para hacer que tenga más importancia, ya que es de las variables más importantes para conocer la Ocupación.

Finalmente, las variables categóricas se transformaron utilizando one-hot encoding, asegurando que todas las variables nominales estuvieran representadas de manera adecuada en el modelo.

Estos son los parámetros que se usaron para el KNN:

- Se utiliza 3 vecinos para la clasificación, `n_neighbors = 3`
- Se usa la distancia euclidiana estándar (sin ponderación), `weights = 'uniform'`
- Se emplea la métrica de Minkowski con orden 2 (equivalente a la distancia euclidiana), `metric = 'minkowski'`
- Se usa el algoritmo de búsqueda automática basado en la cantidad de datos, `algorithm = 'auto'`
- No se establece un valor fijo para controlar la aleatoriedad, `random_state = None`
- Se permite usar todos los puntos en la búsqueda del vecino más cercano, `leaf_size = 30`
- El parámetro de Minkowski está configurado en 2, `p = 2`
- Se consideran todos los vecinos en caso de empate, `metric_params = None`
- No se usa un conjunto específico de etiquetas objetivo, `n_jobs = None`

Luego de la imputación de valores, se obtuvieron el nivel de ingresos de la siguiente manera:

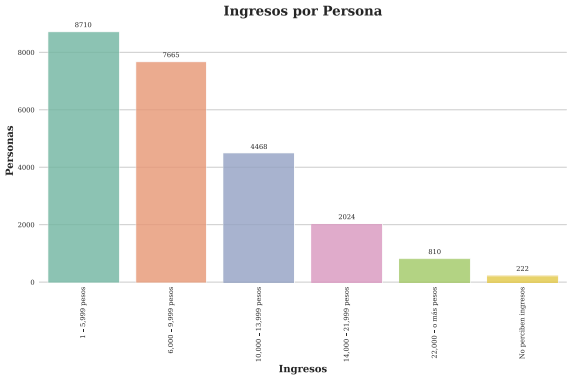

Fig. 11. Gráfico de barras de ingresos por persona, luego de imputación por KNN.

F. Ocupación

Ya no hay valores nulos o de registros que hayan preferido no contestar para todas las variables que ocuparemos para imputar los valores faltantes de la ocupación.

Las transformaciones que ahora se utilizaron son las mismas que en el caso de Ingresos al Hogar, solo que ahora, como tenemos completos los ingresos de los habitantes del hogar, que es una nominal por rangos, a cada rango se le establecerá el punto medio del rango, es decir, a la de 1-5,999 se le puso como 3000, y a la de más de 22,000 se le puso como 25,000, sin transformaciones adicionales. Así se ven las ocupaciones en la población:

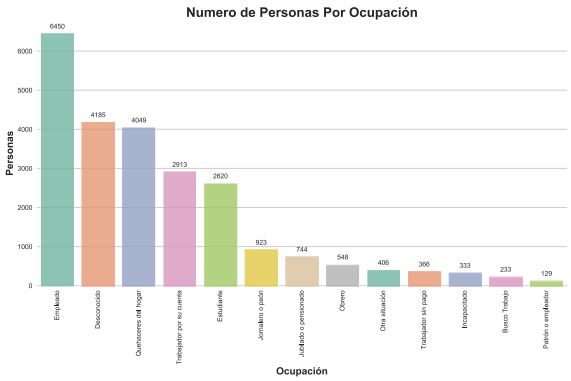

Fig. 12. Gráfico de barras de personas por ocupación, original.

Luego de imputar los valores con KNN, con los mismos parámetros que se usaron para la imputación de Ingresos al Hogar, se obtuvieron los siguientes resultados:

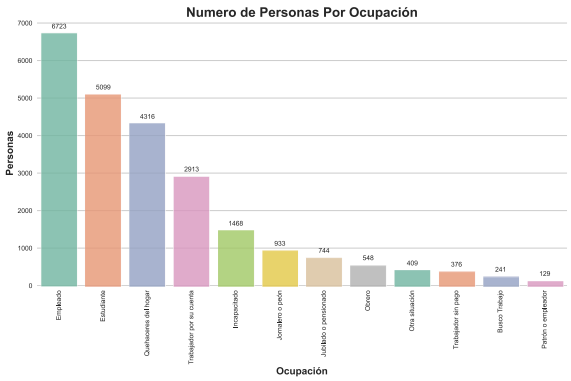

Fig. 13. Gráfico de barras de personas por ocupación, luego de imputación por KNN.

III. Limpieza De Los Datos Para Encuesta de Hogar

Las variables del cuestionario de adultos están completas, no hay valores nulos ni opción para contestar que prefieren no contestar o que no saben.

IV. Conclusión

Ahora, los resultados obtenidos son congruentes con la realidad, ver *fig. 10*. Aproximadamente una cuarta parte de las

edades es menor a 20 años, lo cual es consistente con la presencia de 5,099 estudiantes en la muestra, la categoría más frecuente corresponde a empleados, y la menos representada es la de patrones o empleadores.

Con estos ajustes y validaciones, podemos decir que los datos están listos para realizar pruebas de independencia entre ocupación y enfermedades crónicas.

Referencias

Instituto Nacional de Salud Pública (INSP), 2023. "Información sobre el hogar," distribuida por Instituto Nacional de Salud Pública (INSP), https://ensanut.insp.mx/encuestas/ensanutcontinua2023/descargas.php

Instituto Nacional de Salud Pública (INSP), 2023. "Información sobre los residentes," distribuida por Instituto Nacional de Salud Pública (INSP), https://ensanut.insp.mx/encuestas/ensanutcontinua2023/descargas.php

Instituto Nacional de Salud Pública (INSP), 2023. "Cuestionario de salud de adultos (20 años o más)," distribuida por Instituto Nacional de Salud Pública (INSP), https://ensanut.insp.mx/encuestas/ensanutcontinua2023/descargas.php