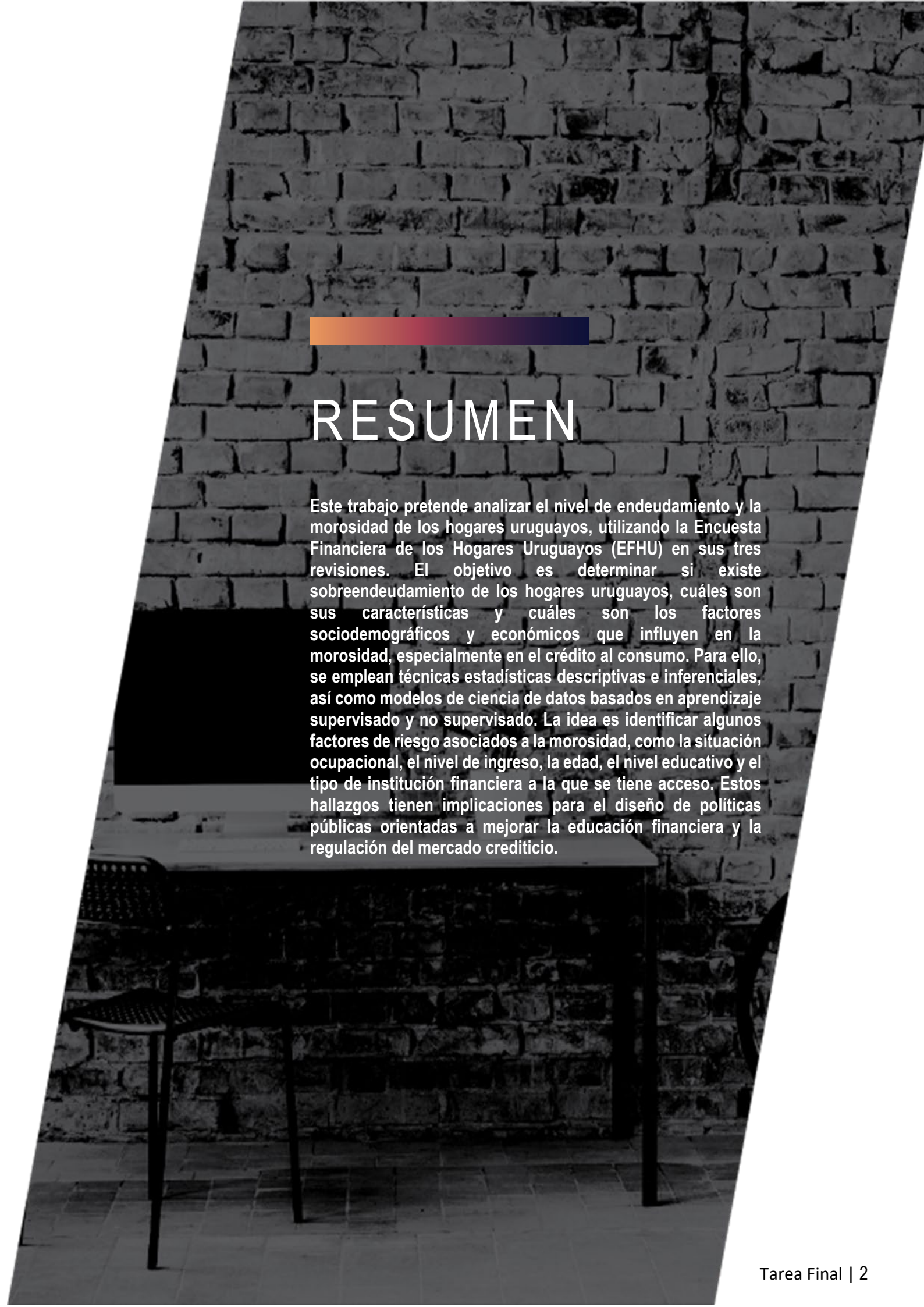
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERIA

TAREA FINAL

CURSO DE INTRODUCCIÓN A LA CIENCIA DE DATOS 2023

Lic.Ec.Federico Bentos.
C.I.: 3.991.400-0
Grupo: 20



RESUMEN

Este trabajo pretende analizar el nivel de endeudamiento y la morosidad de los hogares uruguayos, utilizando la Encuesta Financiera de los Hogares Uruguayos (EFHU) en sus tres revisiones. El objetivo es determinar si existe sobreendeudamiento de los hogares uruguayos, cuáles son sus características y cuáles son los factores sociodemográficos y económicos que influyen en la morosidad, especialmente en el crédito al consumo. Para ello, se emplean técnicas estadísticas descriptivas e inferenciales, así como modelos de ciencia de datos basados en aprendizaje supervisado y no supervisado. La idea es identificar algunos factores de riesgo asociados a la morosidad, como la situación ocupacional, el nivel de ingreso, la edad, el nivel educativo y el tipo de institución financiera a la que se tiene acceso. Estos hallazgos tienen implicaciones para el diseño de políticas públicas orientadas a mejorar la educación financiera y la regulación del mercado crediticio.

INTRODUCCIÓN

JUEGO DE DATOS

El Juego de Datos escogidos para este trabajo incluye la Encuesta Financiera de los Hogares Uruguayos (EFHU) en sus tres revisiones.

La Encuesta Financiera de los Hogares Uruguayos (EFHU) es la primera encuesta en el Uruguay que recoge información sociodemográfica y económico-financiera del hogar, relevando con detalle la tenencia, composición y valor de activos y pasivos, ingresos y egresos, así como también acceso a mercados financieros y uso de medios de pago[1].

La EFHU se implementó en dos etapas. La primera, EFHU-1, consistió en la aplicación de un módulo adicional en la Encuesta Continua de Hogares 2012 del INE. Dicho módulo incluyó 28 preguntas para relevar la tenencia y valor de activos y pasivos para una muestra de 8191 hogares. En la segunda etapa, EFHU-2, se recogió información de 3490 hogares residentes en todo el territorio nacional entre octubre de 2013 y julio de 2014. El cuestionario incluye aproximadamente 500 preguntas que relevan de forma exhaustiva la situación de los hogares en términos de activos, pasivos, productos financieros y uso de medios de pago. En la muestra se sobre ponderaron los hogares pertenecientes a los dos quintiles de mayores ingresos como es usual en este tipo de encuestas. Dando continuidad al proyecto, en 2017 se realizó un nuevo relevamiento utilizando el cuestionario de la EFHU-1, dando lugar a la EFHU-3. Este relevamiento estuvo a cargo del Instituto Nacional de Estadística (INE) y se realizó como un módulo adicional de la Encuesta Continua de Hogares 2017.

CARACTERISTICAS

La ECH aporta información socioeconomica de los Hogares Uruguayos, la EFHU tuvo como objetivo conocer ademas a través de un modulo especializado en algunas ediciones de la ECH, las caracteristicas en cuanto a la información economica-financiera de los Hogares en todas sus dimensiones.

La ECH es una encuesta por muestreo, La unidad de analisis va dirigida a la población que reside en viviendas particulares y que integra hogares particulares, por lo que quedan excluidos tanto las viviendas como los hogares colectivos (hoteles, conventos, cuarteles, hospitales). No obstante lo establecido anteriormente, se incluyen los hogares, que formando un grupo independiente, residen en estos establecimientos, como puede ser el caso de los encargados, caseros, porteros,etc. La unidad geografica recolectada por la ECH, cubre Montevideo, las localidades urbanas de 5.000 habitantes o más, las localidades urbanas de menos de 5.000 habitantes y el área rural. Su Universo de estudio es la totalidad de los hogares particulares que habitan viviendas particulares en el territorio nacional. El procedimiento de muestreo de los hogares que participan en la ECH es seleccionarlos al azar utilizando el marco muestral proveniente del Censo 2011, bajo un diseño muestral complejo que incluye varias etapas de selección y busca brindar estimaciones confiables reduciendo los costos de la recolección de la información al mínimo posible. Las viviendas incluidas en la muestra, en cada uno de los estratos, no son seleccionadas directamente. En una primera etapa son seleccionadas áreas pequeñas bien definidas, las que se denominan unidades primarias de muestreo (UPM). Las UPM corresponden a las zonas censales (conglomerados de viviendas) y son seleccionadas con probabilidad proporcional al tamaño, en base a la cantidad de viviendas particulares. En una segunda etapa se seleccionan cinco viviendas titulares (Unidad última de muestreo <UUM>) y dos viviendas suplentes, con igual probabilidad de selección dentro de cada UPM seleccionada en la primera etapa. La muestra teórica seleccionada para la Encuesta

