Aplicación de técnicas de minería de datos a la Encuesta de Caracterización Socioeconómica Nacional (CASEN) 2022 con R Studio

Integrante: María Paz Santibáñez

Profesor: Pablo Lemus

Fecha: 01 de octubre, 2023

Contenido

INTRODUCCIÓN

En este informe se presenta la aplicación de técnicas de minería de datos a los datos registrados por la encuesta CASEN 2022. La encuesta CASEN tiene como objetivo recopilar información sobre los hogares que habitan viviendas particulares ocupadas y sus residentes habituales en todo el territorio nacional mediante la aplicación de un cuestionario que se divide en 8 módulos. La encuesta CASEN 2022 se aplicó a 202.231 personas durante el periodo de 1 de noviembre 2022 al 2 de febrero 2023 y tuvo una tasa de respuesta del 68,7%.

Para la aplicación de las técnicas de minería de datos se utilizó el software R Studio, el conjunto de datos en formato SPSS (.sav) de la página web del Ministerio de Desarrollo Social https://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-2022, y el libro de códigos Excel (.xlsx). En este informe se consideran los módulos Registro de Residentes, Educación, Trabajo, Ingreso, Salud, Identidad, Redes y Participación, y Vivienda.

En primer lugar, se realizó un análisis exploratorio del conjunto de datos, para esto se eligieron dos preguntas de cada módulo y se analizaron los valores pueden tomar las variables, los valores faltantes y los outliers.

Luego, se analizó en detalle la variable sueldo líquido en el trabajo principal (Ingreso), realizando nuevamente un análisis exploratorio para identificar que valores puede tomar la variable, los valores faltantes y los outliers. Además, se graficó su distribución de densidad y se aplicó una transformación a su distribución.

Por último, se consideraron las variables Edad, Sexo, Jefatura de hogar, Número de personas en el hogar, Nivel de analfabetismo, Nivel educacional, Tipo de institución de educación superior, metros cuadrados de su vivienda y Región para imputar los valores faltantes mediante técnicas de imputación múltiple con la librería mice. Finalmente, se estimaron las tres variables mas relevantes para explicar la variable Ingreso aplicando algoritmos tipo Ranking y Wrapper.

DESARROLLO

DATOS NULOS EN VARIABLES DE CADA MÓDULO

A continuación, se presentan las dos preguntas de cada modulo con el porcentaje de datos nulos, el tipo de dato nulo y la respectiva justificación.

Módulo	Pregunta	Tipo de dato nulo	Porcentaje
Registro de Residentes	¿Cuál es el estado conyugal o civil actual?	MAR	16.35%
	¿Tiene dificultad para caminar o para subir escaleras?	MAR	4.729%
Educación	¿Sabe leer y escribir?	MAR	17.72%
	¿Cuál fue el último año en qué asistió a algún establecimiento educacional?	MAR	94.21%
Trabajo	En su trabajo principal, ¿qué tipo de contrato o acuerdo de trabajo tiene?	MAR	69.65%
	¿Qué medio de transporte utiliza habitualmente para realizar este viaje?	MAR	63.64%
Ingresos	Últimos 12 meses, ¿recibió ingresos por Subsidio Empleo Joven?	MAR	87.38%
	¿Recibe Pensión Garantizada Universal (PGU)?	MAR	98.98%
Salud	¿Cuál es el estado nutricional?	MAR	89.17%
	¿A qué sistema previsional de salud pertenece?	MAR	15.24%
Identidades, Redes y Participación	¿En qué período llegó al país?	MAR	99.85%
	Hasta los 15 años, ¿la jefatura de hogar vivió con alguno de sus padres?	MAR	64.36%
Vivienda	¿Cuál es la situación del título de propiedad de este sitio o inmueble?	MAR	39.36%
	: ¿Compró la vivienda con ayuda de algún programa o subsidio del Estado?	MAR	38.90%

Tabla 1: Variables, tipo de dato nulo, porcentaje de datos nulos y justificación.

ÁNALISIS DETALLADO DE LA VARIABLE INGRESO

La variable sueldo líquido en el trabajo principal (Ingreso) contiene 202.331 observaciones, es de tipo "haven_labelled" la cual contiene una etiqueta de datos que toma valores numéricos y caracteres. A continuación, se presenta una tabla resumen con la información de esta variable, esta información se obtiene aplicando la función "summary" a la variable.

