Maestría Economía Aplicada

MECA 4107: Big Data and Machine Learning for Applied Economics

Presentado por el grupo 5: Carlos Avilán y Francisco Ortiz

Abstract

Conocer variables económicas como el Producto Interno Bruto (PIB) en los municipios de Colombia ha sido una limitante para desarrollar nuestro trabajo de grado *"Evaluación impacto del Proyecto Nacional Conectividad de Alta Velocidad -PNCAV*", ya que las instituciones de estadística como el DANE presentan esta información solamente a nivel departamental. Por ello, en el presente trabajo desarrollamos un modelo utilizando Big Data y Machine Learning para predecir esta información a nivel municipal para los años 2010 a 2020. Para el desarrollo de esta predicción se utilizaron datos de importancia como la intensidad lumínica nocturna, la población de los municipios, la distancia a la capital, entre otros. Se realizaron diferentes pruebas para determinar que el mejor modelo fue el estimado con el hiperparámetro Ridge, con el cual se realizaron las predicciones, se analizaron los resultados y finalmente se concluyó.

Introducción

En nuestro trabajo de grado pretendemos evaluar el impacto del Proyecto Nacional Conectividad de Alta Velocidad -PNCAV el cual nace en el 2013 en el marco del Plan Vive Digital con el objetivo de impulsar la masificación del Internet de alta velocidad, y alcanzar una cobertura del 100% para el 2023 en algunos municipios de las regiones de la Amazonía, Orinoquía y Chocó. Este despliegue de infraestructura pretende generar un desarrollo regional aumentando las oportunidades de transformación socioeconómica.

En el mundo existen estudios del impacto positivo en la economía a causa de la masificación del Internet banda ancha. Sin embargo, en Colombia no existen estudios recientes que puedan determinar el efecto que tienen esta clase de políticas y en particular del PNCAV. Ante esto, nuestro trabajo de grado tiene como fin responder la siguiente pregunta: ¿Cuál es el efecto del aumento de la penetración de Internet banda ancha en el Producto Interno Bruto (PIB) en estos municipios?

Para poder determinar este impacto de la conectividad de internet en la economía es necesario conocer el PIB de los municipios de Colombia. Sin embargo, esta información solamente existe del año 2005 al 2009 (datos CEDE). Para ello en este trabajo final establecimos modelo que fuera capaz de predecir este PIB municipal del año 2010 al 2020 para cada municipio de Colombia en base a diferentes predictores que explican de una u otra forma el PIB.

$$y = f(X) + u$$

$$PIB_{mun} = f(X) + u$$

Para ello se desarrollaron 5 modelos para determinar cual era el mejor en términos de MSE y a partir de este se realizaron las predicciones del PIB municipal.

Datos

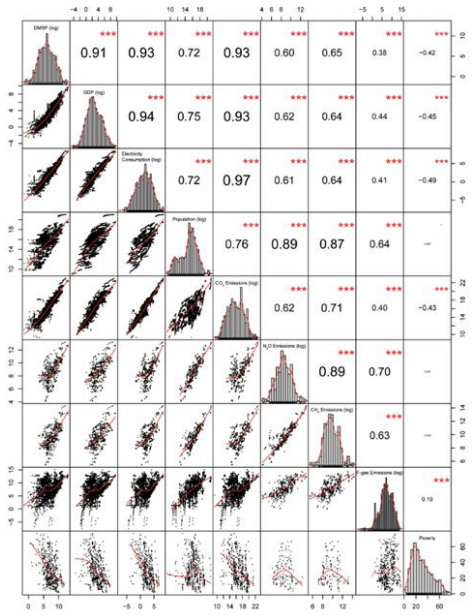
En la tabla 1 se resumen los datos utilizados para cada una de las variables utilizadas. Los datos utilizados son a nivel municipal por cada año y fueron obtenidos del Panel Municipal del CEDE a excepción de los datos de luminosidad que fueron obtenidos a partir del desarrollo realizado por el profesor Eduard Martinez para los años 2005 a 2018. Los datos de luminosidad del 2019 y 2020 fueron estimados en base a este desarrollo previo.

