Universidad de
SanAndrés

Universidad de San Andrés

Big Data

Propuesta Final de Investigación

Profesoras: Noelia Romero y Victoria Oubiña

Integrantes: Vergara Guillen Julián, Sundblad Alfonso, Guerrero Ponzo Felicitas Trinidad

Fecha de entrega: 3 de Diciembre 2023

1. Introducción

Hoy por hoy el bienestar de la población se entiende desde diversos factores tanto socio-económicos (educación, salud, riqueza material) como geográficos (contaminación, recursos recreativos, entre otros). Dentro de la riqueza material, la posibilidad de satisfacer tus necesidades de consumo y tener prestaciones de servicios en la cercanía y variedad de opciones dentro de estas dada por la presencia de competencia entre empresas hacen a un mejor entorno y mejoran la calidad de vida dentro de un centro urbano.

La aparición de *machine learning* y *big data* nos han brindado herramientas que permiten entender el bienestar desde múltiples dimensiones y poder recolectar y trabajar con datos a un nivel de desagregación o especificidad al que la data administrativa muchas veces no llega. El objetivo de esta propuesta de investigación es estudiar la calidad de vida de la población urbana al mayor nivel de desagregación posible de los datos. Las ciudades y los grandes aglomerados urbanos son el espacio en el que transcurre la vida de la mayoría de los ciudadanos argentinos. Ahora bien, dentro de una misma ciudad, la calidad de vida de la población es muy heterogénea. En este sentido, analizaremos si el volumen y oferta de servicios disponibles en distintas partes de una misma ciudad predicen el nivel de bienestar intra ciudad para los grandes aglomerados urbanos de Argentina a través de la utilización de modelos no lineales de *machine learning*. Nuestro prior, es que aquellos lugares con mayor cantidad de opciones de consumo y servicios y volumen de transacciones son también aquellos lugares en donde mejor se vive en el país.

Esta propuesta de investigación creemos que contribuye a la literatura existente en tres dimensiones diferentes: 1) desagregación de datos 2) entender desde otras dimensiones el bienestar de la población 3) herramientas de *machine learning* y uso de información satelital y de *e-commerce*. Con respecto a la primera, nuestra propuesta utiliza datos al mayor nivel de desagregación posible disponible para

Argentina, que es a partir de los radios censales y datos de Google Maps (datos satelitales). Nos permite entender cómo cambian las dinámicas socio-económicas intra-ciudad o intra-localidades. En relación a lo segundo, el bienestar de un individuo también depende del entorno que lo rodea. Esto implica gozar en su cercanía de la mayor disponibilidad de servicios posibles, opciones de consumo y variedad (competencia), opciones de recreación son importantes para el bienestar de las personas por lo que consideramos que extender los análisis de bienestar en esta dimensión es relevante. Finalmente, en relación al tercer punto, podemos establecer la validez de los modelos no lineales para predecir, y en particular, si el volumen de servicios y transacciones a nivel de una determinada zona están relacionadas con el bienestar.

2. Literatura Previa

Las técnicas de *machine* learning y el uso de datos masivos han permitido analizar la pobreza, la riqueza, el bienestar (en múltiples dimensiones) y realizar predicciones a través del uso de imágenes satélites, canastas de consumo a partir del *e-commerce,* registros de llamadas telefónicas y movimientos de una persona a lo largo de una ciudad, entre otros. En esta sección detallaremos trabajos que consideramos relevantes para nuestra propuesta.

En el paper de Wojcik & Krystian Andruszek (2022) se analiza el nivel de bienestar intraurbano. Buscan comparar si los métodos no lineales de *machine learning* predicen mejor que los lineales. Además, buscan responder la duda de que tan útiles realmente son los datos satelitales gratis. Los autores generan un índice de bienestar basado en tres dimensiones: salud, educación y el ingreso per cápita. Usaron datos de *Google Maps*, en el que extrajeron el número y área de los edificios (proxy para la tasa de urbanización) y el área de espacios verdes (*amenities* verdes). Otros posibles predictores se obtuvieron de *OpenStreetMap*, incluyendo datos de edificios, calles, ríos, lagos, áreas verdes, paradas de transporte público, alquiler de bicicletas, estación de servicio, supermercados y centros

comerciales. Finalmente, los autores indican que los modelos no lineales son capaces de predecir diversas dimensiones del bienestar con una precisión mayor que el enfoque lineal. El uso de datos satelitales permite identificar la heterogeneidad urbana. En la misma línea, Meursault et al (2022) en conjunto con datos demográficos a nivel distrital para EEUU busca modelos para mejorar la precisión en la predicción de riesgo creditico en zonas de bajos ingresos.

Por otro lado, en aglomerados urbanos para Indonesia, Wijaya et al. (2020) predicen la pobreza en las ciudades a través de elaborar canastas de consumo representativas del gasto en el hogar con datos de *e-commerce*. En particular, predice el nivel de pobreza en determinadas zonas de una ciudad de acuerdo al volumen de transacciones y el valor de los bienes que se transan en cada zona de una ciudad. En África de Jean, Burke et al (2016) utilizan datos de encuestas e imágenes satelitales de cinco países africanos. Buscan simplificar la búsqueda de pobreza e implementación de políticas. Concluyen que los modelos son altamente predictivos de los consumos promedio de los hogares y la riqueza de activos.

En la misma sintonía, Engstrom, Hersh, & Newhouse, (2017) buscan llenar un data gap de los países con pocos datos. Buscan utilizar las texturas obtenidas de las imágenes satelitales para predecir la tasa de pobreza a niveles locales. Utilizan un área de muestra de 3500 km² en Sri Lanka, observando áreas urbanas rurales y propiedades. Esta interesados en ver qué características están más correlacionadas con las medidas de bienestar, si funciona igual de bien en sectores pobres y ricos, si se puede usar en otras áreas, y que tan robusta es la predicción si usamos menos hogares y corremos una sola simulación.

En cuanto al bienestar en la vida urbana, Michelangeli & Peluso (2016) busca medir la desigualdad entre ciudades en Italia. Para esto, analiza que la distribución de bienes locales y servicios a través de las ciudades mejora el nivel promedio de bienestar en un país dado. Individuos expuestos en mayor

cantidad y calidad a estos servicios y bienes tienen una mejor calidad de vida. Los autores encuentran que hay disparidades significativas debido a la diferencia en la disponibilidad de servicios, infraestructura, condiciones económicas de transporte y de educación.

3. Base de Datos

3.1 Variable Dependiente

Nuestro modelo buscará predecir el índice de calidad de vida (ICV) elaborado por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Este índice utiliza una escala numérica continua del 1 al 10, siendo 1 el menor nivel de bienestar posible y 10 el máximo. Ahora bien, de acuerdo al score en el ICV hay 10 categorías posibles para cada zona $\{D1, D2, D3, D4, D5, D6, D7, D8, D9, D10\}$. Siendo $D1$ los radios con la mejor calidad de vida posible y $D10$ las zonas con la peor calidad de vida posible.

El ICV está elaborado a partir de los 52408 radios censales del país y el indicador está compuesto tanto por indicadores socio-económicos como ambientales. Dentro de la dimensión socio-económica se incluyen las variables de educación (porcentaje de la población con 15 años o más con primario incompleto y universitario completo), salud (tasa de mortalidad infantil y proporción de la población sin cobertura médica) y vivienda (proporción de la población sin inodoro o con inodoro sin descarga y porcentaje de la población en hogares hacinados). Dentro de la dimensión ambiental se consideran variables relacionadas a problemas ambientales (contaminación urbana, inundabilidad, basurales, externalidades negativas, entre otros) y variables relacionadas a los recursos recreativos de la zona ya sea de base natural (playas, relieve, parques y espacios verdes, entre otros) como socialmente construidos como patrimonios urbanos, centros deportivos, centros comerciales entre otros). Las variables que componen la dimensión socio económica son extraídas de los Censos del 2001 Y 2010

mientras que las relacionadas a la dimensión ambiental son extraídas de informes municipales, imágenes satelitales y análisis del terreno.

Para nuestro trabajo, tenemos la posibilidad de agrupar las categorías. Es razonable dada la naturaleza continua de la categoría, mientras más cerco de 1 estoy, peor es la calidad de vida y mientras más cerca de 10 mejor es la calidad de vida. Además, esto nos permitirá darle un mejor tratamiento a la muestra y maximizar el uso de las observaciones. Para esto, redefiniremos en tres categorías: buena, intermedia y mala. Definiremos como buena calidad de vida a aquellos radios censales con un ICV catalogado como $D1, D2 \, o \, D3$. Por otro lado, definiremos como intermedio a aquellos con ICV $D4, D5, D6 \, y \, D7$, mientras que mala a aquellos con un ICV que este en $D8, D9 \, y \, D10$.

Finalmente, debemos considerar que solo utilizaremos los radios censales que esten dentro de ciudades o grandes aglomerados urbanos. Para esto, respetaremos la categorización utilizada por los elaboradores del CONICET que es la de Vapñarsky & Gorojovsky (1990). Para nuestra definición nos quedaremos con los radios censales de metrópolis y grandes ciudades (más de 1 millón de habitantes) que son CABA, algunos partidos de Buenos Aires, Rosario y Córdoba y ciudades intermedias mayores entre 400000 - 999999 habitantes que abarcan ciudades como San Juan, Mendoza, Tucumán, Corrientes, Paraná, más partidos de Buenos Aires, entre otros.

3.2 Predictores

En nuestro trabajo tendremos dos tipos de predictores. En primer lugar, aquellos relacionados a la oferta de servicios de una determinada zona. En segundo lugar, aquellos relacionados al volumen de transacciones. La construcción de la base de datos se realizará a partir de estos dos frentes. Para la construcción de la base estableceremos las zonas geográficas de interés al nivel que están definido en el ICV (radios censales)

Con respecto a la oferta de servicios, se destacarán la cantidad y variedad de empresas relacionadas al consumo y acceso a servicios financieros. Las variables serán: cantidad de cajeros automáticos, cantidad de bancos, cantidad de diferentes empresas distintas de bancos, cantidad de supermercados y farmacias, una *dummy* que indica si posee centro comercial, restaurantes y bares, estaciones de servicio, si tiene templos religiosos, si tiene comisarias, si tiene centros de salud y los puntos de transporte (paradas de colectivos y cantidad de líneas, subte, tren). Notar que las variables seleccionadas son representativas de un amplio espectro de bienes y servicios que se consumen en una ciudad. Si un ciudadano, cuenta con mayor variedad de opciones de consumos en cada una de estas categorías su bienestar debería mejorar.

En cuanto al *scrapping* de los datos, se propone un método eficiente para recopilar los datos detallados utilizando la *API de Places* y respuestas recibidas con la biblioteca *Requests* y *Pandas*. Para estos se deben delimitar los siguientes parámetros: ubicación (*location*), radio de búsqueda (*radius)* y tipo de resultados en la búsqueda(*type*). Una vez recopilados los datos, se agrega la información. Este paso es esencial para obtener resultados estructurados que proporcionan una visión detallada de los servicios en las zonas de interés.

En relación al volumen de transacciones utilizaremos información proveniente de comercio electrónico. En la actualidad, el consumo de bienes y servicios a través de plataformas es un indicador representativo del nivel de consumo que puede existir en una determinada zona (a través del análisis de los compradores y vendedores). Para Argentina, la plataforma de comercio electrónico más grande es Mercado Libre. Por este motivo, hemos optado por considerar información de los vendedores en dicha plataforma. En particular nos resulta relevante estudiar en cada zona: *Cantidad de Vendedores*, *Volumen de Ventas Producido*, *Calidad de Vendedore*s. Zonas con una mayor cantidad de oferentes,

mayor volumen de ventas y vendedores con certificaciones *premium* de Mercado Libre deberían predecir un ICV más alto.

Para recopilar estas variables proponemos, en primera instancia, la utilización de la API de Mercado Libre. Sin embargo, enfrentamos un desafío significativo dado que no existe una funcionalidad directa para visualizar vendedores por zona geográfica. La única manera de acceder a los distintos datos de los vendedores es a través de una publicación de venta. Nuestra estrategia para superar este obstáculo implica, mediante HTTP requests, extraer datos valiosos del vendedor, como su *nombre*, *Id*, *provincia*, *zona*, *cantidad de ventas* (canceladas, completadas, totales), *proporción de ratings* (positivos, neutrales, negativos), e *índice de calidad del vendedor* (que incorpora múltiples parámetros). Primero realizaremos una búsqueda de productos comunes (yerba, remera, jeans, celular, bicicleta, entre otros) para guardar los IDs asociados a cada provincia o gran aglomeración. Esto se encuentra en los *available_filters* de la búsqueda, pero se deberá reiterar múltiples veces hasta que obtengamos todos los IDs que necesitemos. Posteriormente, restringimos la zona en nuestras búsquedas de la API con estos identificadores y obtenemos información detallada de los vendedores en todas las áreas de interés. No obstante, esta solución presenta dos desafíos significativos: en primer lugar, la capacidad de filtrar es limitada y solo permite agregaciones generales, los valores más desagregados solo se pueden obtener una vez seleccionado el vendedor; en segundo lugar, debido a esto existe la posibilidad de que muchas localidades queden excluidas de nuestros datos sin una solución clara.

Ante estas dificultades, hemos explorado una fuente alternativa para obtener los datos. De esta forma nos encontramos con *Nubimetrics*, una empresa especializada en proporcionar datos de mercado y competencia sobre Mercado Libre. Para acceder a estos datos existen dos opciones: de manera gratuita al contar con una cuenta de Mercado Libre que registre al menos 10 ventas en los últimos 6 meses o una opción de pago de 20 USD. Estas bases de datos contienen todos los datos mencionados

anteriormente, categorizados de diversas formas, incluyendo una clasificación detallada por zona, que es de particular interés para nuestro estudio.

4. Metodología

Nuestra propuesta busca predecir Y_i con $Y = 0,1,2,3,\ldots,10$ que es el índice de calidad de vida del CONICET. Como mencionamos previamente, este índice es una variable categórica ordenada. Nuestros predictores X_i son las variables definidas y mencionadas en la sección anterior. Evaluaremos cuatro modelo para predecir Y_i: *Análisis Discriminante Cuadrático* (QDA), *Random Forest* (RF), *Gradient Boosting* (GB) y *Bagging* (BB).

Notar que todos los modelos elegidos son no lineales y dejamos de lados otros potenciales modelos que podemos utilizar como Análisis Discriminante Lineal (LDA) o un modelo de Regresión Logística Multinomial Ordenada (Logit). La razón de esto es que los modelos no lineales tienen mayor flexibilidad que los lineales (relajan supuestos). Además, Wojcik & Andruszek (2021) mencionan que estudios previos de (Bickenbach et al., 2016; Gennaioli 2014; Mellander et al., 2015) sugieren que la relación entre diferentes medidas de bienestar y diferentes tipos de predictores a nivel regional y local no es lineal.

4.1 Análisis Discriminante Cuadrático (QDA)

Con QDA estimaremos $P(Y = c|X)$ siendo *c* una de las tres categorías y X nuestros predictores. Para utilizar QDA suponemos que las X tienen una distribución normal. No obstante, en comparación a LDA, nos permite relajar el supuesto con respecto a igualdad de varianzas dentro de cada categoría. Por esta misma razón es mucho más flexible que LDA ya que tenemos una matriz de covarianza por separado para cada clase.

4.2 Random Forest (RF)

Con *Random Forest*, en cada árbol no elegimos todos los predictores, sino un set de menor dimensión. Tomamos una muestra bootstrapeada de los datos de entrenamiento y para cada una armamos un árbol en el que seleccionamos aleatoriamente el número de predicotres. En cada conjunto de árboles resultantes calcularemos la predicción. Cada observación será asignada a una determinada categoría por el criterio de *voto mayoritario*. Es decir, es asignado a la categoría que más veces es predicha. En este modelo, estimaremos por *validación cruzada* el número de árboles y el número de predictores aleatoriamente seleccionados.

4.3 Gradient Boosting (GB)

Con *Boosting* realizamos algo similar a *bagging* pero no analizamos las predicciones simultáneamente sino que secuencialmente. Computamos esta vez B arboles secuenciales, en el que cada paso le damos más ponderación a aquellas observaciones mal predichas. Nuevamente, a través de *validación cruzada* elegimos el numero óptimo de B árboles. También por *validación cruzada* elegimos la penalización λ que es la tasa a la cual crece cada árbol.

4.4 Bagging Classifier (BB)

Con *Bagging* tenemos una determinada cantidad B de árboles de entrenamiento independientes. El número de árboles lo determinamos a través de *validación* cruzada. Una vez hecho esto, clasificaremos cada observación en una determinada categoría del ICV por *voto mayoritario*. Es decir, las observaciones serán asignadas a la categoría que más se repite en las predicciones.

4.4 Evaluación de desempeño

Para evaluar la performance del modelo utilizaremos diferentes métricas de precisión. Entre ellas, utilizaremos precision (precision), *recall* (ratio de verdaderos positivos), *accuracy* (exactitud)[1], y el Error Cuadrático Medio (ECM). Con respecto a las dos primeras medidas, utilizaremos agregaciones de estas para comparar los modelos. En particular, utilizaremos la macro-agregación y la micro-agregación. Esto se debe a que tenemos 11 categorías a predecir. Esto implica realizar múltiples comparaciones por modelo para cada categoría lo cual puede generar dificultades a la hora de determinar qué modelo es mejor.

Por un lado, la macro-agregación implica calcular el número de verdaderos positivos, falsos positivos y falsos negativos para cada clase. Computar la *precision* y el *recall* para cada clase, y luego hacer un promedio para cada una de las métricas a través de todas las clases. Por otro lado, la micro-agregación implica calcular el número de verdaderos positivos, falsos positivos y falsos negativos a través de todas las clases y luego calcular las medidas de precisión en cuestión.

5. Conclusiones y Limitaciones

Creemos que nuestro trabajo tiene dos limitaciones principales. En primer lugar, el ICV está elaborado a partir de los datos del Censo del 2001 y 2010. Mientras que nuestros predictores están elaborados a partir de la última información disponible. Es decir, la distinta temporalidad puede generar que la realidad que buscan predecir nuestras variables no sea la del radio censal al día de hoy. Frente a esta limitación, proponemos una posible solución para mitigar el problema que es actualizar el ICV con los datos del Censo del 2022 replicando la elaboración que realizó el CONICET. En segundo lugar, los predictores relacionados a las transacciones elaborados a partir de Mercado Libre están a un nivel de

[1] https://www.evidentlyai.com/classification-metrics/multi-class-metrics#binary-vs-multi-class-classification

agregación superior que los radios censales del índice del CONICET. Por ejemplo, en CABA está a nivel comuna, en Buenos Aires está a nivel municipio. Por lo tanto, reflejan la información de varios radios censales. Sin embargo, creemos que la variable debería seguir teniendo poder predictivo, debido a que en muchas localidades la gran mayoría de los radios censales dentro de esta reflejan la misma situación en el ICV.