Etiqueta numérica	Etiqueta carácter	Observaciones
-88	No sabe	1,640
0	No recibió sueldo	1,483
8000 - 25000000	Monto	58,240

Tabla 2: Resumen variable ingreso

Además, la media de la variable Ingreso es de 643.252CLP, y su mediana es de 477.000CLP, como se observa en el histograma.

Imagen 1: Histograma de la variable ingreso

DATOS NULOS VARIABLE INGRESO

Del total de observaciones existen 140.869 datos nulos, lo que corresponde a un porcentaje 69,65% y el tipo de dato nulo es MAR debido a que la perdida del valor de esta observación tiene relación con.

OUTLIERS VARIABLES INGRESO

Se realizó el análisis de valores extremos mediante z-score utilizando dos valores para el umbral z identificando dos tipos de outliers en la variable Ingreso. Previo a la aplicación del algoritmo z-score se realizó el cambio de formato de la variable "No sabe" a variable "NaN" (Not a Number) para luego omitir las observaciones "NA" y "NaN" quedando 59.723 observaciones válidas.

En primer lugar, se identificó una única observación con el máximo valor en la distribución de ingresos correspondiente al monto de 25.000.000CLP, donde z toma valor 23, el cual representa un outlier global y su porcentaje es de 0,0017%. En segundo lugar, se identificaron 202 observaciones correspondientes a outliers colectivos con ingresos en el rango de 5.000.000CLP a 15.000.000CLP, donde z toma el valor 6.7, y su porcentaje es de 0,3382%.

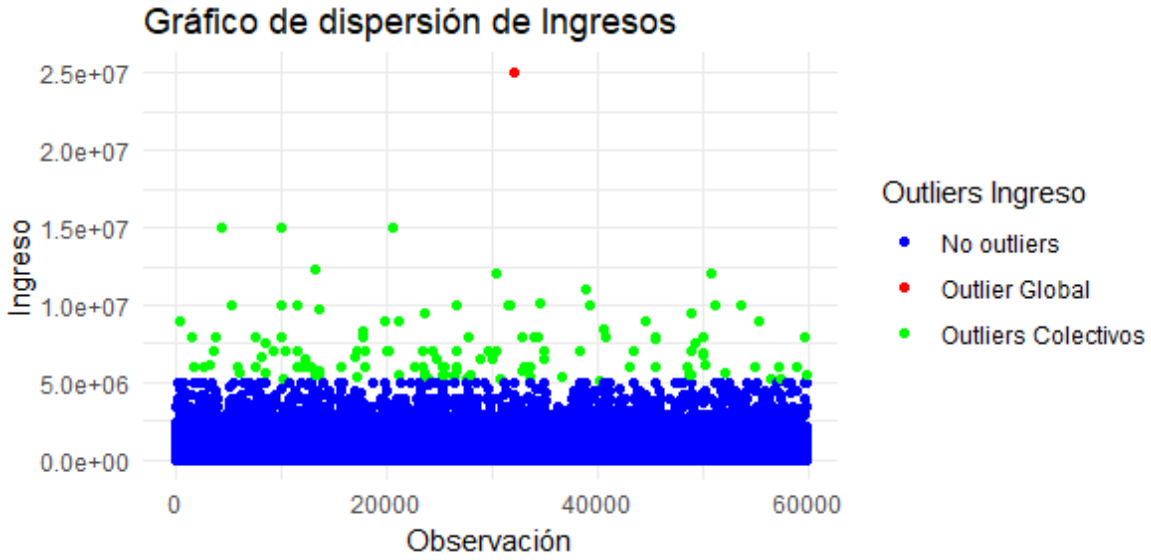

Imagen 2: Dispersión de la variable Ingreso y outliers

DISTRIBUCIÓN DE DENSIDAD VARIABLE INGRESO

A continuación, se presenta el grafico de la distribución de densidad de la variable Ingreso la cual es del tipo asimétrica positiva, de la imagen 3 se observa que el mayor porcentaje de observaciones se concentran en el rango desde 0CLP hasta aproximadamente los 2.500.000CLP. Para luego disminuir su densidad considerablemente para los montos mayores a 5.000.000CLP.

