# Modelo Multivariado para predecir la localización de la población a partir de Factores Sociodemográficos en Colombia

**Carlos Valbuena Acosta**

Director: **Mario Julián Mora Cardona**

Tesis presentada como requisito para optar por el título de Magíster en Ciencia de Datos.


Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana Cali
Colombia
28 de junio de 2024

# Modelo Multivariado para predecir la localización de la población a partir de Factores Sociodemográficos en Colombia

## Resumen

El objetivo de este proyecto era determinar cuáles son los factores que inciden sobre la localización de un individuo en Colombia. Para lograrlo, implementó el algoritmo Propensity Score Matching con base en los datos del Censo 2018 para la población del Valle del Cauca, en los módulos de personas, hogares, viviendas y marco de georreferenciación, con un universo de 3,2 millones de registros y 40 variables seleccionadas.

Para cumplir el objetivo del proyecto, se construyeron 3 bases de datos con sus grupos de tratamiento y control, así: el primero con datos urbanos de Cali y los demás municipios, el segundo, solo con registro urbanos de Cali y el tercero, con registros del área urbana y centros poblados de Cali. Sobre estos algoritmos se entrenó el PSM, partiendo de una preparación de los datos, luego se realizó la estimación del propensity score que es la determinación del problema binario, es decir, la obtención de la probabilidad que un individuo se ubique en un grupo u otro para hacer las muestras comparables, seleccionando los Conjuntos 1 y 2 con el mejor nivel de accuracy con 61 % y 50 % respectivamente debido a la alta variabilidad que reviste una base como el Censo; con estos dos conjuntos se dio paso a la fase de emparejamiento a través de vecinos más cercanos – KNN, donde el conjunto 1 de Cali y los demás municipios obtiene las menores diferencias en las variables observables luego del emparejamiento.

Posteriormente, para predecir la manzana geográfica como unidad mínima de granularidad que ubica al individuo dentro de los Shapes del Censo-DANE, se implementó el clasificador Random Forest, el cual mostró dificultades para predecir la ubicación en una categoría compuesta por 22 caracteres, alcanzando un accuracy de 32 %, luego se hicieron unas transformaciones en la variable a predecir sin afectar su origen, logrando un mejor resultado del 39 % con la predicción de los últimos 8 campos de la localización de los individuos de Cali, pero debido al alto costo computacional este modelo no se pudo replicar para datos nuevos provenientes de SISBEN. Finalmente, se espera que este proyecto contribuya a profundizar los análisis económicos que desarrolle el Centro de Investigación Aplicada Riqueza Completa, mediante la implementación de algoritmos de emparejamiento como el PSM, especialmente dentro del uso de variables sociodemográficas como el Censo y su potencial capacidad para determinar la localización de un individuo a partir de estas.

# Índice general

# Índice de figuras

# Índice de cuadros

# Anexos

**Anexo 1:** Enlace de la carpeta que contiene los archivos utilizando en los diferente procesos de entrenamiento del Modelo.

Bases de Datos

**Anexo 2:** Enlace del script final creados en el proyecto de grado.

- Conjunto 1 – Valle Urbano
- Conjunto 2 – Cali Urbano
- Conjunto 3 – Cali Urbano y Centros Poblados

# Introducción

El estudio de características sociodemográficas de una población específica, es un proceso que se logra gracias al desarrollo de disciplinas como la estadística y la ciencia de datos, por medio de la implementación de distintas técnicas de recopilación, limpieza, depuración, transformación y presentación de grandes volúmenes de información para la toma de decisiones [1]. En Colombia, el Departamento Administrativo Nacional de Estadística – DANE – es la institución encargada de proporcionar datos que se utilizan para el análisis de la situación demográfica y social para la toma de decisiones a nivel nacional y territorial, contribuyendo a la consolidación de un estado social de derecho, con justicia social, económica y ambiental.

El Censo Nacional de Población y Vivienda -CNPV- 2018, realizado por el DANE en Colombia, presenta una radiografía reciente sobre las dinámicas demográficas del país, cuyos resultados indicaron que el 77,1 % de la población se concentra en las áreas urbanas, y que el 28,3 % viven en Bogotá, Medellín, Cali y Barranquilla, lo que ha incrementado los fenómenos de segregación de la población en las áreas periféricas de las principales ciudades [2].

No obstante, los microdatos del DANE no permiten conocer al detalle las características de un individuo y su relación con la ubicación en un territorio específico, lo que constituye una problemática para los análisis económicos que se adelantan desde los principales centros de estudio. Por lo tanto, desde el Centro de Investigación Aplicada – Riqueza Completa de la Facultad de Economía y Administración de la Pontificia Universidad Javeriana de Cali, surgió la necesidad de crear un modelo multivariado, a través de un matching que permitiera emparejar las características sociodemográficas en un territorio determinado a partir de las variables de los microdatos del Censo 2018 [3]. El proyecto generó como resultado una primera aproximación a la manera de abordar este tipo de problemáticas y de esta manera contribuir a la toma de decisiones en materia de política pública en las condiciones de hábitat de la población.

El presente documento contiene los elementos principales del proyecto denominado *"Modelo multivariado para predecir la localización de la población a partir de factores sociodemográficos en Colombia"*, tales como: el planteamiento y definición del problema que se pretende solucionar a través de la pregunta de investigación, objetivo general, objetivos específicos y justificación del proyecto, marco teórico. Posteriormente, se expone todo el desarrollo del proyecto, desde el procesamiento y análisis exploratorio de datos, la implementación del Propensity Score Matching (PSM) y su evaluación, la construcción del instrumento de visualización. Finalmente, el documento expone las conclusiones y las referencias bibliográficas utilizadas para el desarrollo del trabajo de grado.

# Capítulo 1

# Definición del Problema

## 1.1. Planteamiento del Problema

El Departamento Administrativo Nacional de Estadística – DANE de Colombia tiene como objetivo garantizar la producción, disponibilidad y calidad de la información estadística del país [4], en ese sentido dentro de las principales operaciones estadísticas que diseña y ejecuta se encuentra el Censo Nacional de Población y Vivienda -CNPV, que para su última medición en el año 2018 y su posterior corrección en 2020, alcanzó un total de 44,1 millones de personas censadas [2], donde el 77,1 % de estas reside en las ciudades y áreas metropolitanas, siendo Bogotá, Medellín, Cali y Barranquilla las ciudades que concentran un 28,3 % de la población [2] y a su vez aportan el 38,8 % del Valor Agregado Nacional [5].

Este fenómeno obedece a diversas problemáticas que van desde la migración interna de la población proveniente de ciudades intermedias o áreas rurales en busca de mayores oportunidades laborales en las grandes ciudades, hasta factores como el desplazamiento forzado producto del conflicto armado que se ha concentrado en las zonas periféricas del país [6]. A lo anterior, se le suma el éxodo migratorio de venezolanos desde la crisis del 2016, algunas estimaciones indican que para 2020 residían en Colombia un total de 1,9 millones de personas originarias del vecino país [7].

A pesar de que los resultados del Censo bajo unos criterios de anonimización permiten obtener cierta información para la totalidad de la población, pues su carácter es público, su naturaleza misma no permite conocer al detalle cual es la relación que existe entre las características sociodemográficas y la localización del individuo, lo que constituye una dificultad para los principales centros de pensamiento económico del país. Es por ello que se hace necesario construir un modelo multivariado para determinar cuáles son las características sociodemográficas que mayor inciden sobre la ubicación de una persona y poder probar su precisión mediante un emparejamiento, el cual se implementará desde el Centro de Investigación Aplicada – Riqueza Completa de la Facultad de Economía y Administración de la Pontificia Universidad Javeriana de Cali a partir de las variables del Censo 2018 [3], obteniendo un producto que sirve como insumo para la toma de decisiones en materia de políticas que promuevan la calidad de vida y mejoramiento de las condiciones de hábitat en las principales ciudades del país.

En otras palabras, uno de los grandes retos que enfrenta el Centro de Investigación Aplicada es identificar con buena precisión personas que en la base de datos CNPV 2018 no se pueden reconocer de manera única por su naturaleza de anonimización, pero que son indispensables para los desarrollos y estudios que se están haciendo. Para esto la experimentación de diferentes técnicas de minería de datos fue el centro de este trabajo.

## 1.2. Formulación del Problema

El problema que se resolvió dentro de la presente propuesta de investigación respondió a la siguiente pregunta: **¿ Cómo predecir la ubicación geo - censal de una persona a partir de sus características sociodemográficas?**, obteniendo con esto las siguientes preguntas de sistematización: ¿Cómo seleccionar las variables que representan las características sociodemográficas más relevantes para la localización de un individuo?, ¿Cómo construir Propensity score matching (PSM) para realizar el emparejamiento entre las características sociodemográficas y la localización de un individuo?, ¿Cómo evaluar el desempeño del modelo de PSM y afinar sus resultados?, ¿Cómo presentar la información de las características socioeconómicas y los resultados obtenidos sobre el emparejamiento de la población?

# Capítulo 2

# Objetivos del Proyecto

## 2.1. Objetivo General

Predecir la ubicación geo –censal de una persona a partir de las características sociodemográficas que se encuentran en el Censo 2018.

## 2.2. Objetivos Específicos

1. Realizar el procesamiento de la base de datos a partir de los Microdatos del Censo poblacional 2018 – DANE.

2. Realizar el análisis exploratorio para seleccionar las características sociodemográficas de la población que se incluirán en el modelo.

3. Implementar un Propensity Score Matching (PSM) para ubicar personas con características sociodemográficas dadas sobre el mapa geo - censal del Censo Poblacional 2018 – CNPV.

4. Evaluar y afinar el modelo de PSM.

5. Desarrollar un instrumento de visualización que permita analizar los resultados del emparejamiento.

# Capítulo 3

# Justificación

El sector gubernamental requiere diseñar mejores programas y proyectos que propendan por una mayor calidad de vida para sus habitantes, por tanto, se hace necesario contar con información de valor sobre la realidad socioeconómica de un territorio en particular, en este sentido el presente proyecto surge a partir de la problemática que implica realizar un proceso de emparejamiento o matching de las personas a partir de sus características sociodemográficas y su ubicación según el Censo 2018, para aproximar esa persona a un territorio geo-censal, y así determinar que variables inciden en mayor medida sobre la ubicación espacial de un individuo.

Este proyecto es una iniciativa viable debido a que los datos que se utilizarán para su desarrollo corresponden a los microdatos del CNPV 2018 - DANE, los cuales son de dominio público, así mismo, se aportará la experticia del Centro de Investigación Aplicada Riqueza Completa en el estudio de estas problemáticas en el Valle del Cauca a partir de las estadísticas del DANE, lo cual sin duda permitirá obtener una llave para el emparejamiento de la población en términos geoespaciales, que contribuirá en gran medida sobre todos los análisis socioeconómicos que se desarrollan desde este centro, trazando un nuevo camino sobre cómo abordar este tipo de problemas y generando un importante aporte a través de esta herramienta robusta para la toma de decisiones de hacedores de política pública no solo en el departamento del Valle sino también en otras regiones de Colombia. Esto contando con datos suficientemente preprocesados del CNPV 2018 que se encuentran en el Centro de Investigación.