En conclusión, nuestra propuesta de investigación propone predecir y ver la relación que hay entre el bienestar intra-urbano y el volumen de transacciones y cantidad de servicios que tiene la zona. Como mencionamos en la motivación de nuestro trabajo, el bienestar se define desde múltiples dimensiones y creemos que la cantidad de opciones de consumo de bienes y servicios que se dispone intra-ciudad puede servir para predecir y entender la calidad de vida de la zona. La disponibilidad y variedad de opciones de consumo también hacen a un mejor bienestar del individuo.

6. Bibliografía Consultada

Wójcik, P., & Andruszek, K. (2022). Predicting intra-urban well-being from space with nonlinear machine learning. Regional Science Policy & Practice, 14(4), 891-913.

Meursault, V., Moulton, D., Santucci, L., & Schor, N. (2022). One Threshold Doesn't Fit All:Tailoring Machine Learning Predictions of Consumer Default for Lower-Income Areas. In One Threshold Doesn't Fit All: Tailoring Machine Learning Predictions of Consumer Default for Lower-Income Areas: Meursault, Vitaly| uMoulton, Daniel| uSantucci, Larry| uSchor, Nathan. [Sl]: SSRN.

Vapñarsky, C. A., & Gorojovsky, N. (1990). El crecimiento urbano en la Argentina. (No Title).

Bickenbach, F., Bode, E., Nunnenkamp, P., & Söder, M. (2016). Night lights and regional GDP. Review of World Economics, 152, 425-447.

Gennaioli, N., La Porta, R., Lopez De Silanes, F., & Shleifer, A. (2014). Growth in regions. Journal of Economic growth, 19, 259-309.

Mellander, C., Lobo, J., Stolarick, K., & Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity?. PloS one, 10(10), e0139779.

Engstrom, R., Hersh, J. S., & Newhouse, D. L. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being. World Bank Policy Research Working Paper, (8284).

Irvin, J., Laird, D., & Rajpurkar, P. (2017). Using satellite imagery to predict health. Technical report, Stanford University, Department of Computer Science.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.

Velázquez, G. Á., Mikkelsen, C. A., Linares, S., & Celemín, J. P. (2014). Calidad de vida en Argentina: Ranking del bienestar por departamentos (2010).

Velázquez, G. A., & Celemin, J. P. (2019). Geografía y calidad de vida en la Argentina. *Journal de Ciencias Sociales*.

Peluso, E., & Michelangeli, A. (2016). Cities and inequality. *Region: the journal of ERSA*, *3*(2), 47-60.

7. Bases de datos

Google Maps API. https://developers.google.com/maps/documentation [accessed 2023-12-03].

OpenStreetMap API. https://wiki.openstreetmap.org/ [accessed 2023-12-03].

Data and Statistics Page - FEDERAL RESERVE BANK of NEW YORK. https://www.newyorkfed.org/data-and-statistics [accessed 2023-12-03].

USA Census Bureau Data. https://data.census.gov/ [accessed 2023-12-03].

Índice de Calidad de Vida - IGEHCS & ISISTAN (UNCPBA & CONICET). https://icv.conicet.gov.ar/ [accessed 2023-12-03].

Google Places API. Google Maps for Developers. https://developers.google.com/maps/documentation/places/web-service [accessed 2023-12-03].

Requests - Python Library. https://requests.readthedocs.io/en/latest/ [accessed 2023-12-03].

Pandas - Python Library. https://pandas.pydata.org/docs/ [accessed 2023-12-03].

Mercado Libre API. https://developers.mercadolibre.com.ar/en_us/introduction-products [accessed 2023-12-03].

Nubimetrics. https://www.nubimetrics.com/producto/mercado [accessed 2023-12-03].