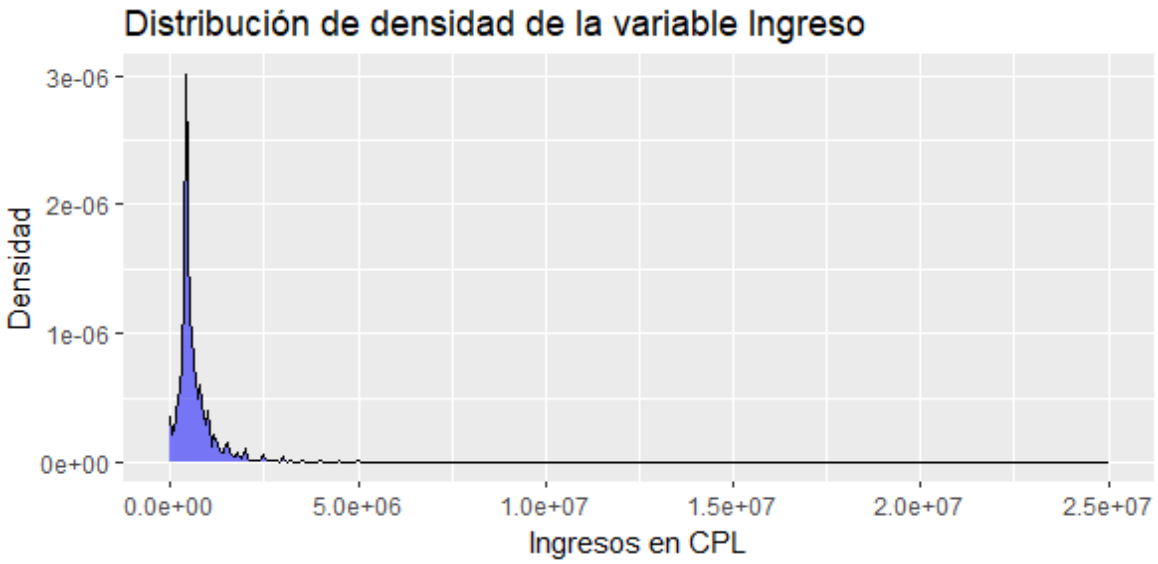

Imagen 3: Distribución de densidad de la variable Ingreso

Con el objetivo de disminuir el sesgo que puede tener el ruido de nuestra distribución y facilitar el ajuste de los parámetros se realiza una transformación de la distribución aplicando la raíz cuadrada a la función de densidad de la variable Ingreso.

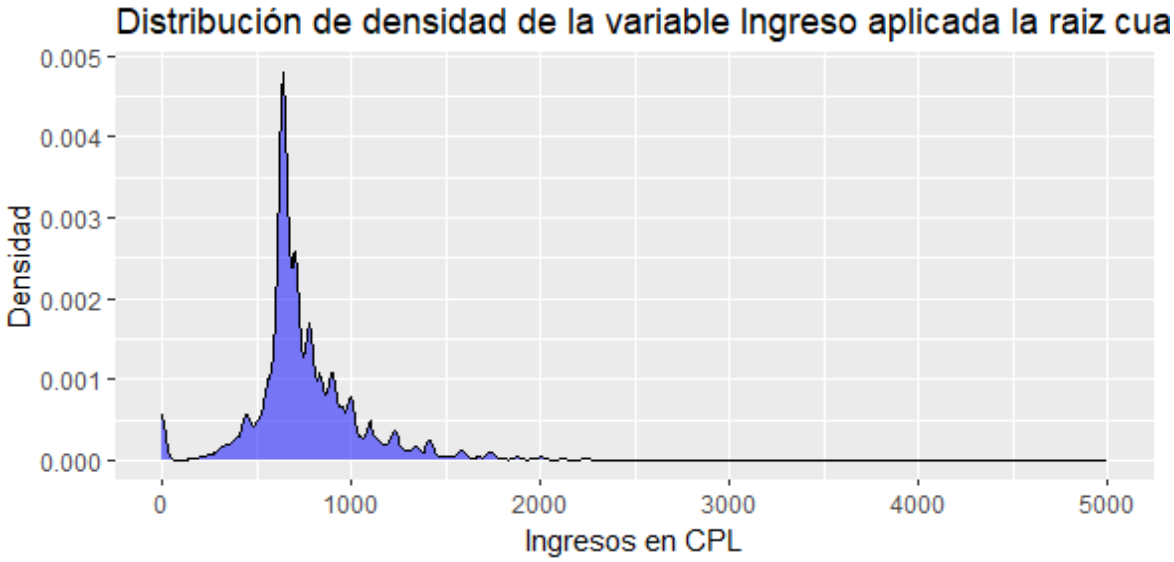

Imagen 4: Transformación de la distribución de densidad de variable Ingreso

IMPUTACIÓN DE DATOS NULOS DE LA VARIABLE INGRESO

Dado que el porcentaje de datos nulos de la variable ingreso es de 65,69% se realizó la imputación de los datos nulos utilizando tres métodos de imputación múltiple y se consideraron las variables edad, sexo, región, jefe de hogar, total personas en el hogar, nivel de analfabetismo, nivel educacional, tipo de institución de educación superior y metros cuadrados de la vivienda.

Para esto fue necesario emplear la librería "mice" la cual permite imputar datos mediante el método pmm (Predictive Mean Matching) que imputa a través de la media, el método cart (Classification and Regression Trees) que utiliza arboles de decisión y, por último, el método Lasso Regression que emplea una regresión Lasso para imputar los datos faltantes.

A continuación, la imagen 4 muestra en rojo el histograma de la distribución original de la variable ingreso, luego, en verde se tiene el histograma de la distribución de la variable ingreso habiendo imputado los datos nulos por el método pmm. Luego, en azul se tiene el histograma de la distribución de los ingresos habiendo imputado los datos nulos por el método cart, y finalmente, en amarillo se tiene el histograma de la distribución de la variable ingreso habiendo imputado los datos nulos por el método de regresión Lasso.

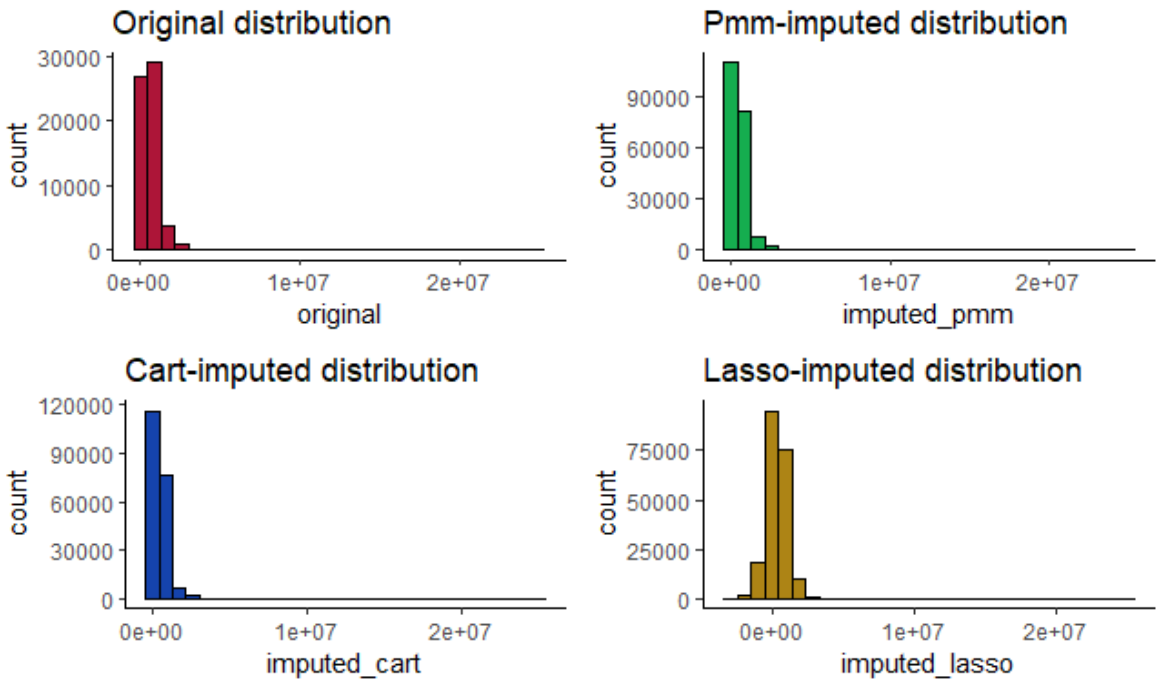

Imagen 5: Imputación de los datos nulos variable Ingreso con librería mice

Para terminar, se observa que el método por regresión Lasso se diferencia mucho de la distribución original y se obtienen valores negativos para los ingresos, lo que no tiene sentido en este contexto, por lo tanto, se descarta. Los dos métodos restantes, pmm y cart presentan

una distribución similar a la variable ingreso, y finalmente, se decide por el método cart ya que la media no representa una buena estimación de la distribución real de los ingresos.

