Universidad de San Andrés
Departamento de Economía
Big Data 2023
TP Nº 4

Análisis de la Encuesta Permanente de Hogares mediante Machine Learning

Parte I: Análisis de la base de hogares y cálculo de pobreza

La Encuesta Permanente de Hogares (EPH), implementada por el Instituto Nacional de Estadística y Censos (INDEC) en Argentina, constituye una herramienta estadística esencial para el estudio de diversos aspectos socioeconómicos de los hogares del país. Este instrumento recopila información integral sobre empleo, ingresos, condiciones de vida, y otros factores críticos que inciden en la estructura social. En este trabajo se propone una metodología detallada para el análisis de los datos de la EPH del primer trimestre de 2023 utilizando técnicas de Machine Learning (ML), con el fin de desarrollar modelos predictivos para la estimación de la probabilidad de caer sobre la línea de pobreza en los hogares argentinos, abarcando escenarios tanto con datos de ingresos como sin ellos.

El primer paso en este análisis fue la integración de los datos de las encuestas a nivel individual y hogar. Se emparejaron los datos relacionados con las características de la vivienda y cada individuo encuestado utilizando dos identificadores únicos: CODUSU y NRO HOGAR. Esta fusión es fundamental para una comprensión completa que vincule las condiciones del hogar con las características individuales de sus integrantes.

En cuanto a la preparación de los datos para el análisis mediante ML, se implementaron varias etapas cruciales. Inicialmente, se procedió a la eliminación de valores codificados como '9', que indicaban respuestas del tipo 'No Sabe / No Responde', a través de una función diseñada específicamente para este propósito. Posteriormente, se utilizó otra función ad hoc para detectar y eliminar columnas duplicadas, garantizando así la unicidad de la información. Además, se descartaron columnas y filas del DataFrame que presentaban un porcentaje de valores faltantes superior al 70%, con el objetivo de preservar la integridad del conjunto de datos.

Otro paso clave fue la transformación de variables categóricas con múltiples respuestas en variables dummy, facilitando así su interpretación y uso en los modelos de ML. Se eliminaron variables temporales, como el año y trimestre de realización de la encuesta y el año de nacimiento de la persona encuestada, por considerar que no aportaban información relevante para la predicción de pobreza. Para atenuar el impacto de outliers, se excluyeron los casos extremos identificados fuera del rango intercuartílico. Finalmente, se descartaron todos los valores faltantes restantes para asegurar la coherencia y completitud del dataset.

