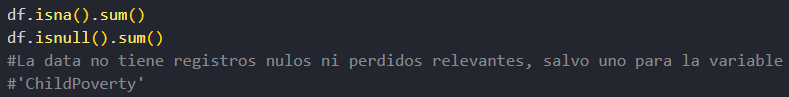
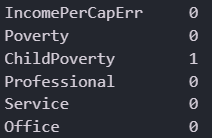
**Preparación de la Data**

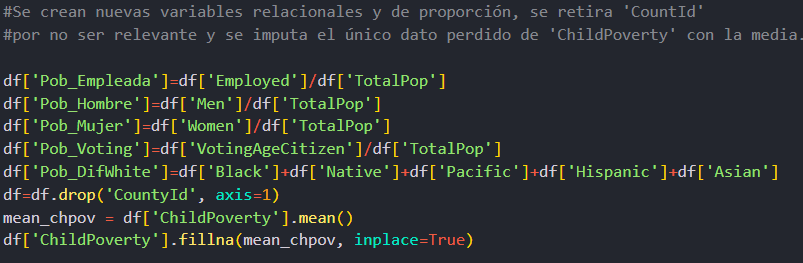
La data del proyecto final se refiere al censo demográfico de los Estados Unidos para el año 2017, la cual reúne una serie de características socioeconómicas de la población, entre las que se encuentra su pertenencia a grupos raciales como: *hispanos, negros, nativos americanos, asiáticos, del pacífico y blancos.*

El proyecto se realiza en Python, específicamente con Visual Studio Code, siendo el paso previo la identificación de data faltante con el siguiente comando:

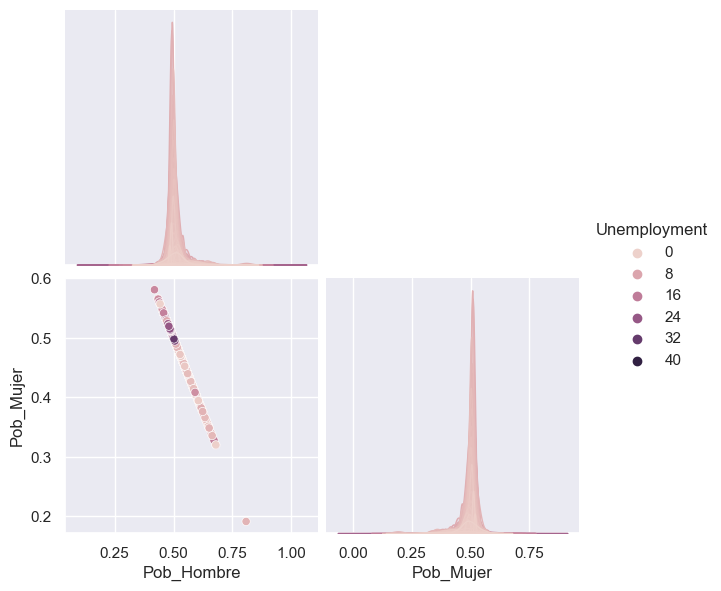
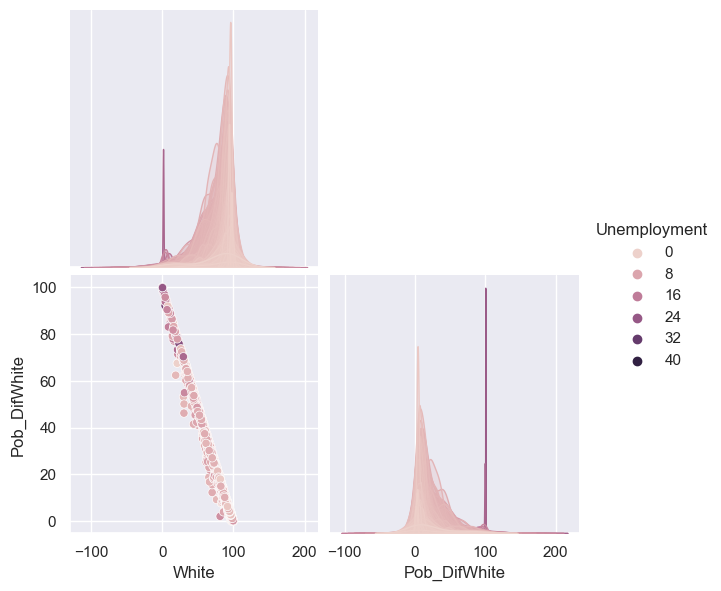


El resultado muestra que la data es en general adecuada y con poca información faltante, el único campo con un solo registro faltante fue ‘ChildPoverty’, para el cual se imputó el dato con la media de los valores que sí están presentes.

