**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 su 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, etc. 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 del DANE y del Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE). En primer lugar, se toma como un problema de clasificación para predecir si el hogar es pobre o no pobre. Posteriormente se toma como un problema de predicción de ingresos, en donde si los ingresos del hogar son menores a la línea de pobreza se considera que el hogar es pobre.

Para evaluar el problema de clasificación se utilizaron modelos logit y LDA xxxxxx. Para el caso del problema por medio de predicción de ingresos se comparó el RMSE de un moldeo de regresión lineal simple, Lasso, Ridge y Elastic Net. Se encontró que el modelo que menor predice la pobreza en nuestro caso es Lasso.

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 nivel de hogares y personas provenientes del DANE, del Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE).

La base de hogares contiene una variable llamada *Pobre* que identifica los hogares en condiciones de pobreza. Adicionalmente, cuenta con información sobre el valor de la canasta básica de bienes que establece el límite de ingresos por debajo del cual un hogar es considerado en pobreza.

La base de hogares 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

Para las predicciones a través de la predicción de ingresos utilizamos como controles

Para nuestras predcicciones a nivel de personas utilizamos variables relacionados con si los individuos viven en un cabecera municipal, sexo, edad, parentesco con el jefe de hogar, seguridad social en salud, régimen del sistema de seguridad social en salud, nivel educativo, si recibió o ganó el mes pasado ingresos por concepto de trabajo (desocupados), si recibió o ganó el mes pasado ingresos por concepto de trabajo (Desocupados), ¿El mes pasado recibió pagos por b. pensiones o jubilaciones por vejez,invalidez o sustitución pensional ?, ¿El mes pasado recibió pagos por c. pensión alimenticia por paternidad, divorcio o separación?, recibió dinero de otros hogares, personas o instituciones no gubernamentales; dinero por intereses, dividendos, utilidades o por cesantias?, Durante los últimos 12 meses, ¿recibió a. dinero de otros hogares o personas residentes en el país?, Durante los últimos 12 meses, ¿recibió b. dinero de otros hogares o personas residentes fuera del país?, Durante los últimos 12 meses, ¿recibió c. ayudas en dinero de instituciones del país?, Durante los últimos 12 meses, ¿recibió d. dinero por intereses de prestamos o CDT´s, depositos de ahorros, utilidades, ganancias o dividendos por inversiones?, Durante los últimos 12 meses, ¿recibió e. dinero por concepto de cesantías y/o intereses a las cesantías?, Durante los últimos 12 meses, ¿recibió f. dinero de otras fuentes diferentes a las anteriores?, si es una persona en edad de trabajar, si esta ocupada, si esta desocupada, si esta inactivo.

A nivel hogar utilizamos variables con información si el hogar vive en cabecera municipal, de cuántos cuartos en total dispone el hogar, en cuántos de esos cuartos duermen las personas de este hogar, el tipo de vivienda del hogar, el número de personas en el hogar, número de personas en la unidad de gasto

Creamos variables de interés

Creamos la variable *Genero* que toma el valor de 1 si el individuo es mujer y 0 si es hombre, este mismo tratamiento se realiza para la variable *Clase* que nos dice si los individuos viven en un cabecera municipal. *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.

Mantenemos la información de los jefes de hogar.

Unimos la base de hogares con la base de personas con los ajustes antes mencionados.

Posteriormente, modificamos la variable *jefe\_hogar\_ina*  para tomar las observacion de 11 años que no clasifica como inactivo, desempleado, ocupado se asigna como inactiva

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).
2. 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).
3. 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:
4. 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.
5. Una comparación con al menos otras 2 especificaciones, para cada enfoque.
6. Una descripción de las variables utilizadas en el modelo y discutir su importancia relativa en la predicción.
7. Una descripción de cualquier estrategia de submuestreo utilizada para abordar los desequilibrios de clase.
8. **Conclusiones y recomendaciones.**

Principales conclusiones del trabajo.

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.**