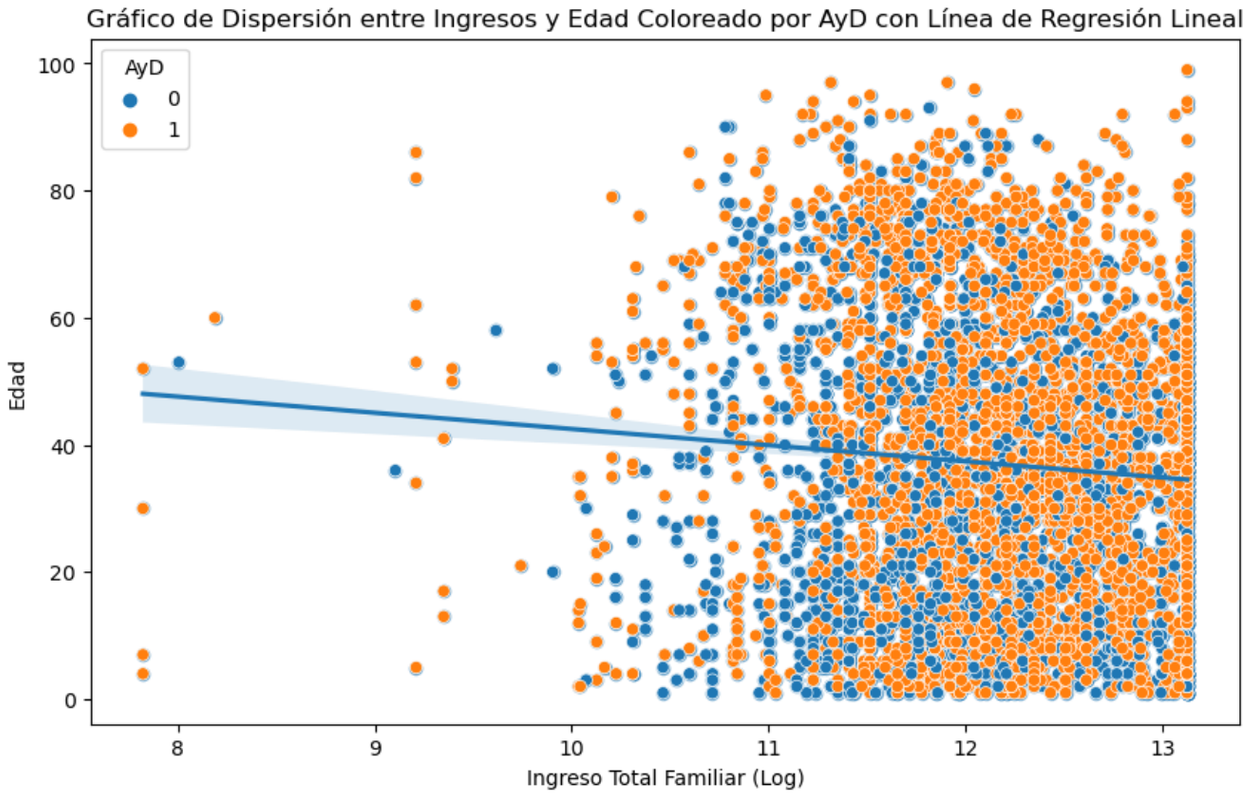
La limpieza y preparación de los datos de la (EPH) resultaron en un conjunto de datos compuesto por 7536 filas y 601 columnas. Esta estructura de datos ofrece una base más sólida para el análisis y modelización subsiguiente.

Posteriormente, se procedió a la creación de una serie de variables adicionales que enriquecen el análisis. Una de ellas fue la fusión de las variables dummies creadas a partir de las categorías 'acceso a agua potable' y 'acceso a desagüe público (cloacas)'. Esta nueva variable compuesta proporciona una visión más integrada del acceso a servicios básicos, lo cual es crucial en el estudio

de las condiciones socioeconómicas. Además, se creó una variable específica para identificar a las personas menores de 15 años que carecen tanto de acceso a agua potable como de servicio de desagüe conectado a la red pública. Esta variable es de particular importancia, ya que pone el foco en un segmento vulnerable de la población y prescinde del reporte de ingresos, muchas veces de carácter sensible.

Un aspecto fundamental del análisis fue la exploración visual de la relación entre el ingreso total familiar (ITF) y la edad de los individuos, mediante un diagrama de dispersión. Este gráfico, que incorpora una diferenciación por color según el acceso a agua potable y a cloacas (AyD), muestra el recorte claro de outliers y revela patrones interesantes. Se observa una correlación lineal leve y decreciente entre ingresos y edad. Explicaciones posibles serían la influencia de grupos de jóvenes con ingresos en dólares, en contraste con una proporción significativa de adultos mayores con ingresos jubilatorios en pesos. O bien el efecto del no reporte de ingresos que podría contar con mayor sesgos en

El análisis cromático e integral del gráfico sugiere la existencia de tres clusters distintivos. Para una completa perspectiva de la relación entre edad e ingresos puede verse el gráfico sin recorte de outliers en Anexo. En el cuadrante inferior izquierdo, se identifica un grupo numeroso de niños y adolescentes que carecen tanto de agua potable como de cloacas, con ingresos familiares entre los más bajos. Por otro lado, en el cuadrante superior derecho, se destaca un conjunto mayoritario de adultos con ingresos superiores al promedio y con acceso a ambos servicios básicos. Sin embargo, esta tendencia no se mantiene en el cuadrante superior izquierdo, donde los adultos con ingresos por debajo del promedio no muestran un patrón definido en cuanto al acceso al agua potable y al desagüe por red pública.



