**Taller 2.**

**Predecir la pobreza.**

* El documento no debe tener más de 10 (diez) páginas e incluir, como máximo, 8 (ocho) pruebas (tablas y/o figuras). La bibliografía y las exhibiciones no cuentan para el límite de páginas.
* Describir en el texto el código utilizado, la forma de replicarlo y todas las tablas e imágenes que se agreguen.
* README

**Integrantes:** Isabella Mendez Pedraza. Cód.: 201814239

Manuela Ojeda Ojeda. Cód.: 201814476

Juan Sebastian Tellez Melo. Cód.: 201513710

Andres Mauricio Palacio Lugo. Cód.: 201618843

**Link del repositorio:** <https://github.com/AndresMPL/Repositorio_PS2.git>

**Modelo predictivo de la pobreza de los hogares.**

I = función de indicador que toma 1 si el ingreso familiar está por debajo de una cierta línea de pobreza.

Predecir la pobreza de dos maneras:

* Problema de clasificación: 0 (no pobre) y 1 (pobre).
* Problema de predicción de ingresos. Con el ingreso previsto, puede usar la línea de pobreza y obtener la clasificación.

1. **Introducción.**

* La introducción expone brevemente el problema y si hay antecedentes (Literatura).
* Describir brevemente los datos y su idoneidad para abordar la cuestión del conjunto de problemas.
* Vista previa de los resultados y las principales conclusiones.

Llegar a entender la pobreza es un reto que ha llamado la atención de diferentes entidades e investigadores. Gracias a que se busca poder medir la efectividad de las diferentes iniciativas que tienen como objetivo combatir la pobreza, para así poder orientar de manera óptima las diferentes intervenciones y políticas que buscan su reducción, maximizando el impacto con el costo más bajo posible.

Para poder medir la pobreza no hay una definición única, existen enfoque monetarios y no monetarios. El primer enfoque, considera que las personas son pobres cuando no tienen suficiente dinero para mantener su sustento. Sin embargo, actualmente existe una discusión

Frente a que la pobreza comprende la falta de oportunidades, educación, atención médica, entre otros. Actualmente, los investigadores concuerdan con que la pobreza es un fenómeno multidimensional que no puede explicarse solo por el dinero (Usmanova, Aziza et al, 2022).

Siguiendo la metodología del Banco Mundial: Pover-T Tests: Predicting Poverty en este documento se busca predecir la pobreza en Colombia. Para esto se utilizaron datos a nivel de hogares y personas, provenientes de la Gran Encuesta Integrada de Hogares (GEIH) y del empalme de las Series de Empleo y Pobreza (MESEP). En primer lugar, se toma como un problema de clasificación para predecir si el hogar es pobre o no pobre dada sus características observables. Posteriormente se toma como un problema de predicción de ingresos, en donde si los ingresos del hogar para posteriormente ubicarlos por encima o por debajo de la línea de pobreza.

Para evaluar el problema de clasificación se utilizaron modelos logit y LDA con ajustes de regularización Lasso, Ridge y Elastic Net. De igual manera, se abordó el problema de balance de clases con metodologías de remuestreo como el Up Samping, Down Samping, y Over samping. Adicionalmente, el problema de clasificación se abordótambien desde los modelos de Arbol. Se busco maximizar la capacidad predictiva de los modelos usando como métrica principal el Accuracy. Sin embargo, no se dejo de lado medidas como Sensitivity y la métrica F1 como indicadores de rendimiento de los modelos calculados.

Para el caso del problema por medio de predicción de ingresos se comparó el RMSE de un molelo de regresión lineal simple con regularización Lasso, Ridge, Elastic Net. También se estimaron diferentes especificaciones de modelos para evitar el sobreajuste y encontrar un equilibrio entre sesgo y varianza. Por otro lado, se calcularon modelos de Arboles, Ramdom Forest y Boosting haciendo tuning de hiperparametros para maximizar la capacidad predictiva. Las medidas de rendimiento en este enfoque son el Accuracy, siendo esta la principal, la Sensitivity y F1.

1. **Datos.**

* Describir la adecuación de los datos para resolver la pregunta predictiva, el proceso de construcción de la muestra, incluyendo cómo se limpiaron y combinaron los datos y cómo se crearon nuevas variables.
* Incluir un análisis descriptivo de los datos. Como mínimo, incluir una tabla de estadísticas descriptivas con su interpretación. Un análisis profundo que ayude al lector a comprender los datos, su variación y la justificación de las elecciones de datos.

Para este ejercicio utilizamos los datos a nivel de hogar y personas provenientes del DANE, del Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE). Estas Bases brindan información sobre características de los hogares como su ubicación, número de personas, numero de cuartos, bajo qué condiciones habitan la vivienda, la suma de los ingresos de todos los integrantes del hogar, entre otros. Adicionalmente, a nivel de personas, ofrece información sobre los integrantes de estos hogares como edad, nivel educativo, sexo, posición en el hogar, actividad o inactividad económica, fuente de ingresos, entre otros.

La base de hogares contiene una variable llamada *Pobre* que identifica los hogares en condiciones de pobreza. Adicionalmente, cuenta con información sobre el límite de ingresos por debajo del cual un hogar es considerado en pobreza (línea de pobreza).

La base de personas contiene la variable *Ingtot* que corresponde al ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas.

