Data Science en Política Pública: hacia decisiones basadas en evidencia

Integrantes:

- Melani Caucota
- Cristina Betancur
- Virginia Chirilá

Buenos Aires Ciudad

Introducción

La Encuesta de Hogares 2019, realizada por CABA, proporciona una visión integral de las condiciones de vida y la estructura socioeconómica. Este tipo de estudio puede ayudar en la toma de decisiones referidas a, por ejemplo:

01 Planeación y Política Pública:

Los gobiernos utilizan los datos de las encuestas para diseñar y evaluar políticas públicas. Identificar necesidades y asignar eficientemente recursos.

02 Desarrollo Económico:

Los datos económicos, como los niveles de ingreso, empleo, y gasto de los hogares, ayudan a comprender la situación económica de un país o región, esto facilita la implementación de políticas que promuevan el crecimiento económico y la reducción de la pobreza.

03 Evaluación de Programas

Las organizaciones y gobiernos pueden utilizar los datos de encuestas para evaluar el impacto de programas específicos, como aquellos destinados a reducir la desigualdad o mejorar la salud pública.

04 Demografía y Planificación Urbana

Permiten a los urbanistas planificar el crecimiento de ciudades y regiones, garantizando que la infraestructura y los servicios se ajusten a las necesidades de la población actual y futura.

Objetivo general

Generar conocimiento para mejorar la toma de decisiones con respecto a la población de la Ciudad Autónoma de Buenos Aires (CABA).

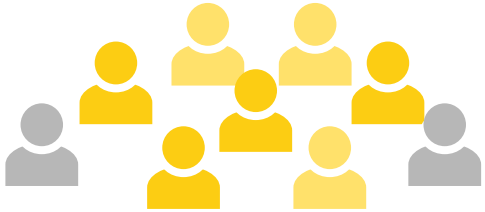
Objetivos específicos

1ºObjetivo:

Analizar la probabilidad de que una persona utilice el sistema de salud privado o público en función a características demográficas, laborales, educativas y sanitarias.

2ºObjetivo:

Estudiar la influencia de características demográficas, laborales, educativas y sanitarias sobre el ingreso económico de una persona.



3ºObjetivo:

Agrupar personas de acuerdo a características socio-económicas, educativas, demográficas y sanitarias para identificar patrones y segmentaciones relevantes.

Metodología

Fuente de datos

Dirección General de ESTADÍSTICA Y CENSOS
Ministerio de Hacienda y Finanzas

Encuesta Anual de Hogares del año 2019 de la Ciudad Autónoma de Buenos Aires (CABA) realizado por la Dirección General de Estadísticas y Censos.

Gran %
en valores nulos

7784 Hijos nacidos vivos
1054 Nivel máximo educativo
64 Años escolaridad

Variables

Procesamiento

01 Tratamiento de valores nulos.

- Imputamos datos
- Creación de nuevas categorías
- Imputamos los registros “No corresponde”
- Eliminación de registros

02 Creación de nuevas variables.

- Ingresos en USD
- Categoría de ingresos (nulo,bajo,alto)
- Atención en sistema sanitario público (si, no).

03 Transformación de variables categóricas.

- Variables binarias -> Get Dummies
- Variable multinivel -> LabelEnconder

Metodología

Estructura del Dataset

Dataset original 31 columnas, 14319 registros

Dataset final 41 columnas, 14310 registros

1ºObjetivo: sistema de salud público

13 Variables predictoras

'dominio', 'sexo_Mujer', 'edad', 'parentesco_jefe', 'situacion_conyugal', 'estado_ocupacional', 'cat_ocupacional', 'nivel_max_educativo', 'años_escolaridad', 'lugar_nacimiento', 'cantidad_hijos_nac_vivos', 'hijos_nacidos_vivos_Si', 'ingresos_totales_USD'

['sist_de_salud_Sistema público']

2ºObjetivo: predecir el ingreso total

12 Variables predictoras

'dominio', 'sexo_Mujer', 'edad', 'parentesco_jefe', 'situacion_conyugal', 'estado_ocupacional', 'cat_ocupacional', 'nivel_max_educativo', 'años_escolaridad', 'lugar_nacimiento', 'cantidad_hijos_nac_vivos', 'hijos_nacidos_vivos_Si',

['ingresos_totales_USD']

3ºObjetivo: Agrupamiento de personas

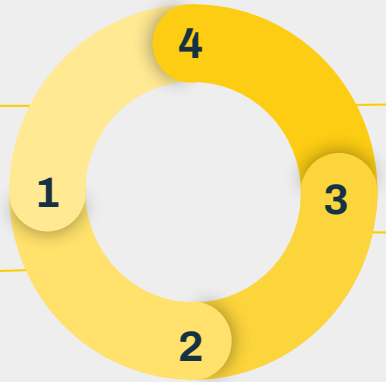
13 Variables

'dominio', 'edad', 'sexo_Mujer', 'situacion_conyugal', 'estado_ocupacional', 'estado_educativo' 'nivel_max_educativo', 'años_escolaridad','lugar_nacimiento', 'cantidad_hijos_nac_vivos', 'hijos_nacidos_vivos_Si', 'ingresos_totales_USD_cat', 'sist_de_salud_Sistema público'

11254 Registros de +18

Algoritmos aprendizaje supervisado

1er objetivo: Salud pública



Train/Test Split

Ajuste de algoritmos

- KNN
- Random Forest

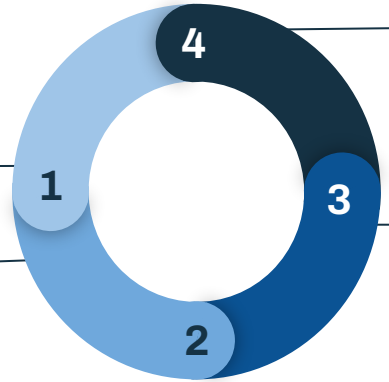
Reajuste y selección del mejor algoritmo

- Aplicación de método SMOTE por desbalance de clases

Evaluación

- Matriz de confusión
- Precisión
- Recall
- F1-score
- Accurency
- Support y balance de clases

2do objetivo: Ingresos



Reajuste y selección del mejor algoritmo

Train/Test Split

Ajuste de algoritmos

- Regresión Lineal (ingresos cuant.)
- KNN y Random Forest (variable de ingresos cual.)

Evaluación

Regresion Lineal:

- R^2
- MSE
- QQPlot
- Homogeneidad de varianzas

KNN y Radom Forest

- Matriz de confusión
- Precisión
- Recall
- F1-score
- Accurency
- Support y balance de clases

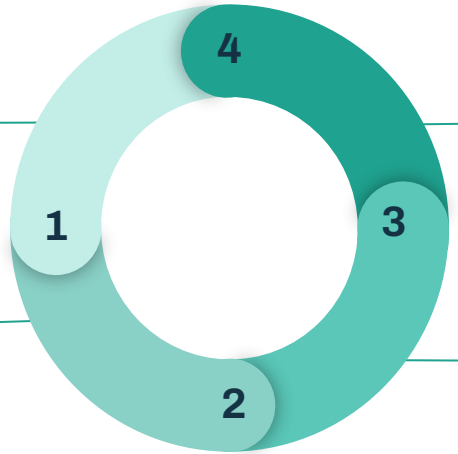
Algoritmos aprendizaje no supervisado

3er objetivo: Agrupamiento de personas

Selección de variables y estandarización de datos

Ajuste de algoritmos

- **K-Means** (mejor k estimado por método Elbow y Silhouette)
- **DBSCAN** (mejor eps estimado gráficamente)