Igualmente, en la tabla 1, contiene la sección de bibliografía en donde se referencian los estudios que dan validez a cada una de las variables como explicativas del PIB. En particular los datos de luminosidad son útiles para

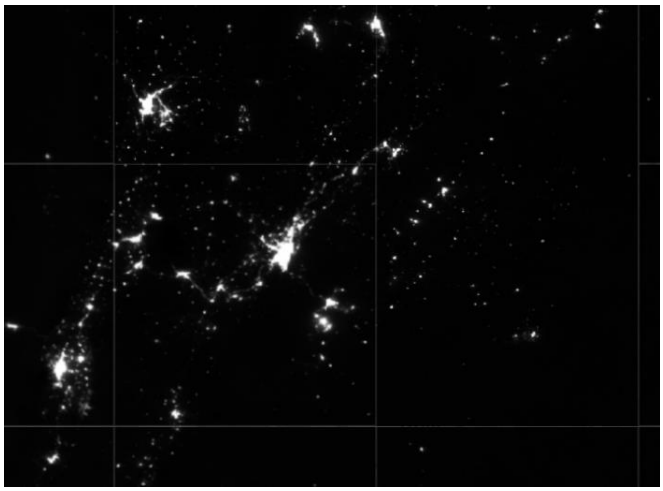
este trabajo ya que existe una fuerte correlación entre las luces nocturnas y las medidas del Producto Interior Bruto (PIB), a nivel nacional, estatal y regional.

Por ejemplo, algunos estudios como el de Proville et al. (2017) encontraron fuertes correlaciones entre los datos de luminosidad nocturna y el consumo de electricidad, las emisiones de CO2 y el PIB. En la figura 1 se aprecia esta correlación.

En la figura 2 se aprecia la luminosidad nocturna de Colombia para el año 2020. En particular en el centro del país donde se encuentra Bogotá se observa un área de alta intensidad lumínica. Esto acorde a la información del DANE (2020), Bogotá D.C. es la región que más aporta al PIB del país.



Gráfica 1. Correlaciones entre los datos de luminosidad nocturna y el consumo de electricidad, las emisiones de CO2 y el PIB



Gráfica 2. Intensidad lumínica nocturna en Colombia. Año 2021.

Variable	Fuente de datos	Bibliografía
PIB municipal (2005 - 2009)	Sánchez, F. y España, I. DANE, Censo 2005. CEDE	No aplica. (Variable de interés)
Datos de luminosidad (satélite VIIRS) 2005 - 2020	A harmonized global nighttime light dataset. Eduard Martinez.	Using luminosity data as a proxy for economic statistics
Número de habitantes (2005 - 2020)	(pobl_tot, DANE. Proyecciones de población)	The Role of Population in Economic Growth
Gastos de capital (Inversión) relacionados con la formación bruta de capital fijo y otros (2005 - 2020)	(g_cap, Buen gobierno CEDE)	ShowProperty (urf.gov.co)
Crédito interno y externo (2005 - 2020)	(finan_credito, Buen Gobierno CEDE)	El efecto del microcrédito en el PIB de Colombia, 2005-2018 (redalyc.org)
Distancia lineal a la capital del departamento (2005 - 2020)	(discapital, Características generales CEDE)	The Contribution of Economic Geography to GDP per Capita (oecd.org)
Valor crédito pequeños productores (2005 - 2020)	(vrf_peq, Agricultura CEDE)	Effects of credit on national and agricultural GDP, and poverty: a developing country perspective | SpringerLink
Área oficial municipio en km² (2005 - 2020)	(areaoficialkm2, Características generales del CEDE)	Competitive Cities Economic Growth - TCdata360 (worldbank.org)

Tabla 1. Variables utilizadas para predecir el PIB

Para el entrenamiento del modelo se utilizaron datos del 2005 al 2009, el cual es el periodo de tiempo donde existe la información del PIB municipal (base de datos *train*). Para las predicciones se utilizaron los datos del 2010 al 2020 (base de datos *test*). En las siguientes tablas se encuentran las estadísticas descriptivas, tanto para la base de datos de *entrenamiento* como la de *prueba*:

Characteristic	**N = 5,260**[1]
PIB Constante	359,782 (3,284,767)
Población total	41,104 (247,374)
Área oficial km2	825 (2,958)
Distancia a la capital	78 (56)
Gastos de capital	20,986 (175,552)
Promedio de luminosidad	4.7 (7.7)
Crédito interno y externo	33 (7,057)
Valor crédito productores	563 (753)

[1] Mean (SD)

Tabla 2. Estadísticas descriptivas base de datos train.

