Análisis de ingresos de personas en CABA:

Período 2019-2020

Aspro, Santiago – Grosclaude, Martín

Introducción:

¿Cuánto son los ingresos promedios de una persona en CABA? ¿El ingreso laboral, varía mucho en función del sexo de la persona? ¿En qué edad aproximadamente alcanzamos el valor máximo de nuestro salario? ¿Es real que hay diferencia entre hombres y mujeres? Todas estas preguntas fue la que nos hicimos a la hora de encarar el proyecto. En primer lugar, realizamos un análisis exploratorio de datos para responder alguna de las preguntas citadas al comienzo, luego, mediante distintos algoritmos de machine learning, buscamos poder estimar los ingresos laborales promedios en función de distintos parámetros.

Descripción de los datasets:

Para llevar adelante nuestro análisis utilizamos un dataset que mostraba una encuesta realizada por el Gobierno de la Ciudad de Buenos Aires en el año 2019. El dataset presenta varias columnas (*features*) y un gran número de personas encuestadas (*samples*). A continuación, mostraremos cuales eran los datos originales de nuestro dataset:

Titulo de la columna	Descripción	Titulo de la columna	Descripción
id	Clave que identifica a la vivienda	ingreso_total_no_lab	Monto del ingreso total no laboral percibido el mes anterior
nhogar	La variable id + nhogar componen la clave que identifica a cada hogar	calidad_ingresos_totales	Calidad de ingresos totales individuales
miembro	Variables id + nhogar+ miembro componen la clave que identifica a cada persona	ingresos_totales	Monto del ingreso total individual percibido el mes anterior
comuna	Comuna donde reside la persona encuestada	calidad_ingresos_familiares	Calidad de ingresos totales familiares
dominio	Variable categórica que indica si la vivienda se ubica en una villa de emergencia	ingresos_familiares	Monto de ingresos totales familiares percibido el mes anterior
edad	Edad de la persona encuestada	ingreso_per_capita_familiar	Monto de ingresos totales familiares percibido el mes anterior
sexo	Sexo de la persona encuestada	estado_educativo	Asistencia (pasada o presente) o no a algún establecimiento educativo
parentesco_jefe	Variable categórica que indica la relación de parentesco entre la persona encuestada y el jefe/a de	sector_educativo	Sector al que pertenece el establecimiento educativo al que asiste
situacion_conyugal	Situación conyugal de la persona encuestada	nivel_actual	Nivel cursado al momento de la encuesta
num_miembro_padre	Número de miembro que corresponde al padre	nivel_max_educativo	Máximo nivel educativo que se cursó
num_miembro_madre	Número de miembro que corresponde a la madre	años_escolaridad	Años de escolaridad alcanzados
estado_ocupacional	Situación ocupacional de la persona encuestada	lugar_nacimiento	Lugar de nacimiento de la persona encuestada
cat_ocupacional	Categoría ocupacional de la persona encuestada	afiliacion_salud	Afiliación de salud de la persona encuestada
calidad_ingresos_lab	Calidad de la declaración de ingresos laborales totales	hijos_nacidos_vivos	Tiene o tuvo hijos nacidos vivos
ingreso_total_lab	Monto del ingreso total laboral percibido el mes anterior	cantidad_hijos_nac_vivos	Cantidad de hijos nacidos vivos
calidad_ingresos_no_lab	Calidad de la declaración de ingresos no laborales totales		

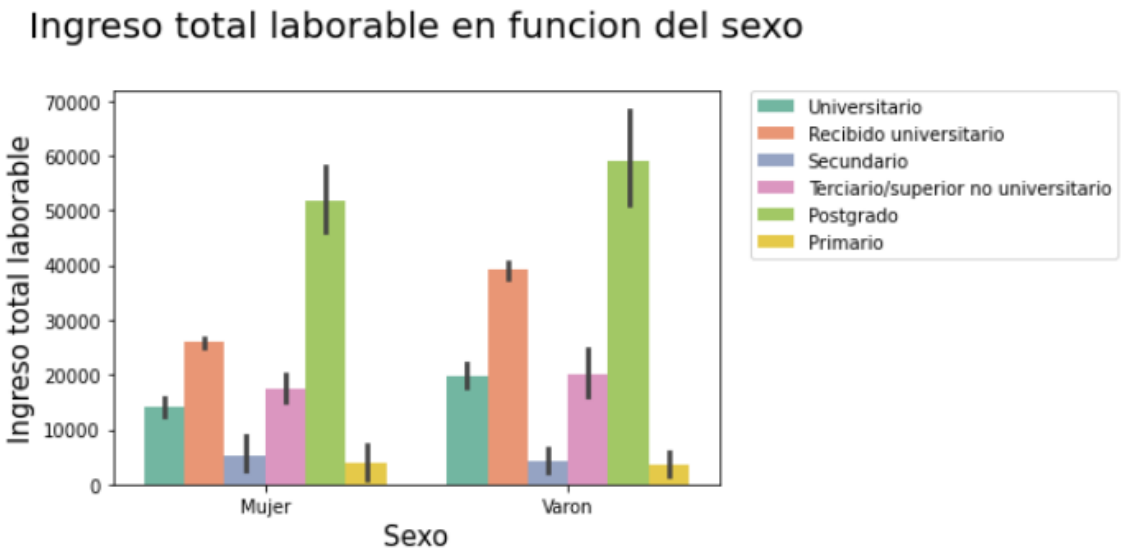
Análisis exploratorio de datos (EDA):

En primer lugar, como parte del EDA y del pre-procesamiento de los datos, eliminamos columnas que no aportaban a nuestro análisis. Las columnas eliminadas fueron:

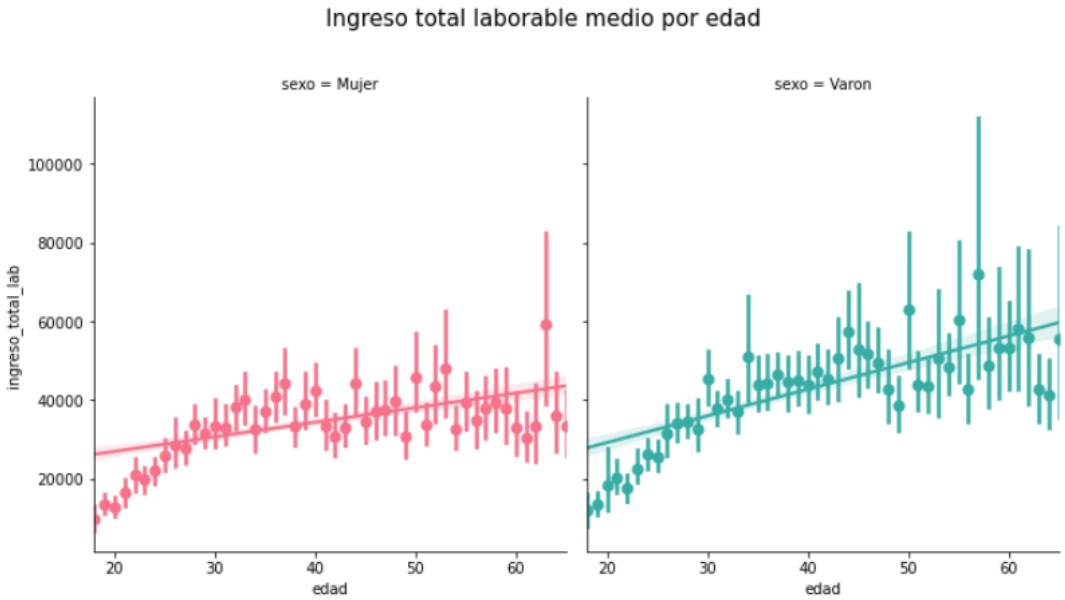
- 'num_miembro_padre'
- 'num_miembro_madre'
- 'calidad_ingresos_no_lab'
- 'ingreso_total_no_lab'
- 'calidad_ingresos_totales'
- 'ingresos_totales'
- 'calidad_ingresos_familiares'
- 'ingresos_familiares'
- 'ingreso_per_capita_familiar'
- 'nivel_max_educativo'
- 'hijos_nacidos_vivos'
- 'calidad_ingresos_lab'
- 'estado_educativo'
- 'sector_educativo'

Luego, aplicando distintas herramientas, cambiamos la estructura de varias features para poder analizarlas de una manera más sencilla. También, reducimos nuestro dataset, al rango de edad laborable de las personas (18-65).

Ahora sí, iniciando con el análisis del dataset, graficamos el ingreso total laborable en función del sexo y el nivel educativo actual de las personas:



Continuando con el análisis, realizamos un gráfico donde se estima, la media del ingreso total laborable de las personas, según la edad y el sexo. Cada barra indica: el salario máximo, mínimo y promedio para cada edad.

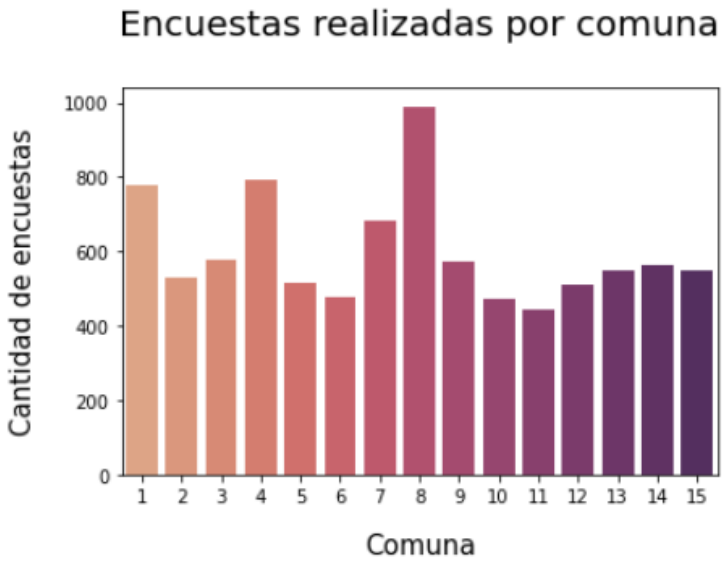


En este grafico pudimos observar la gran diferencia en los salarios promedios de las personas según el sexo. A su vez, se puede observar que la pendiente de crecimiento del salario para los hombres, es mucho mayor que para las mujeres.

En los gráficos anteriores hicimos enfoque a los ingresos de las personas en función a sus niveles educativos y a su sexo. Ahora haremos enfoque en conocer cuál es la comuna donde residen las personas con mayores ingresos. Antes que nada, detallaremos los barrios a los cuales corresponde cada comuna y su respectiva numeración:

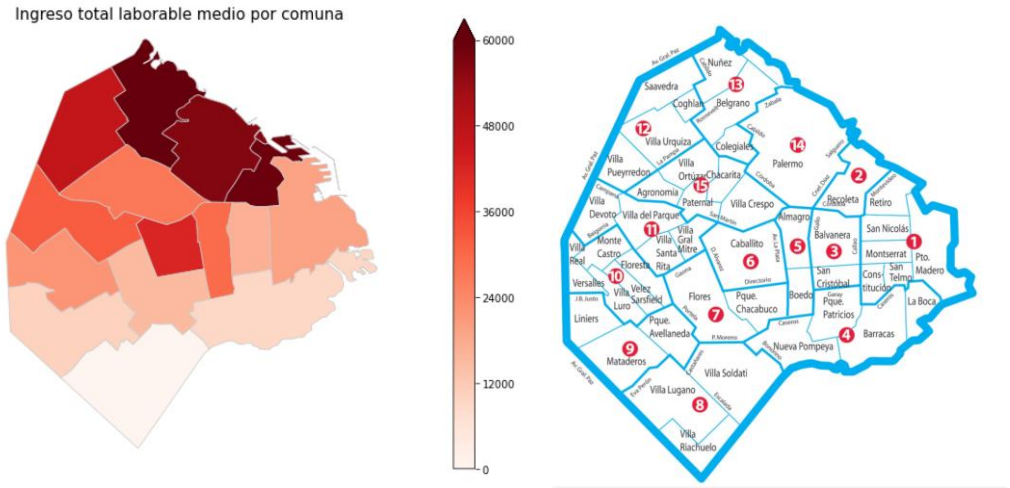
1. Retiro, San Nicolás, Montserrat, Constitución, San Telmo y Puerto Madero.
2. Recoleta.
3. Balvanera y San Cristóbal.
4. La Boca, Barracas, Parque patricios y Nueva Pompeya.
5. Almagro y Boedo.
6. Caballito.
7. Flores y Parque Chacabuco.
8. Villa lugano, Villa Soldati y Villa Riachuelo.
9. Mataderos, Parque Avellaneda y Liniers.
10. Villa Luro, Velez Sarsfield, Floresta, Monte Castro, Villa Real y Versalles.
11. Villa Devoto, Villa del Parque, Villa Santa Rita, Villa Gral. Mitre.
12. Villa Pueyrredón, Villa Urquiza, Coghlan y Saavedra.
13. Núñez y Belgrano
14. Palermo
15. Villa Ortúzar, Agronomía, Paternal, Chacarita y Villa Crespo

Primero observaremos la cantidad de encuestas realizadas por comunas:



Podemos observar que la Comuna 8 fue donde se realizaron más encuestas y que la comuna 11 fue donde menos encuestas se realizaron.

Por último, les mostramos un gráfico, donde se puede visualizar cuál es la comuna donde vive la gente de mayores ingresos y la comuna donde reside las personas de menores ingresos.



Las 3 comunas que presentan mayores ingresos laborales promedios son la 13 (Belgrano, Núñez y Colegiales), la 14 (Palermo) y la 2 (Recoleta).

Variables para analizar los modelos de regresión:

Inicialmente aplicamos LabelEncoder a las variables categóricas. Esta función codifica *features* con valores *string* en variables numéricas.

Luego separamos nuestros datos en datos de entrenamiento y testeo y los escalamos utilizando StandardScaler. Esta función normaliza nuestros datos, aplicándole media=0 y desvió estándar = 1.

Explicaremos un poco los valores que nos interesaran obtener de los modelos:

- **MSE:** Mean Squared Error (Error cuadrático medio). Este valore es sensible a predicciones muy malas. Puede ser problemático en datos ruidosos.
- **RMSE:** Root Mean Squared Error (Raíz cuadrada del error cuadrático medio). [2]
- **MAE:** Mean Average Error (Media del error) Es mejor cuando no queremos penalizar fuerte grandes errores (Por ejemplo: Outliers). [2]
- **R^2:** Esta variable explica la proporción de la varianza de "Y" que explica el modelo de regresión.

Modelos de regresión utilizados:

Regresión Lineal [1][i]:

Es un modelo matemático que se utiliza para aproximar la relación entre una (o varias) variable independiente con una variable dependiente. Este modelo nos entregó los siguientes valores de variables:

R^2	MSE	MAE	RMSE
0.447168	0.094673	0.219484	0.30768978

Ridge Regression:

Es uno de los modelos de regresión más utilizados en machine learning. Estos modelos se utilizan en problemas que no tienen una única solución. Este modelo nos entregó los siguientes valores de variables:

R^2	MSE	MAE	RMSE
0.427556	0.098032	0.224521	0.31310062

Support Vector Regression [3]:

Este modelo de regresión busca maximizar el margen. Construye una función lineal. Determina un margen/radio como función de costo y trata de que todas las muestras caigan dentro del margen (o tubo).

R^2	MSE	MAE	RMSE
0.447168	0.094673	0.219484	0.30768978

KNN Regression:

En el entrenamiento se determinan los más puntos más cercanos por distancia euclidea, es decir, la distancia par-a-par. El valor a predecir se determina por la interpolación de los Y. En este modelo yo no busco aprender un parámetro W y siempre es necesario disponer del dataset para poder hacer la regresión.

R^2	MSE	MAE	RMSE
0.357211	0.110078	0.235997	0.33178005

Regression tree:

Este modelo, como su nombre lo indica, es un tipo de aprendizaje supervisado que utiliza un árbol de decisión para predecir. Es uno de los enfoques de modelado predictivo utilizado en estadísticas, minería de datos y aprendizaje automático.

R^2	MSE	MAE	RMSE
0.402816	0.102268	0.229806	0.31979368

Cuadro final:

	R^2	MSE	MAE	RMSE
Linear	0.427236	0.098086	0.224629	0.31318685
Ridge	0.427556	0.098032	0.224521	0.31310062
SVR	0.447168	0.094673	0.219484	0.30768978
KNN	0.357211	0.110078	0.235997	0.33178005
Tree	0.402816	0.102268	0.229806	0.31979368

En función de los datos citados en el cuadro anterior, determinamos que el modelo que más se ajusta a los datos es Support Vector Regression.

Por último, realizamos un ScatterPlot para poder visualizar la diferencia entre los valores predichos por nuestro modelo y los valores reales del dataset y también hicimos un Displot para ver la distribución de los residuos obtenidos.

Conclusión:

Por medio del análisis de nuestros datos pudimos encontrar comportamientos que normalmente se suponen en la actualidad referidos al ingreso, como, por ejemplo, que una persona recibida de la universidad gana más que una que no terminó la secundaria, que el ingreso en una mujer tiende a ser más bajo que el de un hombre y que el ingreso aumenta en función de la edad hasta que llega a una meseta y luego a un declive.

Para llegar a estas conclusiones realizamos una limpieza y estandarización de los datos, ya que había features que no aportaban información útil para el proyecto y otros que tenían muchas categorías que podían resumirse en categorías binarias.

Luego, avanzando con el EDA, empezamos a ver como se comportaban las diferente features en relación con el ingreso total laborable, sumando también la diferenciación entre hombre y mujer. Para esto hicimos una subdivisión del dataset original, seleccionando a las personas entre 18 a 65 años, es decir, aquellas que podrían llegar a estar laboralmente activas. A su vez, para fines de poder realizar ciertas tareas posteriores, incluimos sólo los casos donde el ingreso total laborable sea mayor a 0. También, agregando la combinación de ingreso total laborable medio en las diferentes comunas, pudimos ver que hay una división horizontal, aproximadamente en la mitad de la ciudad, donde las comunas del norte de la Ciudad tienen un ingreso medio mucho mayor a las del sur, donde se encuentran los barrios de menores ingresos medios por persona.

Pasando a la aplicación de machine learning, incorporamos varios modelos de regresión para tratar de encontrar un regresor que pueda predecir de forma precisa el ingreso total laborable basándose en todos los datos. Logramos encontrar que el SVR de Kernel Gaussiano que fue el que mejor resultados obtuvo, a pesar de su alto costo computacional. A su vez, estudiando la distribución de las predicciones, observamos que con valores bajos de ingreso el regresor tiene mayor error.

[1] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons. 1

[2] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

[3] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. Advances in neural information processing systems, 9, 155-161.

Bibliografía:

El dataset fue obtenido de: https://data.buenosaires.gob.ar/dataset/encuesta-anual-hogares