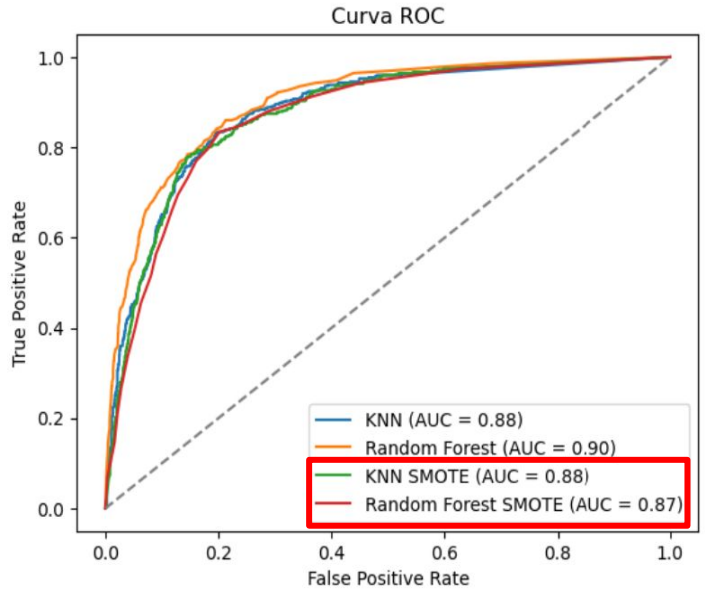
Selección del mejor algoritmo

Basada en la visualización de clusters compactos y bien separados

Evaluación

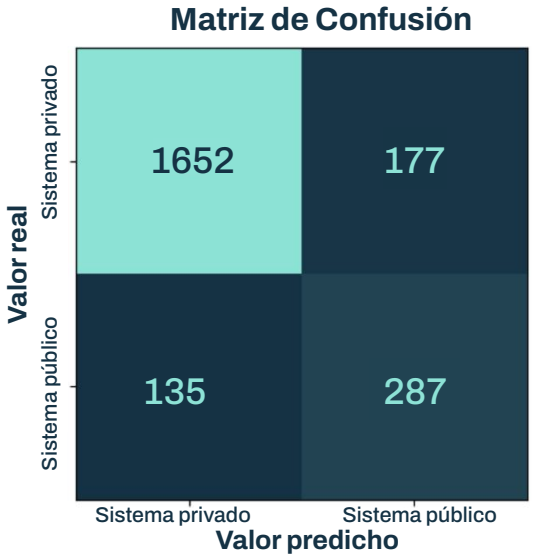
- **K-Means**
  - representante y constitución de cada cluster
  - análisis gráfico con reducción de dimensionalidad (PCA)
- **DBSCAN**
  - análisis de ruido
  - análisis gráfico con reducción de dimensionalidad (PCA)

Aprendizaje Supervisado - 1er objetivo



	Random Forest SMOTE	KNN SMOTE
Precisión General	86.14%	82.63%
Precisión	62%	52%
F1-score	0,65	0,63
Recall	68%	79%
AUC	0,87	0,88

Random Forest



86% de los casos son predichos correctamente

	Sist. público	Sist. privado
Precisión	62%	92%
Recall	68%	90%

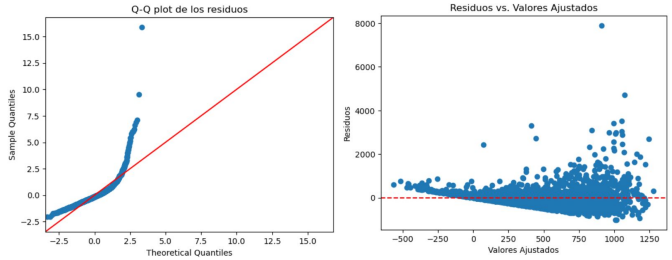
Features importantes

- Ingresos totales 21%
- Edad 18%
- Años de escolaridad 14%
- Nivel máx. educativo 10%

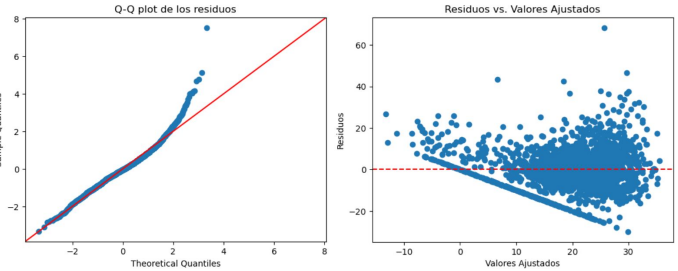
Resultados

Aprendizaje Supervisado - 2do objetivo

Regresión Lineal - va. ingresos cuantitativa



	MSE	R^2
Primer modelo	246141	0.30
Modelo con transformación	81.98	0.41



Variable	β
dominio	-1,83
edad	2,54
estado ocupacional	11,63
años escolaridad	5,1
cantidad de hijos	-0,3
sexo	-2,62

Random forest - va. ingresos cualitativa

Precisión general

56% de los casos son clasificados correctamente

	Nulo	Muy bajo	Bajo	Medio	Alto
Precisión	71%	42%	43%	64%	24%
Recall	73%	32%	44%	72%	10%

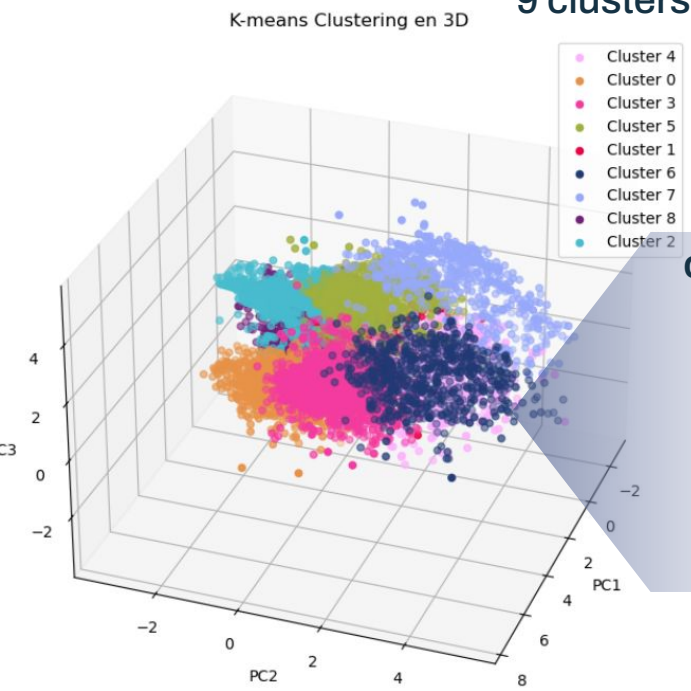
Features importantes

- Edad **35%**
- Años de escolaridad **12%**
- Lugar de nacimiento **9%**
- Situación conyugal **7%**

Aprendizaje No Supervisado

K-Means

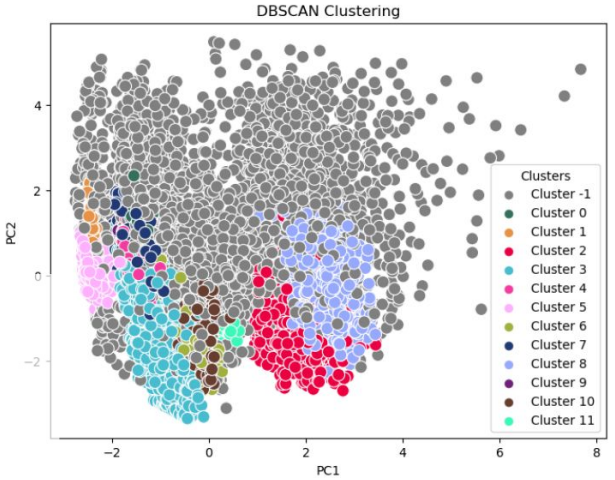
9 clusters



Cluster 6 (770 personas)

- 100 % son varones
- 58% vive en villas de emergencia
- 40% son casados
- 69% está desocupado o inactivo
- 75% tiene un ingreso nulo o bajo
- 45% asistió hasta nivel primario
- 57% es originario de otros países
- 100% posee hijos
- 85% utiliza el sistema de salud público

DBSCAN



12 clusters
66 % ruido

Conclusión

¿Logramos predecir como queríamos?

Si bien los algoritmos de aprendizaje supervisado no fueron muy precisos, nos permitieron distinguir **cuales son los factores más importantes** a la hora de predecir las variables de interés.

Ingresos

Descubrimos que...

Los factores más importantes que influyen sobre los ingresos que percibe una persona son:

- **Edad de la persona**
- **Escolaridad**

Sistema de salud

Descubrimos que...

La mayoría de las personas usan obras sociales o prepagas, y los factores más importantes que modulan el uso o no del sistema público son:

- **El nivel de ingresos**
- **Edad de la persona**
- **Escolaridad**

Se puede considerar que...

La escolaridad es el factor común más importante.

Influye en el ingreso económico de las personas, y en el acceso a proveedores de salud de calidad, o en acceder a una mayor variedad fuera del sistema público.

¿Logramos agrupar a las personas de manera precisa?

En el caso de los algoritmos de aprendizaje no supervisado. La densidad de nuestros datos junto a la presencia de ruido, llevó que los clusters definidos estén muy superpuestos entre sí, y que por lo tanto la pertenencia de un registro a un cluster.

La información generada por este trabajo puede servir al gobierno como soporte para la toma de decisiones futuras, las cuales podrían traducirse en mejoras en la calidad de vida de las personas


¡Gracias!
¿Preguntas?
Buenos Aires Ciudad