Characteristic	**N = 12,064**[1]
Población total	43,556 (263,979)
Área oficial km2	860 (2,897)
Distancia a la capital	79 (56)
Gastos de capital	47,898 (372,767)
Promedio de luminosidad	7 (8)
Crédito interno y externo	455 (21,097)
Valor crédito productores	1,650 (1,864)

[1] Mean (SD)

Tabla 3. Estadísticas descriptivas base de datos test

Como se aprecia en las tablas 2 y 3, los modelos fueron entrenados con 5260 observaciones. La base de datos para predecir tiene 12064 observaciones. Se aprecian datos parecidos en población con una media de 41104 habitantes para *train* y 43556 habitantes para *test*. La distancia a la capital en km se encuentra en 78 y 79, respectivamente. Se encuentran diferencias en los datos de luminosidad, en la de *train* el promedio es de 4.7 *Day-Night Band radiance* y en *test* es de 7 DNB. Esto tiene sentido ya que se espera que el pasar de los años la intensidad lumínica aumente, así mismo ocurre con variables como los "gastos de capital", el "crédito interno y externo" y el "valor de crédito de los productores".

Modelo

La forma funcional de todos los modelos estimados siguieron la siguiente estructura:

$$\ln(pib\ cons) = pobl\_tot + areaoficialkm2 + discapital + vrf\_peq\_productor + lights\_mean$$

Se decidió realizar **5 diferentes modelos de entrenamiento** para así encontrar el mejor modelo predictivo, estos cinco modelos fueron: **OLS, Ridge, Lasso, Elastic Net y Superlearner**. Los **modelos Ridge y Lasso** se construyeron a partir del **Lambda mínimo óptimo** y el **modelo Superlearner** se decidió construirlo en base a **Random Forest, Elastic Net** y **OLS.**

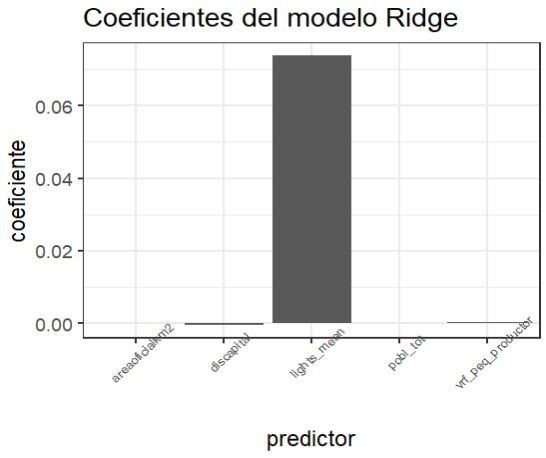
Resultados

Tabla 3. Modelos utilizados en el problema de predicción.

Modelos		MSE	Lambda Óptimo
Modelo 1	Estimación con OLS	1.512511	No aplica
Modelo 2	Estimación con Ridge – Alpha=0	0.02950	0.0663
Modelo 3	Estimación con Lasso – Alpha=1	0.06012	0.00918
Modelo 4	Estimación con Elastic Net	1.51948	0.01325095
Modelo 5	Superlearner (inluye glmnet, ranger, Lm, mean)	0.1458048	No aplica

Como se observa en la Tabla 3, el mejor modelo de predicción del logaritmo natural del PIB municipal es a partir del **Modelo 2 –Ridge (Alpha=0).** Este es el mejor modelo ya que, **es el modelo con menor MSE (0.02950)** lo cual se puede interpretar como el modelo que mejor predice el PIB municipal y el que menor variación presenta dentro de los datos (menor varianza). Un análisis valioso de los diferentes modelos de predicción realizados es que todos los MSE son significativamente bajos y esto se debe a que se decidió aplicar el logaritmo natural al PIB municipal generando esto que los datos tiendan a distribuirse de forma normal y que la varianza disminuya drásticamente.

Gráfica 3. Coeficientes estimados del modelo Ridge



Gráfica 4. Predicciones del PIB por departamento

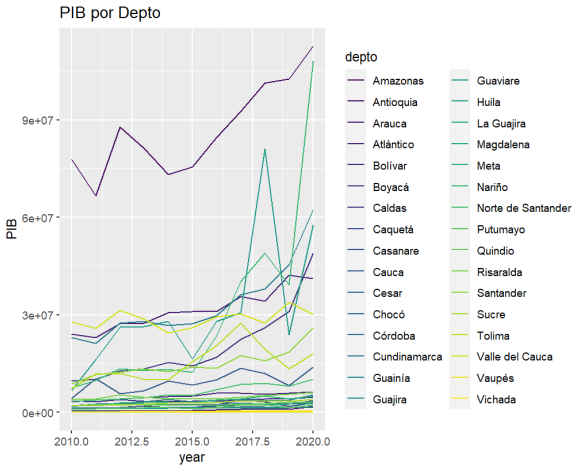


Tabla 4. Estadísticas descriptivas predicciones del PIB municipal predicho

Mínimo	1st cuantil	Mediana	Promedio	3rd cuantil	Máximo
3.856e+04	6.956e+04	9.447e+04	8.310e+07	1.480e+05	3.884e+11