Para las predicciones a través de clasificación se utilizó información a nivel hogar, como controles tenemos información acerca del número de cuartos, número de personas en el hogar, edad del jefe de hogar, número de menores de edad en el hogar y de adultos mayores, tipo de vivienda, sexo del jefe de hogar, mayor nivel de educación alcanzado por el jefe de hogar e información sobre si el jefe de hogar está desocupado o inactivo.

Creamos variables de interés como *Genero* que toma el valor de 1 si el individuo es mujer y 0 si es hombre. *Menores\_edad* es una variable que toma el valor de 1 si el individuo es menor a 14 años y *adulto\_mayor* es una variable que toma el valor de 1 si el individuo es mayor a 65 años.

Adicionalmente, dummificamos las variables que nos dicen si los individuos son desempleados, inactivos, ocupados o si está dentro de la población en edad de trabajar.

A partir de la base de personas tomamos información sobre los jefes de hogar como edad, género, nivel educativo y variables relacionadas al nivel del empleo de los individuos pueden llegar a afectar directamente el ingreso de los hogares o determinar si un hogar es pobre o no pobre. A partir de lo anterior unimos estas dos bases.

Adicionalmente, creamos la variable *Numper\_por\_dor* que nos dice el número de personas por cuartos totales en el hogar. Se crea la variable *Hacinamiento* que toma el valor de 1 si el número de personas por cuartos totales en el hogar es mayor a 3, es decir, que si en promedio más de 3 personas se quedan por cuarto en un hogar esta variable toma el valor de 1.

Se crea la variable *Ocupados\_por\_perhog* en el cual si el número de ocupados en el hogar es mayor a 0 se calcula el número de personas en la unidad de gasto sobre el número de ocupados en el hogar.

Posteriormente, vamos a tratar los NAs tanto de la base de personas como en hogares, mantenemos únicamente las variables que tienen un porcentaje de missing menor o igual a 50%.

1. **Modelos y resultados.**

Especificaciones y modelos utilizados para las tareas predictivas.

Subsecciones:

1. Modelos de clasificación. Describir el enfoque de clasificación, es decir, su intento de predecir directamente ceros (no pobre) y unos (pobre).

Para nuestro problema de clasificación: 0 (no pobre) y 1 (pobre) evaluamos modelos Logit y LDA.

1. Modelos de regresión de ingresos. Describir el enfoque de predicción de ingresos, es decir, su intento de predecir primero los ingresos y luego predecir indirectamente ceros (no pobre) y unos (pobre).

Para el caso de predicción de ingresos evaluamos modelos de regresión lineal, Ridge, Lasso, Elastic Net y árboles. Utilizamos como variable de interés el logaritmo de la variable *Ingtotugarr* que es el ingreso total de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios. Para elegir el mejor modelo para nuestras predicciones comparamos el RMSE en los diferentes casos. A partir de estos ingresos y por medio de la línea de pobreza se busca predecir esta pobreza de los hogares.

Dividimos nuestra base en 70% de la muestra para train 15% para test y 15% para eval y realizamos la estandarización correspondiente.

1. Regresión lineal

Estimamos este modelo de regresión lineal a través del método (lm).

Posteriormente corremos otro modelo de regresión lineal incluyendo más controles y algunas interacciones.

1. Ridge

Para estimar el modelo de Ridge usamos el paquete gmlnet con un

1. Lasso

Para estimar el modelo de Lasso usamos el paquete gmlnet con un

1. Elastic Net

Si es Lasso

Si es Ridge

1. Modelo final. Describir los modelos que seleccionó como su presentación final en la competencia. Hasta 2 envíos contarán para la puntuación de la tabla de clasificación. Si se seleccionan menos de 2, Kaggle seleccionará automáticamente entre los envíos con la mejor puntuación. Esta subsección debe incluir:
2. Una explicación detallada de los modelos finales elegidos para la evaluación en Kaggle. La explicación debe incluir cómo se entrenó el modelo, la selección de hiperparámetros y cualquier otra información relevante.
3. Una comparación con al menos otras 2 especificaciones, para cada enfoque.
4. Una descripción de las variables utilizadas en el modelo y discutir su importancia relativa en la predicción.
5. Una descripción de cualquier estrategia de submuestreo utilizada para abordar los desequilibrios de clase.
6. **Conclusiones y recomendaciones.**

Se concluye que los modelos de clasificación que mostraron mejor rendimiento dentro y fuera de muestra fueron los modelos logit con regularización Elastic Net y remuestreo Up Sampling. Mientras que los modelos de regresión con mejor desempeño prediciendo ingresos para posteriormente clasificar los hogares entre pobre y no pobre, fueron las regresiones lineales con regularización Lasso. Adicionalmente, desde los modelos de Árboles se encontró que la variable más importante para predecir pobreza son el número de cuartos de la vivienda, relacionado con el hacinamiento; la relación entre número de ocupados en el hogar y el número total de personas en la unidad de gasto; y el nivel educativo del jefe del hogar siendo la educación superior la más significativa en diferencia de ingresos. Adicionalmente, la variable de sexo del jefe del hogar no apareció como una variable contundente al momento de predecir pobreza.

1. **Bibliografía.**

Usmanova, Aziza et al. “Utilities of Artificial Intelligence in Poverty Prediction: A Review.” *Sustainability (Basel, Switzerland)* 14.21 (2022): 14238–. Web.

1. **Apéndice.**