# Capítulo 4

# Marco de Referencia

## 4.1. Marco Teórico

El Propensity Score Matching (PSM) propuesto por Rosenbaum y Rubin en 1983 [8], es un método cuasi experimental ampliamente utilizado en el campo de la evaluación de impacto, cuyo objetivo es mitigar los problemas de sesgos de selección en donde los datos no provienen de muestras aleatorias, en este sentido, el método permite encontrar un individuo idéntico (grupo de control artificial) para cada uno de los grupos de tratamiento a partir de un conjunto de características observables que expliquen la asignación de la intervención y los posibles resultados del proceso de selección.

Según [9], este algoritmo se implementa en el marco de los siguientes supuestos:

### 4.1.1. Independencia Condicional

Este supuesto tiene un carácter estricto, pues establece que la participación de un individuo no obedece a variables no observadas, es decir, que esta se atribuye únicamente a las variables observables del individuo.

$$Y(0), Y(1) \perp D|X, \quad \forall X \tag{4.1}$$

Donde:

$Y(0)$ es el valor del grupo de individuos no intervenidos o grupo control
$Y(1)$ es el valor del grupo de intervención o tratamiento
$D|X$ son los distintos valores que puede tomar la variable $X$

Para el caso de variables con muestras muy grandes, puede resultar complejo la implementación del estimador, por tanto, [9] sugiere emparejar los individuos con base en:

$$P(X) = P(D = 1|X) \tag{4.2}$$

Donde:

$P(X)$ Prob. participación en el programa o intervención
$P(D = 1X)$Prob. condic. participar en programa (D=1) por caract. observables (X).

Lo que representa la probabilidad de participación de un individuo en un programa, a partir de sus características observables.

### 4.1.2. Condición de Soporte Común

Este supuesto implica asegurarse que existan similitudes entre los individuos que conforman el grupo de tratamiento y control, es decir, que las personas que cuenten con un mismo vector de características o variables tienen una probabilidad positiva de ser seleccionado o no para participar del programa o intervención en estudio. Lo cual se expresa así:

$$0 < P(D = 1|X) < 1 \tag{4.3}$$

El cumplimiento de esta condición, también denominada *Condición de Superposición* se puede evidenciar en el siguiente gráfico:

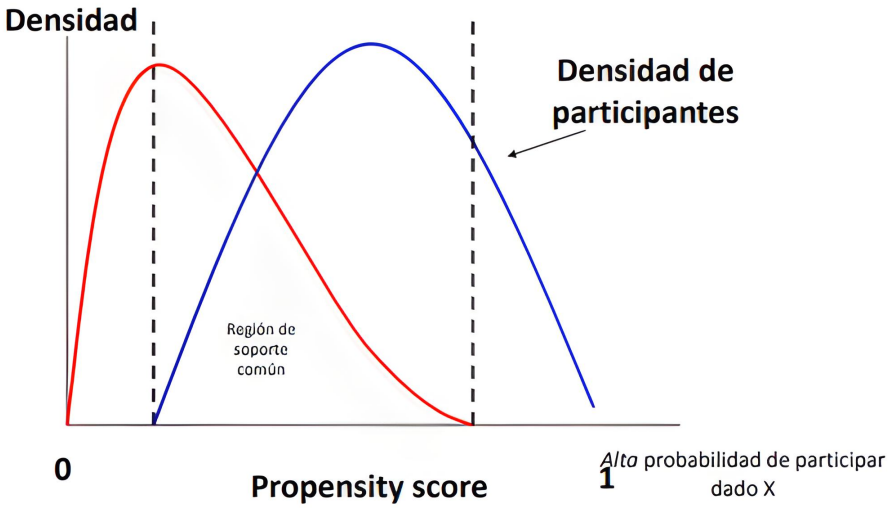


Figura 4.1: Condición de Superposición o soporte común.
Fuente: Tomada de BID [10]

Como se observa la *región de soporte común* determina aquellos individuos que sean participantes o no del programa pero que cuenta con probabilidades similares dadas sus características observables.

Bajo el supuesto de que se cumplan las condiciones en las expresiones (3.2) y (3.3), se obtiene el estimador del efecto promedio del programa para los intervenidos (ATT):

$$\text{ATT} = E(Y_0|X, D = 1) - E(Y_0|X, D = 0) \tag{4.4}$$

Donde:

$E(Y0X, D = 1)$ valor esperado de Y0 para individuos del grupo de tratamiento.

$E(Y0|X, D = 0)$ valor esperado de Y0 para individuos del grupo de control

Lo anterior, denota que se puede utilizar como contrafactual o grupo de control a los individuos que como resultado se clasifican en D = 0, si la diferencia entre ellos y los del D = 1 en promedio es cero [10].

Siguiendo la referencia de la Guía práctica para la evaluación de Impacto de [9], el procedimiento para la implementación del PSM cumple las siguientes condiciones:

### 4.1.3. Estimación de la Probabilidad de Participación en un programa

En esta etapa se utilizan modelos de respuesta binaria como Logit o Probit, donde es comúnmente utilizado el Modelo de probabilidad lineal que es la Regresión Logística propuesto por [11]:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \tag{4.5}$$

Donde:

$Y_i$ es la variable respuesta binaria
$B_0, B_k$ parámetros por estimar
$X_i k$ representa el conjunto de características observables del Individuo

Logrando estimar la probabilidad de participación para cada individuo.

### 4.1.4. Evaluar el cumplimiento de la condición de Soporte Común

Se verifica la condición establecida en la ecuación (3.3), lo cual se puede verificar por medio del análisis gráfico de las densidades de probabilidad de ambos grupos. Las autoras [9], sugieren algunos criterios para identificar el factor común:

- Eliminar las observaciones inferiores al mínimo y mayores al máximo del otro grupo
- Seleccionar variables de probabilidad estimada de participación, donde la densidad sea positiva en ambos grupos

### 4.1.5. Seleccionar los Algoritmos de Emparejamiento

#### Vecino más cercano

A partir de los resultados de cada persona, este algoritmo selecciona un individuo del grupo de intervención y busca dentro del conjunto de control el individuo con la probabilidad más cercana a partir del cálculo de la distancia euclidiana:

$$C(i) = \text{mín}\,||p_1 - p_j|| \tag{4.6}$$

Donde:

$p_1$ es la ubicación de un punto o un individuo
$p_j$ el vecino más cercano

C(i) representa el conjunto de unidades de control emparejadas a una unidad del grupo de tratamiento respectivamente y sobre esa base se estima el ATT así:

$$ATT = \sum (\text{Promedio}((Y_i|D=1) - (Y_c|D=0))) \tag{4.7}$$

Donde:

$(Y_i|D=1)$ conjunto del vecino de tratamiento
$(Y_c|D=0)$ conjunto del vecino de control

Para el caso de este algoritmo, los autores coinciden en que puede mejorar con muestras con reemplazo, pero estas generan mayor varianza y sesgos del estimador.

**Método Kernel**

Su algoritmo realiza un proceso de emparejamiento del grupo de tratamiento con un promedio ponderado para todos los individuos del grupo de control [9].

Este método emplea unas ponderaciones que son inversamente proporcionales a la distancia entre los propensity score de las unidades de tratamiento y control [12].

$$ATT = \sum \left( \frac{1}{n} \text{Promedio}((Y_i|D=1) - (Y_c|D=0)) \right) \tag{4.8}$$

Donde:

$\frac{1}{n}$ medida de normalización, donde n es total de individuos del grupo de tratamiento
$(Y_i|D=1)$ conjunto del vecino de tratamiento
$(Y_c|D=0)$ conjunto del vecino de control

Este algoritmo presenta menores niveles de varianza debido a que utiliza más información.

**Método de Estratificación**

Este algoritmo realiza la partición de las probabilidades estimadas, dividiendo las probabilidades de participación por estrato.

$$ATT = \text{Prom.}(Y_i|D=1) - \text{Prom.}(Y_c|D=0) \tag{4.9}$$

Donde:

$Prom.(Y_i|D=1)$ promedio de resultado de los individuos del grupo de tratamiento
$Prom.(Y_c|D=0)$ promedio de resultado de los individuos del grupo de control.

El impacto se calcula a partir de las diferencias promedio de la variable resultado entre los grupos de tratamiento y control.

### 4.1.6. Balanceo en las probabilidades

En este paso se evalúa la calidad del emparejamiento realizado, evaluado la distribución de las variables relevantes en ambos grupos.

$$Y_i = \beta_0 + \beta_1 \hat{P}(X_i) + \gamma_1 X_{1i} + \ldots + \gamma_k X_{ki} + u_i \tag{4.10}$$

Donde:

$Y_i$ es la variable respuesta binaria
$\beta_0, \beta_1$ parámetros por estimar
$\hat{P}(X_i)$ probabilidad estimada de un individuo recibir tratamiento
$\gamma_k X_{ki}$ representa el conjunto de características observables del Individuo
$Pu_i$ valor de error o residuo

Según [9], el emparejamiento está correcto cuando una vez después de controlar el tratamiento por P(Xi), este no está correlacionado con las características principales.

## 4.2. Antecedentes

A continuación, se enuncian algunos trabajos correspondientes a la implementación de métodos de emparejamiento como Propensity Score Matching (PSM) para evaluar la incidencia de factores socioeconómicos, que sirven como referencia para el presente proyecto:

El trabajo propuesto por [13], tiene por objetivo calcular la probabilidad de ubicación de un hogar en las comunas de Chile a partir de sus características socioeconómicas para determinar las diferencias en su georreferenciación, tomando como base los datos recopilados por la Encuesta de Hogares de 2003 y el Censo Nacional de Población y Vivienda 2002 realizado por el Instituto Nacional de Estadística de ese país.

Debido a que este proyecto incorporó los resultados de dos instrumentos, se tuvo que surtir un procesamiento para homogenizar algunos elementos como los códigos de identificación territorial que cambiaron entre una medición y la otra, sumado a la conciliación de las preguntas que pretendían capturar información socioeconómica en cada uno de los instrumentos. Adicionalmente, el proyecto desarrolló un proceso de ingeniería inversa para integrar las unidades de análisis personas, hogares y viviendas a través de una estructura relacional y posteriormente incorporar un sistema para georreferenciar. La implementación del matching espacial, permitió el emparejamiento de las bases para la creación de una muestra artificial con características homogéneas al grupo de control, arrojó como resultado de la georreferenciación, la heterogeneidad de las comunas, la autocorrelación espacial entre vecinos de comunas adyacentes y la interacción espacial intermunicipal lo cual es una gran contribución como variante a los modelos de estimación en áreas pequeñas.

En Perú se implementó el método de PSM para medir el impacto en el ingreso de los hogares con respecto al acceso a las tecnologías de la información y las comunicaciones – TIC [14], partiendo de una revisión literaria de como esta puede

incidir sobre factores como el crecimiento económico, la eficiencia productiva, el capital humano, el empleo y los ingresos para las personas. Tomando como referencia los resultados de la Encuesta Nacional de Hogares (ENAHO) del Instituto Nacional de Estadística e Informática – INEI del Perú para los años 2002 – 2006 en formato panel.

Para la implementación del modelo PSM se reordenan los datos de Panel convirtiéndose en una base de tipo corte, para la construcción del grupo de control se tomaron en cuenta algunos aspectos relevantes como las restricciones de oferta de estos servicios en algunas regiones del país, realizando el análisis contrafactual tanto de manera conjunta como separado por servicio (Fijo, móvil e internet.). Se realiza el proceso de agrupamiento partiendo de la predicción de una regresión probit para determinar la probabilidad de contar con acceso a las TIC. De manera paralela, se hizo una estimación de datos panel con efectos fijos para minimizar los potenciales sesgos generados en el emparejamiento dada la correlación y doble causalidad entre las TIC y el ingreso. Como resultado final, el modelo de PSM mostró que el acceso a TIC incrementa el ingreso per cápita de los hogares en $ 105 nuevos soles, de manera separada este incremento por servicio fue así: fijo ($ 19), móvil ($ 132) e internet ($365); mientras que el panel con efectos fijos arrojó una mejora conjunta en los ingresos de $ 216 nuevos soles y para el caso específico de telefonía móvil e internet un total de $ 28 y $ 104 nuevos soles respectivamente.

Otra aproximación interesante a la medición de impacto de las características socioeconómicas de la población sobre su bienestar corresponde al estudio para determinar los efectos del capital humano sobre la brecha de ingresos en Ecuador [15], realizando la descomposición de estas brechas o distancias por sexo e informalidad laboral aplicando los métodos Oaxaca - Blinder, luego de implementar el método PSM. Para este estudio se tomaron como referencia los datos de la Encuesta Nacional de Empleo, Desempleo y Subempleo del Instituto Nacional de Estadísticas y Censos – INEC de los años 2004, 2009 y 2014.

Midiendo los tres tipos de ingresos: salarios, ingresos autónomos y total ingresos, por medio de variables como edad, nivel de escolaridad y experiencia laboral. Adicionalmente, se incluyen variables de control tales como: región geográfica, sexo, región, etnia, sector económico, estabilidad laboral y estado civil. El modelo arrojó como resultado que las disparidades de ingreso afectan en principal medida a mujeres, indígenas, negras y trabajadores informales. El método de PSM indicó que la mayor parte de la brecha se explica por las dotaciones de cada individuo, lo cual se confirma con la descomposición salarial Oaxaca – Blinder que le quita fuerza al argumento de que esta desigualdad se explica únicamente por factores de discriminación.

Como la educación es un referente dentro de este proyecto, a través de un estudio realizado en España [16], se implementó PSM para medir el impacto de becas y ayudas de estudio sobre la continuidad de los estudiantes en el nivel secundario post-obligatorio, su modelo se construye a partir de los resultados de la Encuesta de Condiciones de Vida realizada por EUROSTAT con datos longitudinales para 2004-2006 y que fueron publicados en 2009, alcanzando un muestra de 58.740 españoles.

Para la construcción del modelo, se tomó como referencia el nivel educativo del

individuo a sus 19 años y se analizaron variables propias de la persona, de sus padres, su hogar y su entorno particularmente el grado de urbanización donde vive. Primero se implementa un modelo logit, para determinar la probabilidad de acceder a una beca y esta se explica por factores como: la posición entre los hermanos, la renta, la tenencia en propiedad de la vivienda y, por último, el lugar de residencia. Luego se analiza el efecto diferencial de haber accedido a una beca estimando el ATT, evaluado bajo los métodos vecino más cercano, estratificación, kernel y radius que indicaron que las becas aumentan en más de un 20 % la probabilidad de finalizar los estudios con éxito. Finalmente, se comprobó con la condición de soporte común que identificó que el 98,47 % de la base se encuentra dentro de esta área.

A nivel de Colombia un estudio publicado por el Banco de la República [17], mide el impacto del crédito agropecuario sobre el rendimiento de los cultivos y la pobreza multidimensional de los productores de este renglón de la economía nacional. Para este estudio, se tomó por primera vez como referencia los microdatos del Censo Nacional Agropecuario – CNA realizado por el DANE en el año 2014 el cual alcanzó una cobertura del 98,9 % de los 1.101 municipios del país, incorporando datos de otras fuentes desde la Unidad de Producción Agropecuaria y las características geográficas y sociodemográficas del productor.

Para la determinación de la variable de tratamiento, el modelo tomó en cuenta la respuesta de si le fue aprobado un crédito en 2013, no se tuvieron en cuenta los registros de personas que no solicitaron un crédito, limitando así la muestra para hacer más comparables los grupos de tratamiento y control. Se realizó la aplicación integral del PSM con la respectiva validación de los supuestos, como soporte común y el balanceo del emparejamiento, a lo cual se adicionó la estimación por OLS con Clúster que muestra algunas diferencias en los estimadores de impacto, pero no los contradice. Finalmente, el modelo logró determinar que cuando un productor accede a un crédito tiende a incrementar en promedio su rendimiento de cultivo en un 12 % y a disminuir su nivel de pobreza en 0,3 puntos porcentuales.

Otro estudio interesante corresponde a una aproximación para determinar los vínculos que existen entre las condiciones de empleo y la implementación del Sistema de Transporte Transcaribe en el área urbana de Cartagena [18], utilizó para este estudio los datos del Censo 2018, registros de las empresas y el empleo del Informe 500 empresas de Cartagena y la localización de las estaciones de Transcaribe, tomando como referencia el Marco Geoestadístico Nacional (2017).

Este estudio realizó la implementación de dos métodos, el primero de elección binaria, el cual no toma en cuenta factores como la endogeneidad, y el segundo de propensity score matching; las variables para el modelo de respuesta binaria si tiene o no empleo fueron: sexo, edad, lugar de nacimiento, nivel de escolaridad, estudiante, estrato, nivel de pobreza, distancia al centro de empleo más cercano, empresas del sector, vacantes, entre otras.

Dentro del PSM, se estima un modelo logit con la variable independiente distancia, la cual para el emparejamiento se traduce en 4 grupos de tratamientos separados, reduciendo relevancia de los coeficientes para este caso, pero usándolos como herramienta para calcular el efecto promedio de tratamiento, logrando un emparejamiento a través de los vecinos más cercanos donde además, se hizo la prueba de

heterogeneidad para medir si los efectos hallados se mantienen cuando los grupos de tratamiento se dividen en grupos más pequeños aplicándolos a sub muestras por sexo, estrato y grupo etario. Finalmente, el modelo determinó que la cercanía a una estación de Transcaribe incrementa la probabilidad de estar empleado y el acceso a buses de tránsito rápido en un 3 % y 8 % respectivamente.

La presente investigación [19], tiene por objetivo determinar las consecuencias sobre los ingresos de la paternidad o maternidad en los jóvenes menores de 21 años en Colombia, tomando como referencia los datos acumulados de enero a diciembre de 2013 de la Gran Encuesta Integrada de Hogares – GEIH del DANE, que para ese año incluyó un módulo sobre fecundidad aplicada a hombre de 14 a 60 años y mujeres de 12 a 55 años, incorporando cifras de nacimientos de las Estadísticas Vitales e información de la Registraduría Nacional.

Con el fin de evitar la evaluación de cohortes de población distintos, se seleccionaron personas que indicaron haber sido padres antes de los 21 años de una muestra de individuos con edades entre 25 y 35 años evitando los sesgos de medir generaciones distintas. Además, esta edad de 21 años, según el Ministerio de Educación es la edad en la cual un colombiano está culminando sus estudios de educación superior. Debido a que la maternidad se encuentra correlacionada a variables socioeconómicas, incluida el Ingreso identificando el posible problema de endogeneidad, la cual se pretende mitigar a través de la instrumentación de variables. Para la estimación se implementaron dos metodologías, primero el emparejamiento (PSM) para controlar el riesgo del sesgo de selección en variables observables y segundo, evaluando el impacto de un segundo hijo en edad temprana para incorporar también elementos no observables. Como resultado se evidencia que los individuos que son padres jóvenes tienen una penalidad de 11,7 % en sus ingresos, mientras que en las mujeres este efecto es del 12,7 % en los hombres alcanza el 5,3 %.

Otras aplicaciones de los algoritmos de Propensity Score Matching – PSM, pretende calcular un índice de precios espacial para la vivienda urbana en Colombia [20], tomando como referencia las 13 principales ciudades capitales del país que son: Bogotá, Barranquilla, Bucaramanga, Cali, Cartagena, Cúcuta, Ibagué, Manizales, Medellín, Montería, Pasto, Pereira y Villavicencio; como una apuesta ante la ausencia de un indicador capaz de medir el costo de vivienda en las diferentes áreas urbanas. El conjunto de datos proviene de la Gran Encuesta Integrada de Hogares – GEIH 2010, ya que dentro de esta existe información sobre las características físicas de la vivienda (número de dormitorios, materiales de construcción, clase de la unidad habitaciones, entre otras.).

Para el desarrollo del modelo, primero se realizan 3 pasos previos que son: evaluar la diferencia del precio promedio de la vivienda estándar en las ciudades, cuantificar los diferenciales de precios por ciudad y se construye el índice de precios donde la cesta de atributos varía por quintiles del precio. La estrategia econométrica implica la construcción del modelo en dos fases, primero, obtener muestras comparables entre ciudades a través del emparejamiento PSM y segundo, la aplicación de regresiones hedónicas para ambas muestras, logrando una mayor comparabilidad entre los bienes, la cual no refleje la variación de los atributos del bien sino los diferenciales de precios. Como resultado final se obtuvo que, las ciudades con mayor costo de las

viviendas son Bogotá, Cartagena y Villavicencio, cabe anotar que para el caso del arriendo en la capital de país este valor puede estar por encima del 30 % frente a las demás ciudades, así mismo, esta brecha también se evidencia entre las viviendas más costosas en Bogotá frente a las demás ciudades a excepción de Cartagena, donde muestra amplias similitudes dentro de este segmento con las viviendas de la capital de la república.

# Capítulo 5

# Procesamiento de la Base de Datos - CNPV 2018

## 5.1. Identificación de las principales características sociodemográficas y recopilación de la base de datos

Para el procesamiento de la base de datos del Censo Nacional de Población y Vivienda 2018, que corresponde al desarrollo del primer objetivo del proyecto, se inició con un proceso de comprensión y análisis de cómo se componen los microdatos del censo poblacional, su ámbito de aplicación para todos los municipios del territorio nacional, obteniendo información valiosa que se resume en una completa radiografía sobre la realidad sociodemográfica del país.

Los microdatos que componen el Censo Nacional están conformados por 4 grandes bloques que congregan distintos aspectos o características de una población en particular, los cuales se resumen en los siguientes módulos: Personas, Viviendas, Hogares y Marco de Georreferenciación - MGN. El ante-proyecto del presente trabajó contempló en un principio solo tener en cuenta el uso del módulo de **Personas** para obtener la información de cada una de las características sociodemográficas y con estas determinar la localización de un individuo en el universo geo censal del departamento del Valle del Cauca; no obstante, a través de un entendimiento profundo de cada uno de los módulos del censo como instrumento estadístico, se logró identificar a priori dentro de ellos (Vivienda, Hogares y MGN) algunas características que podrían ser relevantes para predecir la localización, por lo tanto, se determinó contemplarlos para la conformación del conjunto de datos y su procesamiento.

Posteriormente, se cargaron los conjuntos de datos para cada uno de los 4 módulos del Valle del Cauca, y se realizó un proceso de limpieza y depuración consistente en la eliminación de variables, pues la mayoría de los registros para estas variables correspondían a valores nulos, los cuales se presentan por módulo a continuación:

1. **Personas:** Elementos de caracterización para la población indígena, condiciones de acceso y calidad en la prestación de servicios de salud, y lo correspondiente al nacimiento de hijos (vivos, sobrevivientes) y si estos residen o no en el país.

2. **Marco de Georreferenciación:** Factores que determinan distintos aspectos de la localización, los cuales se encuentran representados en la variable Código DANE ANM, que es el elemento de representación geográfica del Censo.

3. **Hogares:** Total de personas fallecidas en el Hogar.

4. **Vivienda:** Característica de las viviendas ubicadas en lugares especiales de alojamiento – LEA como centros penitenciarios, albergues infantiles u hogares geriátricos y guarniciones militares; ubicación de la vivienda en un territorio indígena y codificación de las unidades habitacionales en áreas protegidas.

Para el caso de las demás variables se eliminaron los registros que contaban con pocos valores nulos para no depurar otras características que tuvieran información relevante.

Una vez depurada la base, se procedió a analizar la variable geográfica del Censo que se utilizó para predecir la localización de un individuo denominada **COD DANE ANM**, la cual corresponde a un valor categórico con 22 posiciones que resume en cada uno distintos aspectos, tal y como se muestra a continuación:

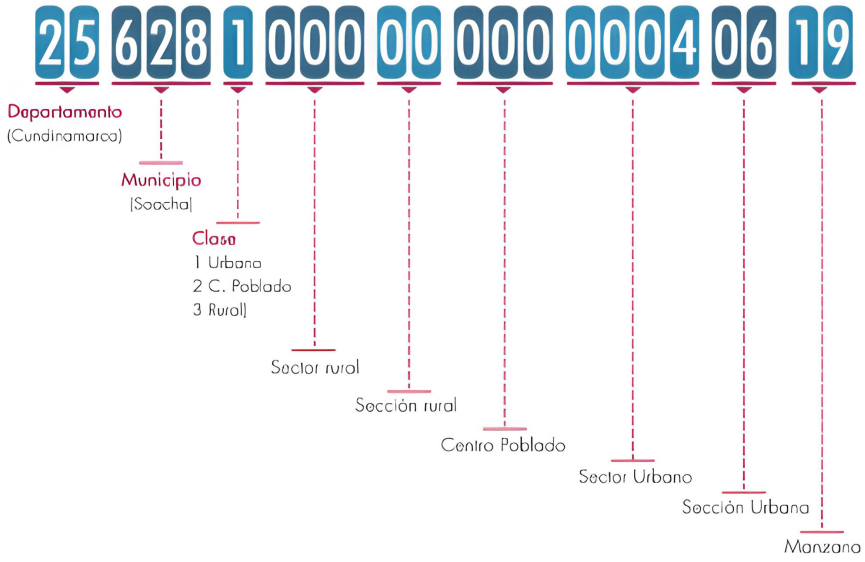


Figura 5.1: Conformación del Código DANE ANM
Fuente: Tomada de DANE [21]

Para el tratamiento de la variable geográfica, se creó una nueva denominada COORD, la cual contiene información de este elemento geográfico desde la clase en adelante. Su tratamiento se presentará al final cuando se prediga el valor y se utilice con los nuevos datos.

Dando como resultado de este proceso lo siguiente:

**Módulo de Personas:** 3,3 millones de registros para el departamento, con un total de 48 variables que resumen aspectos como: la ubicación en el municipio y zona, sexo, edad, pertenencia étnica, lugar de nacimiento y residencia en los últimos años, alfabetismo y máximo nivel de escolaridad logrado, ocupación laboral y estado civil.

**Módulo del Marco de Georreferenciación:** 1,4 millones de registros para el departamento, con un total de 4 variables seleccionadas correspondientes a: Código de la encuesta como variable llave para concatenar con todos los módulos del Censo, Código DANE ANM que representa un valor categórico similar a una clasificación geo espacial, el valor de la manzana censal como unidad geográfica más específica contenida dentro de los microdatos, y por último la variable Coordenada (COORD), la cual es una extracción de los componentes geográficos del código ANM, siendo estas las características geográficas a predecir.

**Módulo de Hogares:** 1,2 millones de registros para el departamento, con un total de 7 variables que contienen elementos tales como: el número de hogares y personas, la cantidad de cuartos y cuantos de estos se destinan exclusivamente como dormitorio, el lugar y la fuente de agua para la preparación de los alimentos.

**Módulo de Vivienda:** 1,4 millones de registros para el departamento, con un total de 17 variables que contienen aspectos de la vivienda como: su ubicación en área protegida, uso de la unidad habitacional, tipo de vivienda y condición de ocupación, materiales en paredes y pisos, estrato socioeconómico, cobertura en todos los servicios públicos como acueducto, alcantarillado, energía, gas, internet y recolección de basuras, entre otras.

A posteriori, se consolidó el conjunto de datos por medio del código de encuesta, lo que permitió relacionar los individuos del módulo personas con las demás características provenientes de los otros módulos; adicional, previendo las posibles dificultades de localización de la población en las zonas rurales dispersas, se priorizó el análisis en las cabeceras municipales y solo se dejó un caso en centros poblados, obteniendo un conjunto de datos de **3,2 millones** de registros u observaciones con un total de **40 variables**, las cuales se presentan a continuación:

1. COD_ENCUESTAS: Código de Encuesta
2. COD DANE ANM: Código de Localización
3. U_MZA: Manzana
4. UA_CLASE: Cabecera Municipal (1), Resto (2)
5. P_NRO_PER: Número de persona en el hogar
6. P_SEXO: Sexo
7. P_EDADR: Edad en grupos Quinquenales
8. P_PARENTESCOR: Relación de parentesco con el Jefe (a) del Hogar
9. P_EST_CIVIL: Estado Civil
10. PA1_GRP_ETNIC: Reconocimiento étnico
11. PA_LUG_NAC: Lugar de Nacimiento
12. PA_VIVIA_5ANOS: Lugar de residencia hace 5 años

13. PA_VIVIA_1ANO: Lugar de residencia hace 12 meses
14. P_ALFABETA: Sabe leer y escribir
15. PA_ASISTENCIA: Asistencia escolar (de forma presencial o virtual)
16. P_NIVEL_ANOSR: Nivel educativo más alto alcanzado y último año o grado aprobado en ese nivel
17. P_TRABAJO: Que hizo durante la semana pasada
18. H_NROHOG: Número del hogar de la vivienda
19. H_NRO_CUARTOS: Número de cuartos en total
20. H_NRO_DORMIT: Número de cuartos para dormir
21. H_DONDE_PREPALIM: Lugar donde prepara los alimentos
22. H_AGUA_COCIN: Fuente de agua para preparar los alimentos
23. HA_TOT_PER: Total de personas en el hogar
24. UVA_ESTA_AREAPROT: Vivienda en un área protegida
25. UVA_USO_UNIDAD: Uso de la Unidad
26. V_TIPO_VIV: Tipo de Vivienda
27. V_CON_OCUP: Condición de ocupación
28. V_TOT_HOG: Total de hogares en la Vivienda
29. V_MAT_PARED: Material predominante en paredes de la vivienda
30. V_MAT_PISO: Material predominante en los pisos
31. VA_EE: Servicio de Energía Eléctrica
32. VA1_ESTRATO: Estrato de la Vivienda (Según servicio de energía)
33. VB_ACU: Servicio de Acueducto
34. VC_ALC: Servicio de Alcantarillado
35. VD_GAS: Servicio de Gas Natural conectado a Red Pública
36. VE_RECBAS: Servicio de recolección de Basura
37. VE1_QSEM: Cuantas veces por semana (recolección de Basura)
38. VF_INTERNET: Servicio de Internet (fijo o móvil)
39. V_TIPO_SERSA: Tipo de Servicio Sanitario (inodoro)
40. TREATMENT: Variable binaria (tratamiento y control)

Y, por último, para el desarrollo del Propensity Score Matching - PSM se debe configurar una población a distribuir entre grupo de tratamiento y control, se crearon los siguientes conjuntos de información que se analizaron en el presente trabajo:

**Conjunto 1 – Valle Urbano:** Contiene los datos sobre las cabeceras de los municipios que componen el Valle del Cauca, se divide el conjunto en Cali para tratamiento y el resto de los municipios el grupo de control.

**Conjunto 2 – Cali Urbano:** Se compone de datos exclusivamente de la cabecera municipal y se construyen los grupos de tratamiento y control de manera aleatoria.

**Conjunto 3 – Cali Urbano y Centros Poblados:** Está compuesto con registros solamente de Cali, distribuyendo como grupo de tratamiento la información de la cabecera municipal y en control lo concerniente a los centros poblados.

### 5.1.1. Evaluar la posibilidad de identificar otras factores o características relevantes que se ajusten a los requisitos de la técnica seleccionada

Con el objetivo de robustecer el estudio se analizaron otras fuentes de información socioeconómica distintas al Censo que existen en el país, una de las más completas es la base del **Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales – SISBEN** , la cual es utilizada por el departamento nacional de planeación – DNP para la priorización y selección de los beneficiarios de programas sociales. Este instrumento está compuesto por 3 módulos, Vivienda, Hogares y Personas, los cuales contienen factores socioeconómicos asociados al nivel de educación, ingreso, salud, condiciones de vivienda, junto con los distintos elementos que conforman el índice de pobreza multidimensional – IPM

Una vez analizado este conjunto de datos frente a la base inicial del Censo DANE 2018, se encontró que debido al proceso de anonimización que surgen en este tipo de instrumentos, no fue posible hallar una variable llave, como, por ejemplo, el número de identificación como un carácter único, que empalmara la base del SISBEN al conjunto procesado del Censo. Adicionalmente, existieron diferencias en las variables de localización de dichos instrumentos, mientras que en el censo la unidad mínima es la Manzana, en el SISBEN corresponde a nivel de Barrio.

Debido a la dificultad presentada, se decidió utilizar la información del SISBEN como el conjunto de nuevas observaciones sobre el cual se estimó la localización para estos individuos, no obstante, atendiendo las limitaciones de información para algunos municipios, se utilizó los datos solo para la ciudad de Cali. Adicionalmente, se encontró la diferencia de algunas variables, que, aunque se ajustaron al modelo, no se pudieron contemplar debido a que estaban presentes en el Censo, pero no en el SISBEN, como fue el caso de la pertenencia étnica de la población.

# Capítulo 6

# Análisis Exploratorio

## 6.1. Exploración de Datos

En el desarrollo del segundo objetivo específico, y mediante una comprensión profunda de cada una de las variables resultantes del conjunto seleccionado, se realizó un análisis exploratorio que permitió elegir las características sociodemográficas más relevantes; este proceso inició con la revisión del Diccionario de los Microdatos que definió cada uno de los valores que componen una variable, como en este caso, el nivel educativo alcanzado por cada uno de los individuos.

| **No.** | **Descripción** |
|---|---|
| 1 | Preescolar |
| 2 | Básica primaria |
| 3 | Básica secundaria |
| 4 | Media académica o clásica |
| 5 | Media técnica |
| 6 | Normalista |
| 7 | Técnica profesional o tecnológica |
| 8 | Universitario |
| 9 | Especialización, maestría, doctorado |
| 10 | Ninguno |
| 99 | No informa |

Cuadro 6.1: Categorías Variable Nivel educativo más alto alcanzado

Fuente: Elaboración propia con base en el script de Colab

Posteriormente, se construyó una tabla resumen que contiene el número de observaciones y la participación porcentual de cada una de las características en las variables, con el fin de evidenciar las más representativas; adicional se realizó un análisis gráfico de todas las variables de la base para lograr una mejor representación de las características y los valores que adoptan.

| P_NIVEL_ANOSR | No. | % |
|---|---|---|
| 4.0 | 798.625 | 31.47 |
| 2.0 | 557.101 | 21.96 |
| 3.0 | 464.438 | 18.30 |
| 6.0 | 309.289 | 12.19 |
| 5.0 | 268.848 | 10.60 |
| 0.0 | 69.925 | 2.76 |
| 7.0 | 67.734 | 2.67 |
| 1.0 | 1.404 | 0.05 |

Cuadro 6.2: Distribución de la Variable Nivel educativo más alto alcanzado

Fuente: Elaboración propia con base en el script de Colab

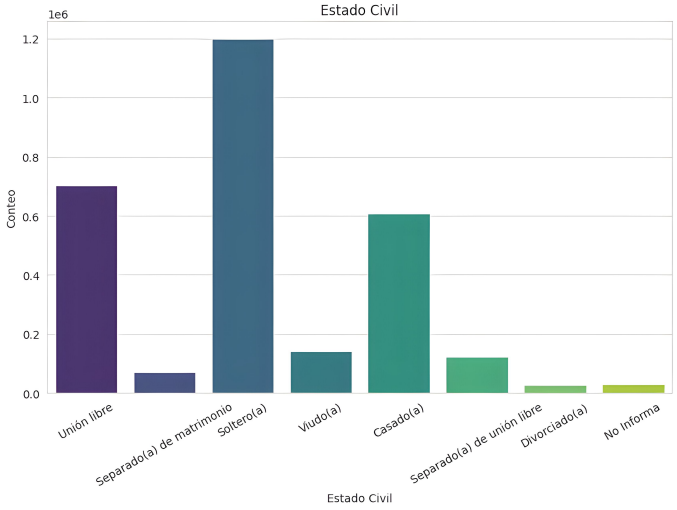


Figura 6.1: Distribución de la Variable Estado Civil
Fuente: Tomada del script de Colab.

Una vez revisadas todas las características que componen la base de datos, se logró identificar y eliminar algunas variables que por su naturaleza no aportaban ningún valor al proceso de modelado, puesto que estaban representadas en su gran por una característica única, como fue el caso del tipo de servicio sanitario, la ubicación de la vivienda en un área protegida, entre otras.

Seguidamente, se logró la identificación de varias características que estaban descritas bajo el rótulo de No informa, representadas por un valor bien definido, imposibilitando su depuración como valor nulo, pero que claramente representaba un elemento que causaría desviaciones en los resultados de la predicción del modelo. Por lo tanto, se procedió a convertir estas características en valores nulos y así poderlos eliminar; este proceso se realizó para 13 variables, como Estado civil, Lugar de Nacimiento, Ocupación Laboral, entre otras.

Luego de la depuración de las características sin información, se revisó la distribución de las clases de tratamiento (1) y control (0) para cada uno de los conjuntos

de datos analizados, identificando en todos unos problemas de desbalanceo de características, lo cual podría generar sesgos en la predicción de modelos de aprendizaje automático, por tanto, se procedió a implementar un Under- Sampling [22] para balancear la distribución de las características, creando un nuevo conjunto de datos, sobre el cual se construyó el objeto RandomUnderSampler que se aplica al conjunto de datos y se construyen un nuevo DataFrame balanceado, como se presenta a continuación:

| Conjunto | Grupo | Original | UnderSampler |
|---|---|---|---|
| 1 | Tratamiento | 1.609.504 | 1.268.682 |
| | Control | 1.298.772 | 1.268.682 |
| 2 | Tratamiento | 805.175 | 782.114 |
| | Control | 804.331 | 782.114 |
| 3 | Tratamiento | 1.609.504 | 28.738 |
| | Control | 29.191 | 28.738 |

Cuadro 6.3: Clases de la variable TREATMENT en los 3 conjuntos de Datos

Fuente: Elaboración propia con base en los resultado del script de Colab

Con los conjuntos de datos balanceados, se replicó el ejercicio de análisis exploratorio de datos y adicional se realizó un proceso de parametrización de las variables del Censo que resultaron seleccionadas en el modelo de clasificación del PSM, las cuales lo requirieron dado que en comparación con la misma característica que contiene la base del SISBEN, adoptan valores distintos para describir la misma variable como fue el caso del nivel educativo más alto alcanzado, situación laboral y el estado civil.

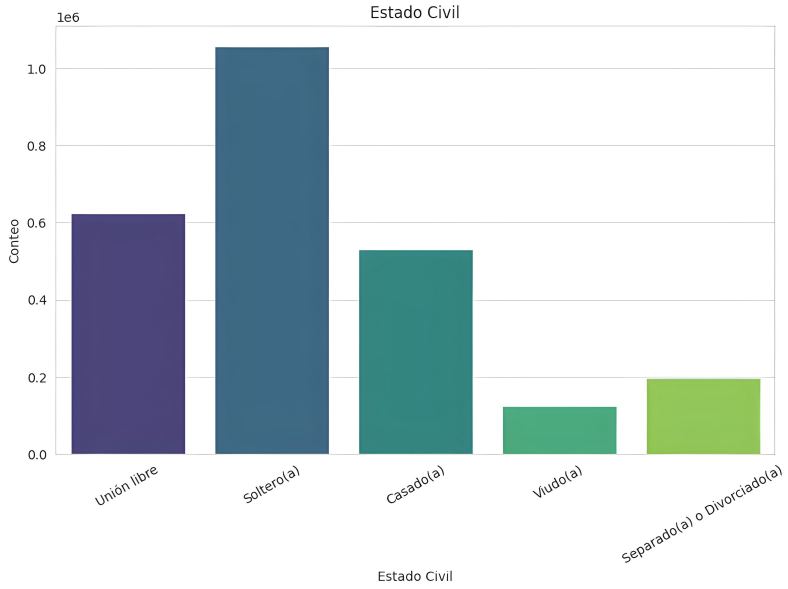


Figura 6.2: Distribución de la Variable Estado Civil Parametrizado
Fuente: Tomada del script de Colab.

# Capítulo 7

# Implementación y Evaluación del Modelo PSM

## 7.1. Propensity Score Matching - PSM

El presente capítulo contiene el desarrollo de los objetivos 3, 4 y 5 del presente trabajo, los cuales consistieron en la implementación del Modelo PSM, su posterior evaluación y afinamiento para obtener los mejores resultados del emparejamiento y la localización de los individuos acorde con las características sociodemográficas seleccionadas.

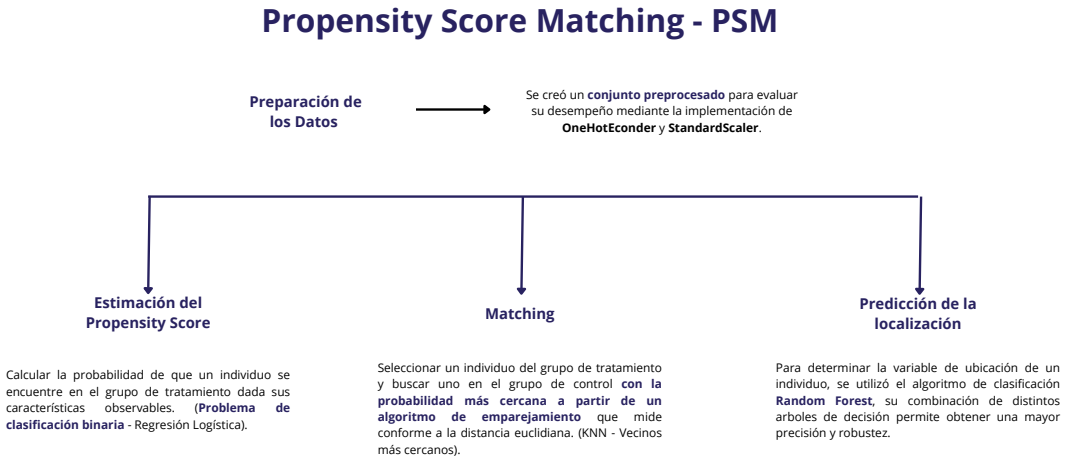


Figura 7.1: Secuencia de implementación del Propensity Score Matching - PSM
Fuente: Elaboración propia con base en [8].

La secuencia de pasos que recorren los distintos conjuntos de datos para la implementación del Modelo PSM se observa en la Figura 7.1. y sus pasos se exponen aqui:

Previo a la implementación, con el objetivo de probar la mejor forma de incorpo-

rar los datos al modelo dada sus características, se realizó un proceso de preparación de los datos, donde al conjunto original se le realizó un preprocesamiento aplicando técnicas de OneHotEncoder y StandardScaler y utilizando ambos conjuntos para evaluar el de mayor desempeño en el PSM.

La aplicación del algoritmo de PSM inicia con la estimación del puntaje de propensión o propensity score, que corresponde a la definición del problema de clasificación binaria, es decir, determinar a qué grupo pertenecen los individuos de acuerdo a la distribución para cada uno de los conjuntos de datos estructurados, mediante la implementación de una Regresión Logística para calcular la probabilidad de que un individuo del grupo se encuentre en el grupo de tratamiento a partir de sus características observaciones (Condiciones sociodemográficas del individuo), obteniendo aquellos individuos que se ubican en una región de soporte común, lo que se traduce en que sus muestras son comparables.

Posteriormente, se realiza el proceso de emparejamiento por medio del algoritmo de Vecinos más cercanos – KNN para calcular la distancia euclidiana entre un individuo del grupo de tratamiento y uno del grupo de control, de acuerdo con sus niveles de probabilidad. Finalmente, una vez se cuenta con el conjunto de datos emparejado, se procede a la estimación de la variable de localización que corresponde a la manzana censal por medio de un Ramdon Forest y así obtener la estimación de este variable geográfica que nos permite ubicar a los individuos en el espacio geo censal.

A continuación, se presentan cada uno de estos acápites a detalle:

### 7.1.1. Preparación de los Datos

Para la implementación de los algoritmos de aprendizaje automático del PSM, se realizó la estimación del modelo sobre el conjunto original y sobre otro preprocesado, para este caso, se aplicaron dos técnicas ampliamente utilizadas para mejorar la eficiencia de los procesos de Machine Learning, los cuales se resumen a continuación:

- La primera fue el *OneHotEncoder*, el cual convierte las variables categóricas en una clasificación numérica, transformándolas en columnas binarias separadas, donde 1 indica si el valor está representado en la instancia y 0 en el caso contrario [23]. Para el caso de los datos del modelo, fue útil dado que las variables seleccionadas estaban representadas en su mayoría por elementos binarios (Ej.: Cobertura de servicios públicos), excepto factores tales como las condiciones laborales y el estado civil de la persona, que generaron incrementos en la dimensionalidad y un potencial riesgo de sobreajuste.

- La segunda corresponde al *StandardScaler*, como un proceso de normalización o estandarización de las variables numéricas, para que se distribuyan dentro de una media cero y una desviación estándar de 1 [24].Este procedimiento se implementó para las variables de estrato socioeconómico y edad, lo cual generó estabilidad y equidad en las características, aunque representó cierta dificultad, pues al realizar esta transformación se presentó una pérdida en la interpretabilidad de dichas características.

Los resultados de este proceso se verán reflejados en la comparación del modelo con ambos conjuntos de datos en la siguiente etapa.

### 7.1.2. Implementación y evaluación del modelo

La implementación del Modelo de Propensity Score Matching – PSM, consta de dos pasos fundamentales. El primero corresponde a un proceso de clasificación dentro del aprendizaje supervisado, para encontrar como lo menciona [9], ese clon de cada individuo tratado en el conjunto de control y contrastar sus variables de resultado. El segundo momento, es la selección de un algoritmo de emparejamiento, que contraste la variable de resultado del individuo tratado con los individuos más parecidos dentro del conjunto de control.

Previo al proceso de clasificación, mediante el Modelo de Regresión Lineal Múltiple, se realizó una prueba piloto de la significancia de las variables del conjunto de datos con respecto a la variable Manzana (U MZA), como la unidad mínima de georreferenciación de las personas, obteniendo como resultado la depuración de las siguientes variables: *GRP ETNIC y P ALFABETA*, las cuales no eran significativas según su p – valor; *PA LUG NAC, PA VIVIA 5ANOS, PA VIVIA 1ANO*, por su correlación geográfica y las características de *VE RECBAS y VA EE*, ya que su cobertura era total y no mostraban ningún elemento diferenciador al momento de predecir la localización.

#### Estimación del Propensity Score

El PSM es un método cuasi experimental, es decir, por su naturaleza no se construye a partir de un experimento aleatorio, sino que se realiza mediante la construcción de un grupo de control artificial, que hace coincidir a las unidades tratadas con aquellas que no, dadas unas características observables. En este sentido, se calcula el propensity score que representa la probabilidad de que un individuo se encuentre en el grupo de tratamiento dado las covariables o características observables [25].

Lo anterior, se traduce en un problema de clasificación binaria, el cual se resolvió mediante la implementación del Modelo de Regresión Logística [26], como una técnica ampliamente utilizada en los procesos de clasificación, funciona muy bien con conjuntos de datos robustos como los Microdatos del Censo que se utilizaron en el presente trabajo, los cuales quedaron balanceados en sus características. En contraste, una de las desventajas que presenta este modelo es que asume la linealidad de las variables, lo cual es previsible que no ocurriera con los datos del Censo y adicionalmente, ante conjuntos de alta dimensionalidad puede presentar overfitting.

Para ejecutar el Modelo se realizaron las siguientes acciones: se separó el conjunto de datos en tratamiento y control, se probaron las distintas características para predecir la variable TREATMENT, se seleccionaron las características más relevantes, una vez estimada la regresión logística y su predicción, se realizó el cálculo del propensity score (ps) y su respectiva transformación a una escala logarítmica para su normalización. Este procedimiento se iteró para cada uno de los conjuntos de datos, a continuación, se presentan los resultados:

Los conjuntos de datos 1 y 2, que representan a Valle Urbano y Cali Urbano respectivamente, son los más grandes dentro de los datasets analizados. Se observa en

la gráfica de la condición de soporte común que existe una importante superposición en los valores de probabilidad en cada uno de los grupos, tanto en tratamiento como en control; para el Conjunto 2, se aprecia un comportamiento similar a una distribución normal en los propensity score obtenidos, mientras que el Conjunto 3, muestra una dispersión en las probabilidades, pero con una alta superposición entre los dos grupos de control y tratamiento.

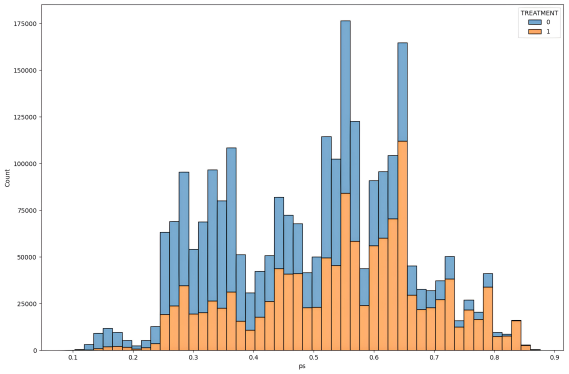


Figura 7.2: Propensity Score - Conjunto de Datos 1. Valle Urbano
Fuente: Tomada del script de Colab.

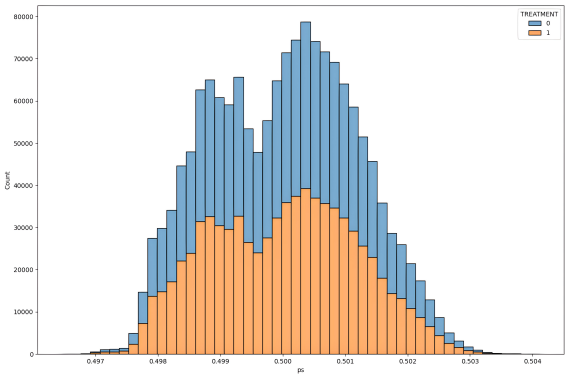


Figura 7.3: Propensity Score - Conjunto de Datos 2. Cali Urbano
Fuente: Tomada del script de Colab.

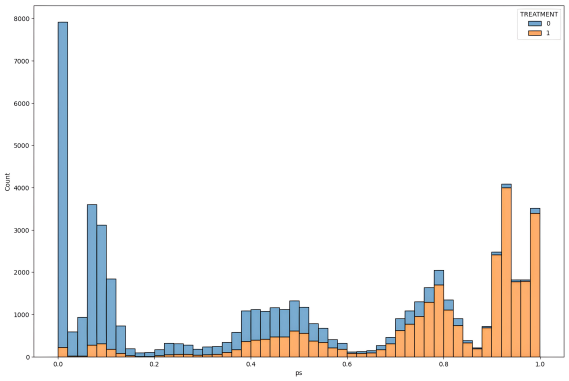


Figura 7.4: Propensity Score - Conjunto de Datos 3. Cali Urbano y Centros Poblados
Fuente: Tomada del script de Colab.

| Conjunto | Grupo | Precision | Recall | F1 - Score | Accuracy |
|---|---|---|---|---|---|
| 1 | Trat. | 0.62 | 0.56 | 0.59 | **0.61** |
| | Contr. | 0.60 | 0.66 | 0.63 | |
| 2 | Trat. | 0.50 | 0.45 | 0.47 | **0.50** |
| | Contr. | 0.50 | 0.55 | 0.52 | |
| 3 | Trat. | 0.83 | 0.85 | 0.84 | **0.84** |
| | Contr. | 0.85 | 0.83 | 0.84 | |

Cuadro 7.1: Resultados Modelo de Regresión Logística en los 3 conjuntos de Datos Fuente: Elaboración propia con base en los resultados del script de Colab

Con respecto a los resultados comparativos que se observan en la tabla anterior, como primera medida revisamos el Accuracy, que mide la proporción de predicciones correctas, donde el Conjunto 2 - Cali Urbano obtuvieron un valor del 50 %, mientras que el Conjunto 1 - Cali y los demás municipios del Valle alcanzaron un valor de 61 %, el Conjunto 3 – Cali Urbano y Centros Poblados lograron un valor de 84 %; al evaluar este último resultado que es el de mejor desempeño, se resalta que de los tres conjuntos de datos este es el más pequeño, lo cual infiere claramente un proceso de sobreajuste. Por tanto, se descartó este conjunto analizado y se continuó el proceso con los otros dos datasets estudiados.

Para el caso de los conjuntos de datos 1 y 2, se realizó la predicción del modelo de regresión logística con los conjuntos preprocesados utilizando las técnicas OneHotEncoder y con el StandardScaler, mostrando unos valores de Accuracy de 61 % y 50 %, los mismos de los conjuntos originales, por tanto, en vista de que no obtuvieron un mejor resultado y que seleccionarlos implicaría una dificultad en la interpretabilidad de los modelos al final del proceso, estos fueron descartados del análisis.

Continuando con el proceso de análisis de los resultados de las métricas de desempeño para los conjuntos de datos 1 y 2 en su versión original, se aprecia lo siguiente:

- **Precision:** El conjunto 1 muestra una mayor precisión que el conjunto 2, tanto en el conjunto de tratamiento (0,62 frente a 0,50) como en control (0,60 frente a 0,50).
- **Recall:** El conjunto 1 es mejor que el conjunto 2, al momento de encontrar todos los casos positivos tanto en tratamiento (0,56 frente a 0,45) como en control (0,66 frente a 0,55).
- **F1 – Score:** Que representa una media armónica de los resultados de recall y precision, denota un mejor desempeño del Conjunto 1 con respecto al conjunto 2.

Finalmente, a partir de los resultados obtenidos en el cálculo del propensity score, el estudio continuó con estos dos conjuntos de datos (1 y 2) y pasan al proceso de matching que se expone en el siguiente ítem.

**Emparejamiento**

Para el Matching se realizó la implementación del *Vecinos más cercanos – KNN* [27], que es un modelo no paramétrico, es decir, que este no intenta ajustar los datos a una forma funcional especifica, lo que indica que las predicciones sobre el conjunto de prueba dependen de los datos de entrenamiento. Este algoritmo es sencillo de entender e implementar, es muy efectivo con un número adecuado de datos y una elección optima de los K vecinos a identificar, su principal desventaja es el costo computacional que genera con conjuntos de datos grandes.

Para ejecutar el Modelo KNN, se definieron los parámetros de K = 2 vecinos y el caliper que corresponde a la distancia máxima permitida para considerar dos puntos como vecinos, que en este caso fue del 0,25 de la desviación estándar del propensity score, como una medida estricta para obtener una correspondencia más precisa entre los puntos analizados. Posteriormente, se calcularon las distancias entre los vecinos más cercanos, se realizó un ciclo para identificar que, aplicando el principio de muestras sin reemplazo, por cada punto en tratamiento se encuentre un punto en control, lo que indica que una vez emparejados estos individuos no se utilizarán para emparejar a otras personas. Luego, se identificaron las observaciones en tratamiento y aquellas emparejadas en control, dejando solo las observaciones que fueron emparejadas, para así calcular las diferencias entre los conjuntos de tratamiento y control antes y después del emparejamiento, obteniendo los siguientes resultados:

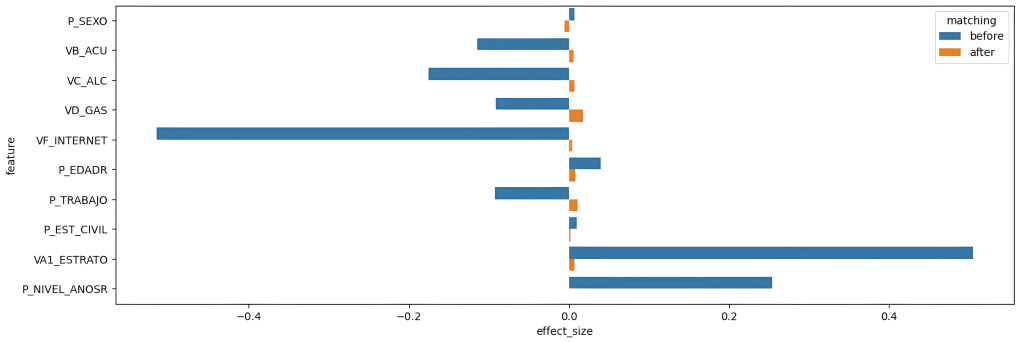


Figura 7.5: Diferencias antes y después del Emparejamiento - Conjunto 1
Fuente: Tomada del script de Colab.

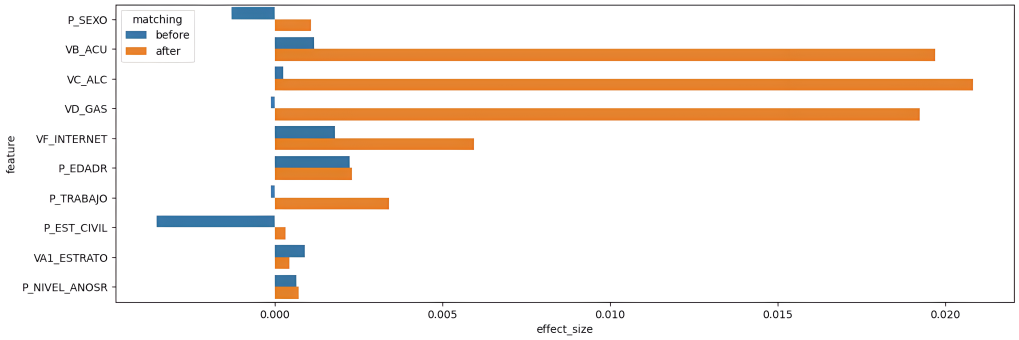


Figura 7.6: Diferencias antes y después del Emparejamiento - Conjunto 2
Fuente: Tomada del script de Colab.

Como se observa en las figuras anteriores, luego del proceso de emparejamiento en el Conjunto 1, todas las categorías se encuentran más equilibradas, mientras que en el caso del Conjunto 2, ocurrió lo opuesto, tras el matching las categorías muestran un mayor desequilibro frente a como se encontraban antes de implementar este proceso. Por tanto, para el proceso de predicción se continuó con el Conjunto de datos 1 - Valle Urbano, que contiene la información para Cali y los demás municipios.

### 7.1.3. Predicción de la localización

Para el proceso de estimación de la ubicación de los individuos a partir de las características sociodemográficas seleccionadas, se utilizó el algoritmo de clasificación al Random Forest [28], es un modelo que combina distintos árboles de decisión para mejorar su precisión y robustez, donde cada árbol es un modelo que divide iterativamente el conjunto de datos en subconjuntos, y bajo un umbral de decisión maximiza los proceso de clasificación o regresión, ya que funciona muy bien para ambos procesos de aprendizaje supervisado.

Estos árboles manejan muy bien predictores tanto numéricos como categóricos, así que no siempre requieren para su estimación preprocesamientos, tal y como sucede con la técnica de OneHotEncoder y al ser un método no paramétrico los datos no requieren tener una distribución especifica. En contraste, una de sus mayores desventajas y que se evidenció en este proyecto es su alto costo computacional, el cual obligó a limitar su desempeño para lograr la estimación.

Para esta etapa, el mayor reto fue el cálculo del modelo de Clasificación que permitiera determinar la ubicación exacta de los individuos del Conjunto 1, es decir, la población urbana de Cali y los demás municipios del departamento, tomando como referencia el Código ANM DANE, una variable de tipo objeto que contiene 22 caracteres donde resume todas las instancias geográficas, desde el código del departamento hasta la manzana específica donde se localiza el individuo, este tipo de datos por su naturaleza y magnitud representó una dificultad para el proceso de estimación pues provocó un alto consumo de memoria que impedía avanzar en el modelo.

De lo anterior se infiere que, por el alto costo computacional que dificultaba la estimación del modelo para el total del conjunto de datos 1, se determinó realizarla solo para la ciudad de Cali, pues representaba la mitad del set de datos, además, facilitaba la predicción y los elementos asociados a un código ANM.

Esto no ocurría al momento de entrenar el modelo de Cali y todos los municipios con los nuevos datos del SISBEN, que solo representaban individuos de la ciudad de Cali, por tanto, con esto se evitó el riesgo de predecir datos para todos los municipios y que, al momento de implementar con muestras de la capital vallecaucana, este los ubicara en otros municipios.

Para la implementación del Random Forest se realizaron las siguientes acciones: se seleccionaron las características y la variable objetivo de la clasificación, se dividieron los datos en el conjunto de tratamiento y control, para la estimación del modelo se ajustaron una serie de hiperparámetros que lograran un adecuado balance entre una buena predicción y la optimización del uso de la memoria, detallas así:

- **n_estimators = 100**: Indica el número de árboles con el cual se construyó el bosque.
- **max_depth = 10**: Determina la profundidad máxima para cada árbol, esto permitió prevenir procesos de sobreajuste.
- **max_features = 'sqrt'**: Define el número máximo de características a considerar para dividir un nodo.
- **min_samples_split = 10**: Define el número mínimo de muestras requeridas para dividir el nodo interno; si es más alto puede generar sobreajuste.
- **min_samples_leaf = 4**: Especifica el número mínimo de muestras que toma un nodo hoja, reduce que el modelo aprenda ruido del conjunto de entrenamiento.
- **n_jobs = -1**: Menciona el número de procesos que se ejecutan en paralelo en entrenamiento y predicción; en este caso, indica que se utilizan todos los núcleos de CPU disponibles.

Una vez entrenado el modelo, para evitar fallos por exceder la capacidad de memoria, se realizó la predicción por partes, y finalmente se obtuvo la métrica de desempeño adaptando tres alternativas del Código o ubicación geográfica a predecir, el de la Figura 7.7, corresponde al resultado con el Código DANE ANM en su versión original alcanzando un accuracy de 32 %; una segunda estimación tomando como referencia una extracción de la variable a predecir con los últimos 8 caracteres del campo original de 22, eliminando los primeros elementos que eran similares, para este caso de solo Cali, logrando un accuracy de 39 %. Y por último, presumiendo posibles falencias en el dato original del Código ANM por los valores en cero de la parte central del carácter, se eliminaron estos y se realizó la tercera estimación obteniendo como resultado un accuracy de 32 %.

| | | | | |
|---|---|---|---|---|
| 76001100000000021060601 | 1.00 | 0.00 | 0.00 | 40 |
| 76001100000000021060603 | 1.00 | 0.00 | 0.00 | 31 |
| 76001100000000021060605 | 0.35 | 0.56 | 0.43 | 25 |
| 76001100000000021060607 | 0.28 | 1.00 | 0.43 | 41 |
| 76001100000000021060609 | 0.45 | 0.57 | 0.50 | 23 |
| 76001100000000021060611 | 0.10 | 1.00 | 0.18 | 24 |
| 76001100000000021060613 | 1.00 | 0.00 | 0.00 | 32 |
| 76001100000000021060614 | 1.00 | 0.00 | 0.00 | 36 |
| 76001100000000021060616 | 0.12 | 1.00 | 0.21 | 32 |
| 76001100000000021060617 | 1.00 | 0.00 | 0.00 | 41 |
| 76001100000000021060618 | 1.00 | 0.33 | 0.50 | 27 |
| 76001100000000021060620 | 1.00 | 0.00 | 0.00 | 41 |
| 76001100000000021060622 | 1.00 | 0.00 | 0.00 | 33 |
| 76001100000000021060623 | 1.00 | 0.00 | 0.00 | 38 |
| 76001100000000021060624 | 1.00 | 0.00 | 0.00 | 31 |
| 76001100000000021060625 | 1.00 | 0.00 | 0.00 | 41 |
| 76001100000000021060638 | 1.00 | 0.00 | 0.00 | 40 |
| 76001100000000021060639 | 1.00 | 0.00 | 0.00 | 34 |
| 76001100000000021060640 | 1.00 | 0.00 | 0.00 | 33 |
| 76001100000000021060641 | 1.00 | 0.00 | 0.00 | 32 |
| 76001100000000021060642 | 1.00 | 0.00 | 0.00 | 38 |
| 76001100000000099060105 | 1.00 | 0.00 | 0.00 | 31 |
| 76001100000000099060115 | 1.00 | 0.00 | 0.00 | 30 |
| 76001100000000099060126 | 1.00 | 0.00 | 0.00 | 39 |
| 76001100000000099060127 | 1.00 | 0.04 | 0.07 | 28 |
| accuracy | | | 0.32 | 369802 |
| macro avg | 0.81 | 0.32 | 0.22 | 369802 |
| weighted avg | 0.82 | 0.32 | 0.22 | 369802 |

Figura 7.7: Desempeño Modelo Random Forest Código ANM DANE
Fuente: Tomada del script de Colab.

| | | | | |
|---|---|---|---|---|
| 21060609 | 0.64 | 1.00 | 0.78 | 25 |
| 21060611 | 0.48 | 1.00 | 0.65 | 20 |
| 21060613 | 1.00 | 0.00 | 0.00 | 26 |
| 21060614 | 1.00 | 0.00 | 0.00 | 27 |
| 21060616 | 0.16 | 1.00 | 0.27 | 27 |
| 21060617 | 0.67 | 0.32 | 0.43 | 31 |
| 21060618 | 1.00 | 0.00 | 0.00 | 25 |
| 21060622 | 1.00 | 0.00 | 0.00 | 32 |
| 21060623 | 0.31 | 1.00 | 0.48 | 24 |
| 21060624 | 1.00 | 0.00 | 0.00 | 17 |
| 21060625 | 1.00 | 0.00 | 0.00 | 23 |
| 21060638 | 1.00 | 0.00 | 0.00 | 27 |
| 21060639 | 1.00 | 0.00 | 0.00 | 22 |
| 21060640 | 1.00 | 0.00 | 0.00 | 25 |
| 21060641 | 1.00 | 0.00 | 0.00 | 27 |
| 21060642 | 1.00 | 0.00 | 0.00 | 21 |
| 99060105 | 1.00 | 0.05 | 0.09 | 21 |
| 99060115 | 1.00 | 0.00 | 0.00 | 25 |
| 99060126 | 1.00 | 0.00 | 0.00 | 28 |
| 99060127 | 1.00 | 0.09 | 0.16 | 23 |
| accuracy | | | 0.39 | 243561 |
| macro avg | 0.80 | 0.40 | 0.29 | 243561 |
| weighted avg | 0.81 | 0.39 | 0.29 | 243561 |

Figura 7.8: Desempeño Modelo Random Forest Código Coordenada Específica
Fuente: Tomada del script de Colab.

| | | | | |
|---|---|---|---|---|
| 7600121060603 | 1.00 | 0.00 | 0.00 | 31 |
| 7600121060605 | 0.35 | 0.56 | 0.43 | 25 |
| 7600121060607 | 0.28 | 1.00 | 0.43 | 41 |
| 7600121060609 | 0.45 | 0.57 | 0.50 | 23 |
| 7600121060611 | 0.10 | 1.00 | 0.18 | 24 |
| 7600121060613 | 1.00 | 0.00 | 0.00 | 32 |
| 7600121060614 | 1.00 | 0.00 | 0.00 | 36 |
| 7600121060616 | 0.12 | 1.00 | 0.21 | 32 |
| 7600121060617 | 1.00 | 0.00 | 0.00 | 41 |
| 7600121060618 | 1.00 | 0.33 | 0.50 | 27 |
| 7600121060620 | 1.00 | 0.00 | 0.00 | 41 |
| 7600121060622 | 1.00 | 0.00 | 0.00 | 33 |
| 7600121060623 | 1.00 | 0.00 | 0.00 | 38 |
| 7600121060624 | 1.00 | 0.00 | 0.00 | 31 |
| 7600121060625 | 1.00 | 0.00 | 0.00 | 41 |
| 7600121060638 | 1.00 | 0.00 | 0.00 | 40 |
| 7600121060639 | 1.00 | 0.00 | 0.00 | 34 |
| 7600121060640 | 1.00 | 0.00 | 0.00 | 33 |
| 7600121060641 | 1.00 | 0.00 | 0.00 | 32 |
| 7600121060642 | 1.00 | 0.00 | 0.00 | 38 |
| 7600199060105 | 1.00 | 0.00 | 0.00 | 31 |
| 7600199060115 | 1.00 | 0.00 | 0.00 | 30 |
| 7600199060126 | 1.00 | 0.00 | 0.00 | 39 |
| 7600199060127 | 1.00 | 0.04 | 0.07 | 28 |
| accuracy | | | 0.32 | 369802 |
| macro avg | 0.81 | 0.32 | 0.22 | 369802 |
| weighted avg | 0.82 | 0.32 | 0.22 | 369802 |

Figura 7.9: Desempeño Modelo Random Forest Código ANM (Procesado)
Fuente: Tomada del script de Colab.

Finalmente, se cargó el conjunto de datos del SISBEN para el municipio de Cali realizandole a este el mismo procedimiento de los datos mencionado en la Figura 7.1, pero al momento de ejecutar el Random Forest para predecir la manzana geográfica con estos nuevos datos, el colab en su versión Pro no logró culminar la ejecución ya que el entorno se desconectaba de manera persistente, debido a la alta demanda computacional sumado al tamaño y tipo de los datos de las variables para predecir este elemento, por tanto, no se logró culminar este paso.

### 7.1.4. Visualización de la localización

Para la visualización de los resultados obtenidos en la predicción del modelo se tomaron como referencia los datos de las manzanas de la ciudad de Cali, cabe anotar que previo a la implementación del PSM se identificaron un total de 13.140 manzanas en la ciudad, así mismo, cuando se realizó la ejecución del Random Forest estas fueron 11.810 manzanas y con los resultados de la predicción del modelo se lograron ubicar en el esquema de Tableau **11.006 manzanas censales** como elemento de granularidad más fino dentro del marco de georreferenciación del Censo 2018 en la capital del Valle del Cauca.

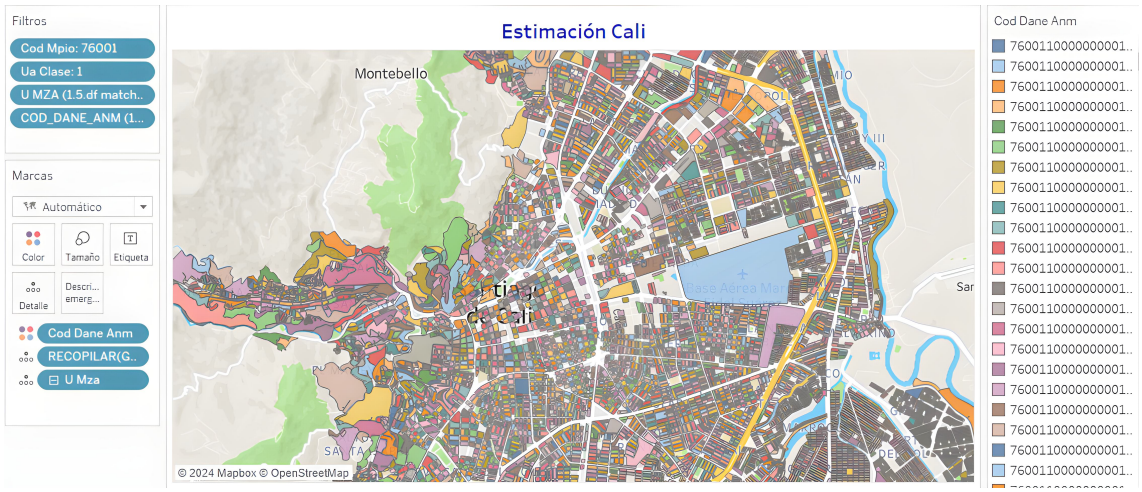


Figura 7.10: Visualización de Datos en Tableau
Fuente: Tomada de Tableau.

# Capítulo 8

# Conclusiones y Trabajos Futuros

## 8.1. Conclusiones

- Los microdatos del Censo Poblacional en Colombia se constituyen como la principal fuente de información sociodemográfica para la toma de decisiones en materia de política económica y social, es una base de datos muy robusta y representativa de la población del país, brindando información en distintos aspectos como de la persona, las condiciones de la vivienda, las características del hogar y los distintos elementos geográficos de los individuos.

- El modelo de Propensity Score Matching – PSM es una técnica muy robusta para los procesos de emparejamiento, muestra de ello es que, a pesar de la alta variabilidad que existe en los registros del Censo y los tipos de datos que contiene, el algoritmo logró establecer los puntajes de propensión que es la probabilidad que un individuo se ubique en el grupo de tratamiento o control, demostrando el cumplimiento de la condición de soporte común, y luego obtener mediante método de vecinos más cercanos – KNN los conjuntos de datos emparejados, sobre las muestras más robustas.

- De los tres conjuntos de datos a los cuales se les aplicó el algoritmo de PSM, el que demostró el mayor desempeño fue el conjunto 1 de Valle Urbano, representado por la población de Cali en el grupo de tratamiento y en control por los habitantes de los demás municipios. De lo anterior, Cabe resaltar que era la base más representativa del departamento, pues obtuvo un Accuracy de 61 %, y posteriormente con la implementación del KNN mostró una disminución de las diferencias entre las variables del conjunto emparejado frente al no emparejado.

- Finalmente, la predicción del Random Forest para clasificar a la población según el Código ANM DANE, determinó la ubicación exacta de un individuo en Cali dadas las características de: Sexo, edad, estado civil, niveles educativo máximo alcanzado, cobertura de servicios públicos (acueducto, alcantarillado, gas e internet), condiciones de ocupación laboral y estrato socioeconómico, logrando como resultado un nivel de precisión del 39 %, previa transformación de la variable de ubicación para superar la dificultad de la categoría y reconociendo dato este resultado la posible existencia de otras variables observables para determinar la localización.

## 8.2. Trabajos Futuros

Frente la oportunidad de desarrollar trabajos futuros que se podrían derivar del presente proyecto, se identifican las siguientes:

- **Técnicas de aprendizaje profundo:** La implementación de otras técnicas de clasificación para abordar este tipo de problemáticas, como es el caso de las redes neuronales, son muy versátiles para aprender y modelar relaciones no lineales, son escalables y se desempeñan muy bien con conjuntos de datos grandes y de alta dimensionalidad.
- **Identificación de nuevos elementos de georreferenciación y características relevantes:** Es claro que el Censo y el SISBEN son los instrumentos estadísticos que mayor información brindan para el desarrollo de políticas económicas y sociales en el país, sin embargo, como sus elementos de localización son únicos tienen la dificultad para enlazarse entre si o con otras fuentes, por tanto, surge la necesidad de encontrar nuevos elementos de localización, quizá más regionales pero que guarden una adecuada ubicación geográfica, que sea representativa y escalable, que permita el levantamiento de otros tipos de indicadores representados en factores socioeconómicos, como el ingreso, nivel de pobreza, desigualdad, entre otros.

# Bibliografía


[1] A. Barreto Villanueva, "El progreso de la estadística y su utilidad en la evaluación del desarrollo," *Papeles de Población*, vol. 18, no. 73, pp. 1–31, 2012.

[2] DANE, "Censo nacional de población y vivienda 2018." `https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018`, 2020. Fecha de acceso: 08/07/2023.

[3] DANE, "Archivo nacional de datos (anda)." `https://microdatos.dane.gov.co/index.php/catalog/643`, 2022. Fecha de acceso: 08/07/2023.

[4] P. de la República de Colombia, "Decreto 262 de 2004: Por el cual se modifica la estructura del Departamento Administrativo Nacional de Estadística DANE y se dictan otras disposiciones.," 2004.

[5] DANE, "Cuentas nacionales departamentales." `https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales`, 2023. Fecha de acceso: 08/07/2023.

[6] L. C. Riaño Bermudez, "Análisis espacio-temporal del desplazamiento forzado en colombia," Master's thesis, Universidad Nacional de Colombia, Bogotá, Sept. 2022.

[7] DANE, "Integración de la población venezolana en colombia: Impactos de las características de las personas y los hogares en la participación laboral," Tech. Rep. ISBN 978-958-5437-21-0, Departamento Administrativo Nacional de Estadística (DANE), Fondo de Población de las Naciones Unidas (UNFPA), Bogotá, 2022.

[8] D. Rubin and P. Rosenbaum, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[9] R. Bernal and X. Peña, *Guía práctica para la evaluación de impacto*. Universidad de los Andes, 2011.

[10] A. Malffioli, "Métodos experimentales y no-experimentales," in *Taller de evaluación de impacto de programas de ciencia, tecnología e innovación*, Ciudad de México, 2014.

[11] J. Shambaugh, A. Klein, and J. Herbert, "Structural measures as predictors of injury in basketball players," *Med Sci Sports Exerc*, vol. 23, no. 5, pp. 522–7, 1991.

[12] K. Caballero and J. Ferrer, "Evaluación de políticas públicas con microsimulaciones." 2011.

[13] M. Navarrete, P. Aroca, and J. Bernal, "Matching espacial para georreferenciar datos de encuestas de hogar," *Estudios de Economía*, vol. 44, no. 1, pp. 53–80, 2017.

[14] R. Fernández Machado and P. Medina Quispe, *Evaluación de impacto del acceso a las TIC sobre el ingreso de los hogares*. IRSI, Lima, 2011.

[15] P. Ponce, S. Robles, R. Alvarado, and C. Ortiz, "Efecto del capital humano en la brecha de ingresos," *Revista Economía y Política*, vol. 29, pp. 25–47, 2019.

[16] M. Mediavilla bordalejo, *Las becas y la continuidad escolar en el nivel secundario post-obligatorio*. Universitat de Barcelona, Barcelona, 2010.

[17] J. J. Echavarría, S. Restrepo-Tamayo, M. Villamizar-Villegas, and J. D. Hernandez-Leal, *Impacto del crédito sobre el agro en Colombia*. Banco de la República de Colombia, Bogotá, 2017.

[18] K. O. Muñoz Martínez, "Análisis del vínculo entre el transporte público y el empleo en cartagena, colombia," 2021.

[19] C. C. Gómez Cañon, "Consecuencias de ser padre a temprana edad sobre los ingresos," *Ensayos sobre Política Económica*, vol. 34, pp. 103–125, 2016.

[20] B. de la República, "Un índice de precios espacial para la vivienda urbana en colombia: Una aplicación con métodos de emparejamiento," working paper, Banco de la República, Cartagena, 2012.

[21] D. A. N. de Estadística (DANE), "Manual de uso del marco geoestadístico nacional en el proceso estadístico," working paper, Departamento Administrativo Nacional de Estadística (DANE)Banco de la República, Bogotá D.C., 2018.

[22] imbalanced-learn Development Team, "imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," 2023. Accedido: 2024-06-25.

[23] Scikit-learn, "Onehotencoder," 2024. Accedido: 2024-06-28.

[24] Scikit-learn, "Standardscaler," 2024. Accedido: 2024-06-28.

[25] H. Wang, "Propensity score matching in python," 2024. Accedido: 2024-06-28.

[26] Scikit-learn, "Logisticregression," 2024. Accedido: 2024-06-28.

[27] Scikit-learn, "Kneighborsclassifier," 2024. Accedido: 2024-06-28.

[28] Scikit-learn, "Randomforestclassifier," 2024. Accedido: 2024-06-28.