[1] Encuesta Financiera de los Hogares Uruguayos – Facultad de Ciencias Sociales

Continua de Hogares 2012 fue de 50.523 hogares. La tasa de respuesta fue de 86,8%. Se cuenta con variables que continen los expansores del año, del semestre, del trimestre y del mes[2].

En particular este trabajo pretenderia centrarse en estudiar el nivel de endeudamiento de los Hogares Uruguayos y estudiar los determinantes sociodemograficos y economicos de los diferentes niveles de endeudamiento y en particular la morosidad. La pregunta rectora de este trabajo consiste en cuestionarse si existe sobre endeudamiento de los Hogares Uruguayos, cuales son sus caracteristicas, y cuales son los determinantes de la morosidad para los diferentes pasivos tomados por los Hogares Uruguayos, en particular interesaria estudiar los Niveles de Morosidad de los Hogares Uruguayos en cuanto al Crédito al Consumo.

La EFHU en sus diferentes versiones captura información respecto a las tenencias de Activos y Pasivos de los Hogares, en cuanto a "Vivienda Principal", "Otras Propiedades", "Vehículos", "Negocios Propios", "Deudas no relacionadas con la adquisición de la vivienda principal", "Percepción de carga financiera", "Medios de pago", y "Activos financieros"[3].

Siguiendo con el interes en estudiar los Niveles de Morosidad de los Hogares Uruguayos en cuanto al Crédito al Consumo, se podría Analizar dentro de "Deudas no relacionadas con la adquisición de la vivienda principal", las deudas con instituciones financieras no bancarias, contando en el cuestionario con información respecto a si las tiene, cual es su monto mensual y si esta al día con ella. Tambien se puede analizar dentro de esta misma categoría si mantiene deudas con bancos (fuera de la adquisición de vivienda principal), casas comerciales, automotoras, familiares, amigos u otras personas.

C A L I D A D D E D A T O S

En cuanto a la información y su disponibilidad, se hace incapie en que esta información no se encuentra online para todas las EFHU, ya que solo la EFHU-1 se encuentra disponible online en la pagina de UDELAR-Ciencias Sociales, pero se puede acceder a los microdatos de las EFHU-2 y EFHU-3, solicitandolos previamente a la dirección de correo: efhu_decon@cienciassociales.edu.uy (**Microdatos de Módulo Financiero y ECH-2017).**

Otro problema asociado a esto, es los recursos que se cuentan para realizar este relevamiento y la cantidad de instituciones publicas e investigadores que involucra, contandose actualmente con la información mas reciente a nivel del año 2017, sería bueno contar con este módulo de la ECH con información mas actual.

De todos modos la idea sería ver como con los diferentes conjuntos de datos a lo largo del tiempo (2012-2014-2017) evoluciono esta realidad en el marco del proceso nacional e internacional de inclusión financiera y la posibilidad de parte de la población de acceder a estos servicios y mercados, analizando como cambian los comportamientos de los hogares a la luz de este proceso.

Atendiendo asimismo a la calidad de los Datos se encuentra que el DECON cuenta para la última revisión de este módulo (Año 2017) con 2 bases de datos; "efhu3 revisada.dta" y "efhu3 revisada imputada.dta". La segunda es creada en base a agregar valores imputados a los casos de no respuesta al item en las variables de valores de activos y pasivos. Se realizó un procedimiento de imputación estocástica múltiple, generando diez juegos. La base incluye también las variables necesarias para trabajar con bases con imputación múltiple en STATA, y variables sombra (flags) que indican si la observación fue imputada[4][5]. La cantidad y proporción en el total de datos imputados podría poner en duda los resultados que se obtuvieran del analisis.

[2] ine.gub.uy/Anda5/index.php/catalog/725

[3] Cuestionario_EFHU3.pdf (cienciassociales.edu.uy)

[4] En el caso de los flags el valor "0", true missing, indica que el hogar no tenía que responder esa pregunta, "1" respuesta válida y "2050" no sabe/no contesta y corresponde imputar el valor.

[5] Cuestionario_EFHU3.pdf (cienciassociales.edu.uy)

PROBLEMA/PREGUNTA

QUE SE PUEDE RESOLVER UTILIZANDO LAS HERRAMIENTAS PRESENTADAS EN EL CURSO

Morosidad

- ¿Es un fenomeno creciente a lo largo del tiempo?

Analizar los determinantes sociodemograficos y economicos de la Morosidad de los Hogares Uruguayos en cuanto a el Mercado de Crédito al Consumo.

Analizar la relación entre niveles de ingreso, endeudamiento y morosidad de los Hogares Uruguayos en relación al crédito al consumo.

Analizar la relación entre situación ocupacional, endeudamiento y morosidad de los Hogares Uruguayos en relación al crédito al consumo.

Credito al Consumo

- ¿Qué caracteristicas tienen los Hogares Uruguayos que se endeudan con el sistema financiero no bancario?

Analizar el porcentaje de Hogares Uruguayos que estan en morosidad en relación al crédito bancario y no bancario, y su evolución en el tiempo.

Caracterizar a los Hogares Uruguayos en cuanto al tipo de acceso y utilización del sistema financiero bancario y no bancario.

Caracterizar a los hogares Uruguayos en base al mayor endeudamiento con el sistema financiero bancario y no bancario.

Acceso al Crédito

- ¿Cómo ha evolucionado el acceso a instrumentos financieros de crédito?
- ¿El sobreendeudamiento puede afectar el acceso futuro al mismo?

Caracteristicas sociodemograficas y economicas de los Hogares Uruguayos por tipo de Acceso al Crédito y Motivo del mismo.

Predecir en base a las caracteristicas sociodemograficas y economicas de los Hogares, cuales pueden caer en situación de Morosidad y que porcentaje de los Hogares Uruguayos pueden caer en esta categoría.

PROCESO

QUÉ MÉTODOS SE PODRÍAN APLICAR, CON QUÉ CUIDADOS, QUÉ VISUALIZACIONES USARÍA, TANTO PARA EXPLORAR LOS DATOS COMO PARA EXPLICAR RESULTADOS

Aprendizaje Automático

ANALISIS EXPLORATORIO

La idea es ver si existen datos faltantes, duplicados, imputados y realizar un join con la información de la ECH que soporta la muestra de esta encuesta EFHU, para obtener información relacionada con las caracteristicas sociodemograficas y economicas de los hogares.

Aprendizaje No Supervisado

ANALISIS DESCRIPTIVO

La idea es realizar un summary de las principales estadisticas para las diferentes variables de analisis, y mostrar visualizaciones de esta estadisticas como tablas, visualizaciones gráficas a través de graficos de distribución, histogramas, box-plot y graficos de violines para las principales variables analizadas y diferencias existentes de acuerdo a variables categoricas de interes. Por ejemplo como se distribuye la morosidad por niveles de ingreso, niveles de estudio, situación ocupacional del hogar, variables categoricas que discriminen los hogares donde hay empleados publicos y privados, caracteristicas sociodemograficas como cantidad de integrantes del hogar, situación de migración, etc.

ANALISIS MULTIVARIADO

En base a la cantidad de variables con las que cuenta la ECH (para propiciar las dimensiones de Analisis) y la EFHU (variables de interes a explicar), sería bueno aplicar tecnicas de reduccion de la dimensionalidad del problema al intentar cruzar tantas variables y realizar un Analisis Multivariado del fenomeno.

Aprendizaje Supervisado

Luego de un Analisis exploratorio de las bases de datos y un Analisis Descriptivo considerando diferentes variables, se sugiere realizar un Analisis de Componentes Pincipales (PCA) para intentar reducir la dimensionalidad del problema y obtener los primeros componentes principales de manera de capturar la mayor cantidad de información (medida a través de la variabilidad de la misma) en la menor cantidad posible de ejes, analizando y caracterizando estos componentes en base a las correlaciones con las variables originales, dimensiones de este analisis.

Se deberia revisar si el plano factorial (primeros 2 componentes) es capaz de capturar suficiente variabilidad y caracterización de las variables originales utilizadas para el analisis. Si esto no fuera posible debería incluirse la cantidad de componentes necesarios para intentar capturar un porcentaje elevado de la varianza total.

CARACTERIZACIÓN DE LOS HOGARES URUGUAYOS

En base a la realización de Analisis de Aprendizaje no Supervisado como tecnicas de clustering se busca caracterizar a los Hogares Uruguayos en cuanto a las dimensiones sociodemograficas y economicas de interes recogidas en ambas encuestas como ser la

situación ocupacional, el nivel de ingresos, el nivel de endeudamiento del hogar, la cantidad de integrantes, las caracteristicas del jefe de hogar, el nivel de estudios, el nivel de acceso al mercado créditicio no formal y formal (bancario y no bancario), situación de migración, etc, en relación a los niveles de endeudamiento y cumplimiento con sus deudas en el mercado de créditos.

MODELOS DE APRENDIZAJE

La idea seria caracterizar y responder; "Cuales son los determinantes de la Morosidad de los Hogares Uruguayos respecto al Crédito al Consumo" medido como tener "Deudas no relacionadas con la adquisición de la vivienda principal" y "No estar al día con ellas". En particular interesa en este trabajo las deudas consultadas en la pregunta: *MF12 2 1. ¿Tiene alguna deuda con instituciones financieras no bancarias (como OCA, ANDA, Pronto!, Creditel, Cooperativas de ahorro y crédito, etc.)?* , aunque tambien deberían considerarse como un Proxy al crédito al consumo toda la dimensión "Deudas no relacionadas con la adquisición de la vivienda principal".

En base a este objetivo se proponen realizar diferentes modelos de predicción de que un Hogar mantenga deudas impagas, con el sistema financiero no bancario, pero tambien con otro tipo de agentes dentro de la categoría mencionada anteriormente, en relación a las variables sociodemograficas y economicas del hogar relevadas en la ECH, asi tambien como con relación a variables propias del módulo de la EFHU donde por ejemplo se tiene la información del monto de deuda mantenida y en la dimensión "Percepción de carga financiera del Hogar" se cuenta con información respecto a: *MF18. ¿Qué porcentaje del total de los ingresos mensuales del hogar destina al pago de sus deudas (incluyendo el pago de préstamos para la compra de su vivienda principal)? Recuerde que se pregunta por deudas, no se deben incluir aquellos gastos corrientes tales como compra de alimentos, ropa, pago de alquiler, pago de luz, agua, impuestos, etc; excepto que estos pagos estén atrasados.*

La idea es utilizar como variable dependiente la variable binaria que identifica si el hogar mantiene deudas impagas, tanto con el sector financiero no bancario, como con el bancario.

Como variables independientes se plantea utilizar el mayor conjunto de variables sociodemograficas y economicas posibles que refieren a, situación ocupacional, nivel de ingresos, nivel de endeudamiento del hogar, cantidad de integrantes, caracteristicas sociodemograficas del jefe/a de hogar, nivel de estudios, nivel de acceso al mercado créditicio formal y no formal, situación de migración.

En una segunda instancia sería bueno agregar variables independientes vinculadas a la situación macroeconomica del país; nivel de producto, tasa de variación del producto, nivel de tasa de interes, variación de la misma, tipo de cambio, su tasa de variación, nivel del salario nominal, su tasa de variación, inflación, etc. Estas variables de contexto podrían estar influyendo en la situación de morosidad del hogar y podrían ser explicativas de los porcentajes generales de morosidad registrados a lo largo del tiempo.

La idea es aplicar algunos métodos de aprendizaje supervisado, que consiste en entrenar un modelo con datos etiquetados, es decir, con la variable de morosidad conocida, para luego aplicarlo a nuevos datos y predecir la morosidad. Este método requiere tener una muestra representativa y suficiente de datos históricos con la variable de morosidad observada, así como definir una medida de error o pérdida para evaluar el desempeño del modelo.

Dentro de los algoritmos utilizados para entrenar esos modelos podemos escoger:

1. ***La regresión logística***, que es un modelo lineal que estima la probabilidad de que un hogar sea moroso o no, en función de un conjunto de variables explicativas. Este

modelo es simple, interpretable y fácil de implementar, pero puede tener limitaciones para capturar relaciones no lineales o interacciones entre las variables.

2. ***Los árboles de decisión***, que son modelos que dividen el espacio de los datos en regiones homogéneas según unas reglas basadas en las variables explicativas. Estos modelos son intuitivos, flexibles y capaces de manejar variables categóricas y numéricas, pero pueden ser sensibles al ruido o al sobreajuste.
3. ***Modelo Multinomial Naive Bayes,*** aplicado con los componentes principales que son ortogonales, el modelo multinomial naive Bayes puede tener un sesgo mayor y una varianza menor que el modelo multinomial bayesiano, debido a la suposición de independencia entre las variables explicativas. Esto significa que el modelo multinomial naive Bayes puede ser más simple y estable, pero también menos flexible y preciso que el modelo multinomial bayesiano.
4. ***Las redes neuronales***, que son modelos que simulan el funcionamiento del cerebro humano, con capas de neuronas artificiales que procesan la información y aprenden de los datos. Estos modelos son potentes, escalables y capaces de capturar relaciones complejas y no lineales entre las variables, pero pueden ser difíciles de interpretar, entrenar y optimizar.

Para evaluar el rendimiento o la calidad de los modelos, se pueden usar diferentes criterios, dependiendo del tipo y la distribución de los datos. Algunos ejemplos son:

1. ***La precisión***, que es la proporción de casos correctamente clasificados por el modelo. Este criterio es simple y fácil de calcular, pero puede ser engañoso si hay un desbalance entre las clases (por ejemplo, si hay muchos más hogares no morosos que morosos).
2. ***La sensibilidad y la especificidad***, que son las proporciones de casos positivos (morosos) y negativos (no morosos) correctamente clasificados por el modelo. Estos criterios son útiles para medir el poder discriminativo del modelo, pero pueden estar en conflicto entre sí (por ejemplo, si se aumenta la sensibilidad se puede disminuir la especificidad).
3. ***El AUC*** (área bajo la curva ROC), que es una medida que resume el rendimiento del modelo a lo largo de diferentes umbrales de clasificación. Este criterio es robusto frente al desbalance de clases y permite comparar diferentes modelos independientemente del umbral elegido.

Respecto a las visualizaciones para este tipo de modelos se pueden plantear tablas con mapas de calor mostrando los true positive y true negative arribados por los diferentes modelos, asi como mostrar en reportes de tablas los indicadores de bondad del ajuste de cada modelo especificado, estimado, tanto en los datos de train como de test. Para esto se recomienda el uso de Matrices de Confusión y versiones de mapa de calor de estas.