La gráfica 3 nos muestra la importancia de las variables dependientes para predecir el logaritmo natural del PIB municipal analizándose que el promedio de la luminosidad es la variable más importante en la predicción. Específicamente, se puede interpretar que un aumento de 1 *Day-Night Band radiance* en el promedio de la luminosidad genera que el PIB municipal aumente en 7.38%, manteniendo las demás variables dependientes constantes. Una implicación de este resultado es que esta fuerte correlación entre la luminosidad y el Producto Interior Bruto (PIB) esta acorde con la literatura pues los investigadores consideran la luminosidad una variable clave para predecir el PIB a nivel estatal, regional y/o nacional.

Otras dos variables explicativas con importancia en la predicción del logaritmo del PIB municipal son la distancia del municipio a la capital del país y el valor de los créditos a los pequeños productores. Específicamente, se interpreta que un aumento en 1 km de la distancia a la capital por parte del municipio implica que el PIB municipal disminuya en 0.0385%, manteniendo las demás variables independientes constantes. Además, un aumento en 1 COP del valor de los créditos a los pequeños productores genera que el PIB municipal aumente en 0.311%, manteniendo las demás variables independientes constantes.

Por otro lado, en la Gráfica 4 se puede apreciar las predicciones realizadas del PIB por departamento (se decidió dar una representación gráfica departamental puesto que son miles de municipios para graficarlos todos). En esta gráfica se puede analizar de que Bogotá y Antioquía son los departamentos con mayor PIB predicho mientras que departamentos como Bolívar, Nariño y Chocó presentan los menores PIB predichos.

La Tabla 4 nos muestra las estadísticas descriptivas de las predicciones del PIB municipal en millones de COP (se decidió aplicar Euler a los logaritmos naturales del PIB predichos para obtener los valores en millones de COP) pudiéndose determinar de que el PIB municipal mínimo predicho es de 3.856e+04 COP mientras que el PIB municipal máximo predicho es de 3.884e+11 COP. El promedio del PIB municipal predicho es de 8.310e+07 con el mayor PIB municipal predicho de los siguientes municipios: Bogotá D.C, Antioquía, Valle del Cauca, Huila y Atlántico.

Conclusiones

El desarrollo del mejor modelo de predicción del PIB municipal se realizó por medio del entrenamiento a diferentes modelos predictivos y de la comparación de estos por medio del MSE. Este procedimiento nos llevó a determinar que **el mejor modelo predictivo realizado fue el modelo Ridge.** Los resultados de la predicción por medio del modelo Ridge nos muestra que los mayores PIB`s municipales predichos se encuentran en los siguientes municipios: Bogotá D.C, Medellín, Cali, Pitalito y Barranquilla mientras que los menores PIB's predichos se encuentran en: Norosí, Montecristo, San Jacinto del Cauca, Regidor y Montecristo (todos estos municipios están situados en el departamento de Bolívar).

Una conclusión valiosa de este ejercicio fue la transformación realizada a la variable dependiente (PIB municipal) al aplicarle el logaritmo natural, ya que al realizar las predicciones sin realizar esta transformación los modelos de entrenamiento presentaban unos MSE significativamente altos y, así mismo, los resultados de los coeficientes de los modelos no eran los mejores. De la misma forma, se encontró que la elección de las variables predictoras es crucial, ya que algunas de estas variables como "crédito interno y externo" y "gastos de capital" generaron predicciones erróneas, por lo cual fueron omitidas. Lo dicho con anterioridad es una muestra de que el manejo de los datos es esencial para lograr los mejores modelos y predicciones posibles.
Así mismo, para realizar estas predicciones del PIB municipal se utilizaron procedimientos innovadores como fue el uso de mapas de luminosidad para poder determinar la luminosidad promedio por municipio. De esta manera confirmamos que el Big Data no solo se refiere al manejo de grandes volúmenes de datos sino también de la obtención de datos de fuentes complejas y poco convencionales.

Datos y código

En el siguiente repositorio de GitHub se encuentran los datos y el código utilizado:

https://github.com/faor10/Proyecto-Final/

Referencias

Proville J, Zavala-Araiza D, Wagner G (2017) Night-time lights: A global, long term look at links to socio-economic trends. PLoS ONE 12(3): e0174610. https://doi.org/10.1371/journal.pone.0174610

DANE (2021). Cuentas nacionales departamentales. https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales