Analisis de ingresos en CABA

Grosclaude Martin , Aspro Santiago

Universidad Tecnológica Nacional, Facultad Regional Buenos Aires

Introduction

La idea del Proyecto es poder estimar el ingreso promedio de las personas en función de distintas variables, como por ejemplo, el sexo, la edad, el nivel educativo actual, entre otras. Mientras realizábamos el análisis exploratorio de datos fuimos descubriendo varias cuestiones interesantes.

Datasets

El dataset utilizado proviene de una encuesta a hogares realizada por el Gobierno de la Ciudad de Buenos Aires. Las encuestas fueron realizadas en el año 2019. El numero total de personas encuestadas es de 14.319

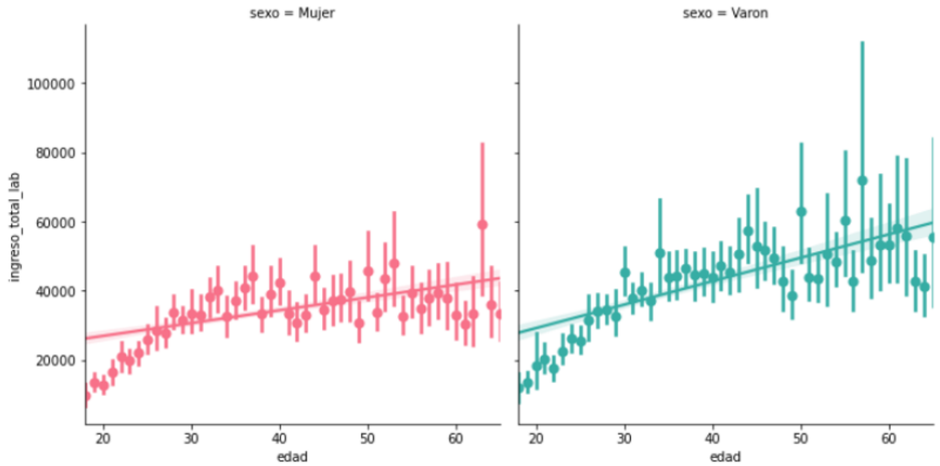
Métodos

Los modelos utilizados para la regresion fueron los siguientes:

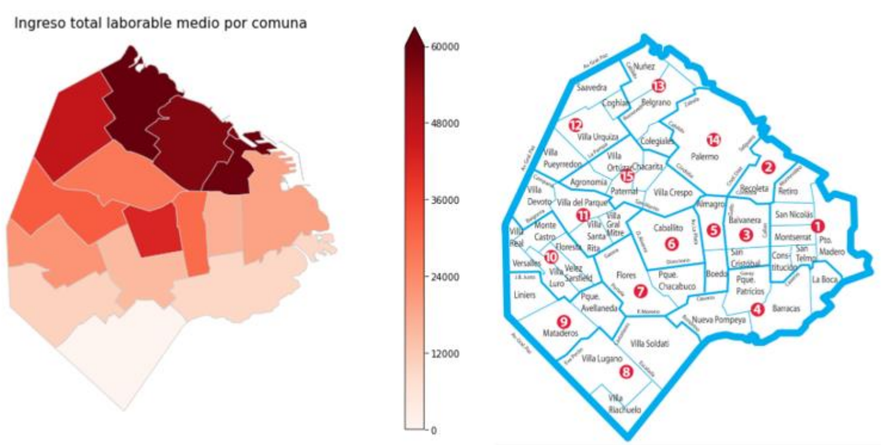
- Linear Regression.
- Ridge Regression.
- KNN Regression
- Regression Tree
- Support Vector Regression

Analisis Exploratorio de datos

Dentro de los datos mas relevantes obtenidos se desprenden la comparación de los sueldos entre hombres y mujeres y a su vez la evolución en el tiempo del salario.



Por otro lado, determinamos cuales son las comunas donde residen las personas de mayores ingresos.



Resultados

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

En función de los resultados antes mencionados, el modelo elegido es el Support Vector Regression.

Conclusiones

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 otros casos también interesantes.
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 fue el que mejor resultados obtuvo, a pesar de su alto costo computacional.