Adicionalmente, se definieron nuevas variables o campos con la información actual, cuya finalidad fue analizar la proporción de las distintas poblaciones, como, por ejemplo: población empleada, proporción de hombres y mujeres, en edad de votar y se consolidó toda la población considerada como minoría en una sola variable llamada ‘Pob\_DifWhite’, que es la sumatoria de los campos que describen la raza de las personas, excluyendo a la población blanca.



Antes de modelar la data, se hicieron algunos gráficos para entender si, de manera preliminar, la hipótesis de la investigación es afirmativa, es decir, que pertenecer a una minoría racial influye en la situación socioeconómica de las personas.

Los gráficos anteriores muestran que la tasa de desempleo es mayor (colores más oscuros) en la población con menor población blanca. Por su parte, si bien no es el objetivo de la investigación, ser mujer también tiene una incidencia en cuanto a mayores tasas de desempleo que las que presentan los hombres. En ese sentido, de manera intuitiva, se empieza a visualizar de forma preliminar que pertenecer a una minoría, por ejemplo, racial o sexual, siendo mujer, repercute en los indicadores económicos y sociales.

**Modelado con Aprendizaje Automático**

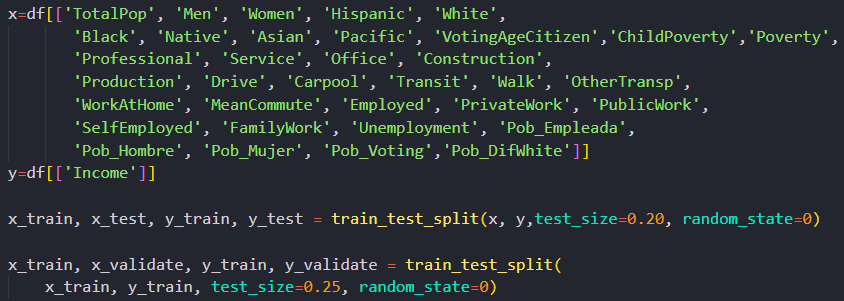
Para responder la pregunta de investigación se pueden tener dos enfoques al modelar la información: el supervisado y no supervisado.

Frente al primer enfoque, se busca determinar o proyectar los ingresos de una persona en función de las variables independientes que componen el set de datos. Así, el primer modelo trabajado es una regresión lineal.

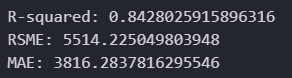
**Regresión Lineal Múltiple**

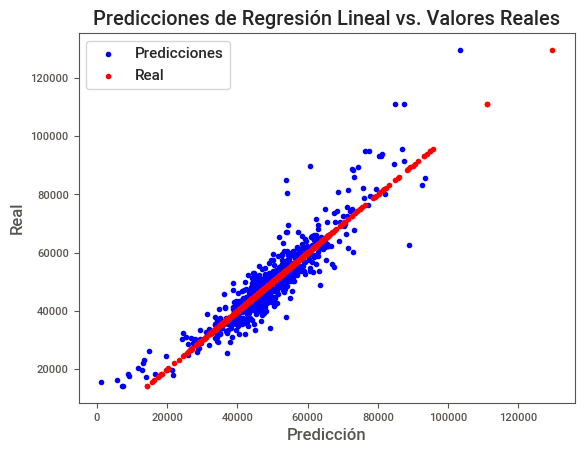
El primer paso es preparar la información, mediante la separación de la variable dependiente o de respuesta, del resto de variables explicativas, las cuales también se dejan aparte en un data frame distinto.

Posteriormente, la data es segmentada en bases de entrenamiento, testeo y validación, las cuales servirán para determinar que comportamiento del modelo con nueva data ingresada.



Posteriormente el modelo surte los siguientes pasos: i) Llamada al modelo de regresión lineal, ii) Entrenamiento con la data de entrenamiento o *train* y iii) predicción o ajuste del modelo, para lo cual se emplea la data de validación o *validate.*

Una vez se ejecuta el modelo, se plantean los principales indicadores de validación, que dan razón de un buen modelo para predecir ingresos, lo anterior dado el desempeño de R2 (84.2%) y los errores, que para el caso del error medio absoluto (MAE) es de 3.816 dólares.



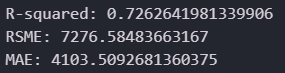
La mirada gráfica del modelo, confirma los resultados que ya se comentaron, los valores reales del ingreso de las personas (línea roja), cruzan o cortan de manera generalizada los datos predichos por el modelo (puntos azules). De igual modo, se puede apreciar algún tipo de curva en los datos predichos en los ingresos más bajos, lo que nos llevó a considerar el modelamiento de una *Regresión Polinómica Múltiple.*

En términos de la *media de los ingresos*, el modelo promedia ingresos de USD 48.977 frente a USD 48.995 de la data real, muy cercanos entre sí.

**Regresión Polinómica Múltiple**

Buscando la optimización del modelo y la aproximación más fiel a la data, se definió un modelo de regresión polinomial, el cual permite modelar una relación que no es del todo lineal entre las variables, tanto de las explicativas como de la variable de respuesta que se busca pronosticar.

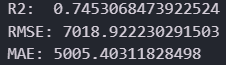
En primera medida, se debe ajustar la data original para que reúna las características polinomiales, para este caso, de grado 2; dicha transformación permitirá incorporar características como el término constante, el término lineal y término polinómico o cuadrático. Esta transformación se hace necesaria cuando los datos parecen no ser del todo lineales, por lo que ajustar una línea recta (modelo de regresión lineal) puede que no sea el modelo óptimo, en contraste, la línea polinómica permite una aproximación más cercana a la realidad de la data. Frente a dicho modelo, los resultados son los siguientes:

Como se aprecia, frente a los resultados de la regresión lineal simple, la regresión polinómica no obtiene un mejor ajuste para los datos, dado que el R cuadrado es menor (73%) y el error absoluto medio (4.103 dólares de ingreso) fue superior en 288 dólares de ingreso, por lo cual, este modelo no se considera como alternativa.

**Árbol de Decisión Regresor**

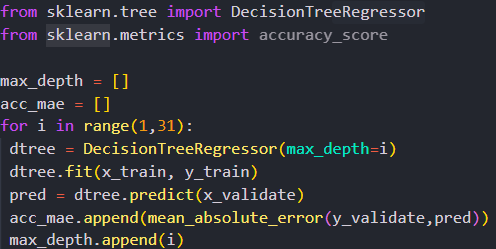
La siguiente aproximación para un modelo que pronostique los ingresos de la población de los Estados Unidos es un árbol de decisión. En este modelo el reto es distinto, pues se buscará hacer una optimización que permita definir el número de ramas optimo que disminuya el error de los modelos previos.

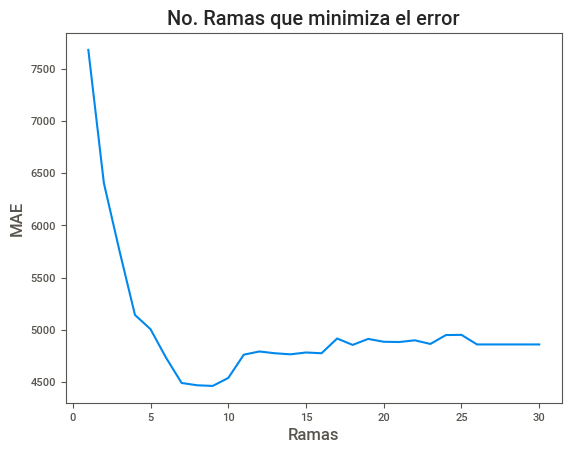
En primera instancia, se desarrolló un modelo para el cual se definió un árbol de cinco ramas, para el cual se determinaron los mismos indicadores de validación que para los modelos de regresión lineal y polinomial:

Si se tienen en cuenta los resultados de los modelos previos, particularmente de la regresión lineal, se concluye que el árbol de decisión sin optimización de ramas, no obtiene mejores métricas, por lo cual vamos a validar los resultados de optimizarlo.

**Optimización del Árbol de Decisión**

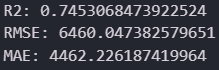
La optimización consiste en buscar el número de ramas que minimizan, en este caso, el error medio absoluto, con las ramas que se obtengan de este proceso, se ejecuta un nuevo árbol de decisión y poder comparar sus métricas con las de los modelos previos.

En primera medida, se crean dos vectores vacíos (para almacenar el No. de ramas y los errores del modelo) en los cuales se va a almacenar el resultado de las iteraciones de la función. En cuanto a la función, se definieron 31 experimentos, esperando fueran suficientes para visualizar gráficamente el número de ramas óptimo. Posteriormente el modelo se entrena y sus resultados aplicados a la data de validación, cuyos errores se almacenan y se van acumulando en uno de los vectores creados inicialmente. El resultado de dichas iteraciones se grafica para determinar el número de ramas que minimiza el error del modelo, según se ilustra:

El gráfico muestra que el error se minimiza en nueve ramas, obteniendo en este nivel un MAE de 4.462, el cual es inferior al error de la octava y décima rama, de 4.468 y 4.538 respectivamente.

**9**

Al ajustar el nuevo modelo con estas nueve ramas optimizadas, los indicadores de ajuste de modelo muestran los siguientes resultados:

El nuevo modelo optimizado reduce el MAE en un 10,8%, por lo que se confirma la eficacia del proceso de optimización, no obstante, ninguno de los modelos de árbol de decisión logra mejorar el ajuste obtenido por el modelo de regresión lineal.

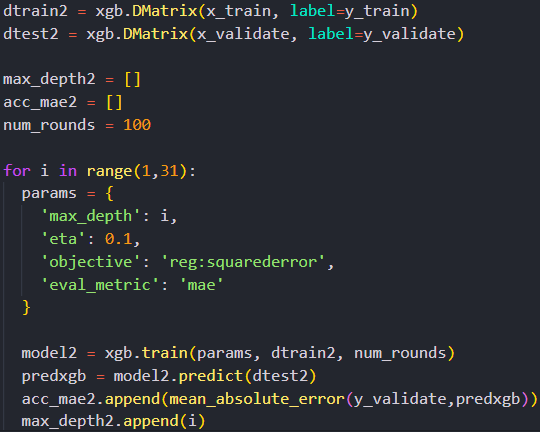
**XGBoost**

Como ultimo modelo evaluado, se consideró el *extreme gradient boosting* el cual, en términos generales, es una actualización y mejora del modelo de arboles de decisión, pudiendo resumirse como un modelo que itera de manera secuencial varios arboles de decisión que pretenden corregir los errores de los árboles generados previamente.

Previo a su ejecución, este modelo requiere la transformación de la data de entrada en una forma matricial, tanto de la data de entrenamiento como de la data de validación, esto mediante la función DMatrix. Posteriormente, se deben fijar los parámetros principales del modelo, como i) la función de perdida *(objective)* o, en otras palabras, la función que busca la minimización del error que se definió en ‘*squarederror’* ii) el parámetro *‘eta’* que asigna el peso o participación que se le asigna a cada árbol de decisión que itera el modelo; se deja el valor por defecto en 0.1, para evitar cualquier posibilidad de sobreajuste, iii) *‘eval\_metric’*, definida en *‘MAE’*, es decir, que el error medio absoluto será la medida para medir el desempeño del modelo.

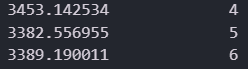
Definidos estos pasos previos, se sigue la misma dinámica comentada en el apartado de árbol de decisión, en ese sentido, abordaremos de manera directa la optimización realizada al modelo, cuyo parámetro de ajuste fue la profundidad de las ramas, la cual mostró como resultado cinco (5), que se definieron en la entrada de *‘max\_depth’*.

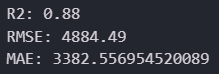
Definición de las matrices de entrada



Entrenamiento y validación del modelo optimizado

Número de ramas para optimizar

Del desarrollo anterior, se aprecia que el MAE es inferior cuando las ramas de los arboles iterados por el algoritmo XGboost se definen en 5, pues es el menor error obtenido entre los errores de las ramas que le anteceden y suceden (4 y 6 respectivamente).

Al ejecutar el modelo XGboost, definiendo el parámetro del número de ramas o ‘max\_depth’ en cinco (5), de acuerdo con la optimización, se muestran mejores resultados que los obtenidos en la regresión lineal, con un mejor R Cuadrado (88% frente a 84,2%) y un mejor MAE (3.383 frente a 3.816).

De lo anterior y como conclusión, se opta por el modelo *XGBoost* *optimizado* para el propósito de proyectar los ingresos de los estadounidenses, en función de las variables sociodemográficas y étnicas que componen la data original.