SELECCIÓN DE VARIABLES

A continuación, se realizó la selección de atributos o variables que mejor explican a la variable ingreso mediante el uso de metodologías de Ranking y Wrapper considerando las variables edad, sexo, región, jefe de hogar, total personas en el hogar, nivel de analfabetismo, nivel educacional, tipo de institución de educación superior y metros cuadrados de la vivienda.

Para la metodología de Ranking se aplicaron dos métodos de filtros de entropía basados en ganancia de información.

En primer lugar, se utilizó “information gain” el cual indica que las variables más significativas para estimar la variable ingreso son el nivel de educación, tipo de institución de educación superior y la edad.

En segundo lugar, se utilizó “gain.ratio” el cual indica que las variables más significativas para estimar la variable ingreso son tipo de institución de educación superior, el nivel de educación, y el nivel analfabetismo.

Luego, para los tipos Wrapper, se utilizaron algoritmos de envoltura Greedy Search: Forward, Backward, debido a problemas con el formato de la base de datos utilizada no se pudo aplicar satisfactoriamente estos algoritmos. Es fundamental para este informe identificar el error para poder obtener las variables que mejor explican a la variable ingreso.

Para finalizar, se considera que el método de filtro basado en información “gain.ratio” es el mejor método de selección de atributos, donde considero que la importancia de las variables tipo de institución de educación superior es 0.1070, el nivel de educación es 0,.0867 y el nivel analfabetismo es 0.0505.

CONCLUSIÓN

Para concluir, se encontró que las variables que mejor explican la variable de ingreso son el nivel educativo alcanzado, el tipo de institución de educación superior donde se estudió, y el nivel de alfabetismo. Estos factores son indicativos de la importancia de la educación en la determinación del ingreso de los chilenos. Además, se determinó que el mejor método para imputar los datos faltantes es la regresión y clasificación de árboles de decisión. Este método proporciona una técnica robusta y eficiente para manejar los datos incompletos, lo que mejora la calidad y precisión del análisis.

La minería de datos demostró ser una herramienta indispensable para extraer conclusiones significativas de este conjunto de datos. La capacidad de identificar patrones y relaciones en grandes conjuntos de datos es fundamental en la era actual de la información. Sin embargo, se encontraron dificultades al aplicar algunas técnicas en R Studio. Por lo tanto, se recomienda continuar estudiando los algoritmos y los tipos de datos en R Studio para superar estos desafíos. La mejora continua en el manejo y análisis de datos es esencial para mantenerse al día con las tendencias actuales en ingeniería aplicada y minería de datos.

REFERENCIAS

Buuren, S. v. (2018). *Flexible Imputation of Missing Data.* Obtenido de https://stefvanbuuren.name/fimd/

Lemus, P., & Madariaga, G. (2023). Apuntes Data Mining [Clase 1 - 8] [Ayudantias 1 - 4].

Ministerio de Desarrollo Social y Familia. (s.f.). *Encuesta de caracterización socioeconómica nacional 2022*. Obtenido de https://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-2022

Rojas-Jimenez, K. (2022). *Ciencia de Datos para Ciencias Naturales.* Obtenido de https://bookdown.org/keilor_rojas/CienciaDatos/

Romanski, P., Kotthoff, L., & Schratz, P. (23 de agosto de 2022). *Package 'FSelector'*. Obtenido de Documentación para el paquete 'FSelector' versión 0.34: https://search.r-project.org/CRAN/refmans/FSelector/html/00Index.html

Therneau, T., Atkinson, B., & Brian, R. (21 de octubre de 2022). *rpart: Recursive Partitioning and Regression Trees*. Obtenido de https://cran.r-project.org/web/packages/rpart/index.html

van Buuren, S. (5 de julio de 2023). *Multivariate Imputation by Chained Equations*. Obtenido de Documentation for package 'mice' version 3.16.0: https://search.r-project.org/CRAN/refmans/mice/html/00Index.html