Continuando con el análisis de la Encuesta Permanente de Hogares, el siguiente paso consistió en segmentar la base de datos en dos grupos: los casos que reportaron ingresos y aquellos que no lo hicieron. Esta división resultó en un total de 4171 casos que reportaron ingresos y 3334 casos que no. Esta distinción es crucial para una evaluación precisa de la situación económica de los hogares, especialmente en lo que respecta a la estimación de la pobreza.

Para la base de datos correspondiente a los hogares que reportaron ingresos, se procedió a cal-

cular la proporción de hogares que caen por debajo de la línea de pobreza. Este cálculo se basó en la Canasta Básica Total (CBT), ajustada según la cantidad de personas en cada hogar y sus características específicas. La Canasta Básica Total es un indicador económico fundamental en Argentina que estima el umbral mínimo de ingresos necesarios para satisfacer un conjunto básico de necesidades alimentarias y no alimentarias.

El resultado de este análisis fue la identificación de 1,478,555 hogares como pobres para el Gran Buenos Aires, lo que representa un 28.22% del total. Esta cifra muestra una leve discrepancia con los valores oficiales reportados por el INDEC (30,3%) para el mismo período. Tal diferencia sugiere que el proceso de limpieza y preparación de los datos, aunque necesario para la aplicación de técnicas de Machine Learning, pudo haber introducido modificaciones sutiles en la distribución de los datos comparada con la metodología utilizada por el INDEC.

Parte II: Construcción de funciones

En el proceso de análisis de la Encuesta Permanente de Hogares (EPH) para la construcción de modelos predictivos de pobreza, se implementó una fase crucial de preprocesamiento de datos. Este paso es esencial para asegurar la adecuada preparación del conjunto de datos antes de aplicar técnicas de Machine Learning.

Se inició separando las características (X), es decir, las variables independientes, de la variable dependiente (y), la cual en este contexto es la clasificación de pobreza de los hogares. Este procedimiento implicó la eliminación de la variable 'pobre' del conjunto de datos, seguido de la selección de solo aquellas características numéricas, fundamentales para los algoritmos de aprendizaje automático.

Posteriormente, se abordó el tratamiento de los valores faltantes en las variables independientes. Para ello, se desarrolló una función denominada impute_column, diseñada para imputar valores faltantes de manera diferenciada según la naturaleza de cada variable. En variables con más de dos valores únicos, se reemplazaron los valores faltantes por la media, mientras que en variables con dos o menos valores únicos, se utilizó la mediana. Este enfoque mixto de imputación asegura un tratamiento más preciso y adaptado a la distribución de cada variable.

Una vez tratados los valores faltantes, se aplicó un escalador robusto (RobustScaler) para estandarizar las características. Este escalador es especialmente efectivo en la gestión de outliers, lo cual es vital para evitar distorsiones en los modelos predictivos. La transformación resultante garantiza que todas las variables independientes compartan una escala uniforme, facilitando así su interpretación y análisis en el modelo de Machine Learning.

Finalmente, para mantener la integridad del análisis, se sincronizó la variable dependiente 'pobre' con las filas del conjunto de características escaladas. Esto condujo a la etapa de división de los datos en conjuntos de entrenamiento y prueba, una práctica estándar en la construcción de modelos predictivos para evaluar su desempeño. El conjunto de datos se dividió de manera que el 70% de los datos se utilizó para el entrenamiento y el 30% restante para la prueba, con un total de 2919 observaciones en el conjunto de entrenamiento y 1252 en el conjunto de prueba, tanto para las variables independientes como para la variable dependiente.

Luego se diseñaron dos funciones específicas para evaluar y validar los modelos de Machine Learning aplicados al análisis de la Encuesta Permanente de Hogares (EPH) de Argentina. Estas funciones son cruciales para comparar el rendimiento de diferentes enfoques y elegir el modelo más adecuado para predecir la pobreza.

La primera función, evalua_modelo, se enfoca en la evaluación de una variedad de modelos de clasificación. La función toma como entrada los conjuntos de datos de entrenamiento y prueba, junto con una serie de parámetros específicos (predeterminados por defecto) para cada tipo de mod-

elo. Los modelos considerados incluyen regresión logística, Análisis Discriminante Lineal (LDA), k-Vecinos más cercanos (kNN), Random Forest (RF), Árbol de Decisión, Bagging y Gradient Boosting. La función está diseñada para calcular y retornar un DataFrame con métricas clave de rendimiento como el Área Bajo la Curva ROC (AUC), la precisión (Accuracy), el Error Cuadrático Medio (ECM), y los valores de la matriz de confusión (Verdaderos Positivos, Falsos Positivos, Verdaderos Negativos, Falsos Negativos) para cada modelo. Esto permite una comparación detallada y rigurosa de los modelos en términos de su capacidad para predecir la pobreza.

La segunda función, cross_validation, implementa un proceso de validación cruzada para los mismos modelos de clasificación. Esta función divide el conjunto de datos en k particiones y realiza la evaluación de modelos en cada partición, proporcionando así una evaluación más robusta y menos sesgada del rendimiento del modelo. Este enfoque es esencial para garantizar que los modelos no solo se ajusten bien a un conjunto particular de datos de entrenamiento, sino que también tengan una buena capacidad de generalización en nuevos datos. Es importante destacar que, en el contexto de la validación cruzada, existe un trade-off de sesgo y varianza. Mientras que el método LOOCV puede proporcionar estimaciones casi imparciales del error de prueba, el uso de K-Fold CV, especialmente con un número de particiones como 5 o 10, a menudo resulta en estimaciones más precisas del error de prueba debido a este trade-off. La función cross_validation retorna un DataFrame con las métricas de rendimiento de cada modelo en cada partición, ofreciendo una visión integral del comportamiento del modelo en diferentes segmentos del conjunto de datos. Al utilizar K-Fold CV en la evaluación de modelos para analizar la pobreza en Argentina, es posible alcanzar un equilibrio entre sesgo y varianza, lo cual es crucial para asegurar la validez y la fiabilidad de los resultados. Este equilibrio ayuda a evitar el sobreajuste a un conjunto específico de datos y mejora la capacidad del modelo para generalizar a nuevos conjuntos de datos, lo que es fundamental en aplicaciones de machine learning orientadas a la comprensión y mitigación de la pobreza.

A continuación, la función evalua_config fue diseñada para evaluar una amplia gama de configuraciones de hiperparámetros a través de un proceso de validación cruzada, con el objetivo de optimizar el rendimiento de los modelos de clasificación aplicados al análisis de la Encuesta Permanente de Hogares (EPH).

Esta función acepta como entrada los conjuntos de características (X) y la variable objetivo (y), así como una serie de listas que contienen rangos de valores para diferentes hiperparámetros: lambda_values para la regresión logística, k_neighbors_values para kNN, penalidades para la penalidad en la regresión logística, n_estimators_values para el número de árboles en métodos como Random Forest y Gradient Boosting, max_features_values para el número máximo de características en estos modelos, max_depth_values para la profundidad máxima de los árboles, n bagging estimators values para el número de estimadores en Bagging y learning_rate_values para la tasa de aprendizaje en Gradient Boosting.

El objetivo de evalua_config es identificar la configuración de hiperparámetros que minimiza el error, en este caso, el Error Cuadrático Medio (ECM). Para ello, se realiza por defecto una búsqueda aleatoria a través de 10 configuraciones distintas, seleccionando aleatoriamente valores de los hiperparámetros dentro de los rangos proporcionados. Cada configuración generada se evalúa utilizando la función cross_validation, que implementa validación cruzada en los modelos especificados, con el número de particiones k.

Dentro de este proceso, la función evalúa el ECM promedio para cada método de clasificación. La configuración que logra el menor ECM promedio se convierte en la "mejor configuración". Los resultados óptimos, basados en la aleatorización de valores, fueron:

lambda_ : 1.0,
k_neighbors : 5,
penalidad : 'l2',
n_estimators : 100,
max_features : 'sqrt',
max_depth : None,
n_bagging_estimators : 10,
learning_rate : 0.1

Al final de la ejecución, evalua_config devuelve un tuple que contiene el mejor ECM encontrado y la configuración de hiperparámetros correspondiente, incluyendo información sobre cuál de los métodos de clasificación produjo este resultado. Esta función es esencial para afinar los modelos de Machine Learning, asegurando que se operen con los hiperparámetros más efectivos para el conjunto de datos específico y la tarea de predecir la pobreza en Argentina.

A partir de allí construimos la función evalua_multiples_metodos, que es una extensión del proceso de optimización de hiperparámetros iniciado con evalua_config, diseñada para aplicar este enfoque de manera integral a varios métodos de clasificación.

Esta función integra y amplía la evaluación de diferentes configuraciones de hiperparámetros para una variedad de modelos, incluyendo regresión logística, LDA, kNN, Random Forest, Árbol de Decisión, Bagging y Gradient Boosting. La función acepta rangos de valores para diversos hiperparámetros, como lambda_values, k_neighbors_values, penalidades, entre otros, así como la lista de modelos a evaluar.

El núcleo de evalua_multiples_metodos consiste en utilizar la función evalua_config para identificar la configuración óptima de hiperparámetros y el correspondiente ECM mínimo para el conjunto de modelos especificados. La función evalua_config se ejecuta con los parámetros proporcionados y retorna el mejor ECM y la configuración de hiperparámetros asociada, que juntos representan la combinación más efectiva para la predicción de la variable dependiente en el conjunto de datos.

Posteriormente, evalua_multiples_metodos crea un resumen de los resultados obtenidos, incluyendo la identificación la "Mejor Configuración Global" para cada modelo y, al final, aquel con el ECM mínimo, junto con los detalles de la configuración de hiperparámetros que condujeron a este resultado.

Parte III: Clasificación y regularización

Con la optimización de hiperparámetros y la eliminación de las variables de ingresos, la función `evalua_multiples_metodos` ha logrado una mejora significativa en la predicción en comparación con el Trabajo Práctico 3 (TP3). En el TP3, los modelos LASSO y RIDGE presentaron un Error Cuadrático Medio (ECM) de 0.144 cada uno. Los resultados actuales muestran una notable mejora con Gradient Boosting.

La mejora en el ECM, de 0.144 en LASSO y RIDGE a 0.094 en Gradient Boosting, indica una mayor precisión en la predicción de la pobreza utilizando el conjunto de datos de la EPH sin las variables de ingresos.

Tras seleccionar el modelo óptimo de Gradient Boosting con hiperparámetros específicos, se procedió a su ajuste usando los conjuntos `X_imputed` y `y_cleaned`. Posteriormente, se preparó el conjunto de datos `clean_no_respondieron` para la predicción. El proceso seguido se resume a continuación:

1. Ajuste del modelo Gradient Boosting:

```
gb_model = GradientBoostingClassifier(n_estimators=100, learning_rate=0.1,
max_depth=10, random_state=42)
gb_model.fit(X_imputed, y_cleaned)
```

2. Preparación de `clean_no_respondieron` y predicción de pobreza:
   - Se añadió una columna constante para asegurar la compatibilidad.
   - Se imputaron los datos faltantes tal como antes.
   - Se predijo la probabilidad de pobreza y se calculó la proporción de individuos clasificados como pobres.
3. Resultado:

```
La proporción de hogares pobres predicha es: 33.98%
```

Este resultado muestra una mejora notable en comparación con el Trabajo Práctico 3, acercándose al valor reportado por el INDEC (30.3%). Cabe resaltar que este resultado fue sin un barrido exhaustivo de hiperparámetros y métodos. Esta eficacia refleja una aparente adecuación del modelo de Gradient Boosting y del tratamiento de los datos.

Anexo

Gráfico 1