Read me

Este repositorio contiene los códigos y data sets utilizados para la exploración, análisis y construcción de modelos de machine learning basados en la Encuesta Anual de Hogares del año 2019 de la Ciudad Autónoma de Buenos Aires (CABA).

Estructura del repositorio

Código:

1. Parte 1. Exploración y procesamiento de datos.ipynb
2. Parte 2. Aprendizaje supervisado.ipynb
3. Parte 3. Aprendizaje no supervisado.ipynb

Data sets:

* 1. Original (usado en código 1): “encuesta-anual-hogares-2019.csv”
  2. Limpio con variables categóricas originales (generado por código 1): “encuesta-anual-hogares-2019-sintransf.csv”
  3. Limpio con variables categóricas numéricas (generado por código 1 y empleado en código 2 y 3): “encuesta-anual-hogares-2019-transf.csv”

Introducción del tema

Las encuestas socioeconómicas son herramientas fundamentales para los gobiernos porque proporcionan datos clave para entender la realidad de una población en términos de ingresos, educación, empleo, acceso a servicios, condiciones de vivienda y otros indicadores vitales. La información generada por este tipo de herramienta no solo permite diagnosticar la realidad social de una población, sino que también puede utilizarse para el diseño y la evaluación de políticas públicas, el monitoreo de tendencias sociales y la planificación y el desarrollo económico, permitiendo como objetivo final la toma de decisiones informadas e inteligentes basadas en evidencia.

Objetivo del proyecto

Con este trabajo planemos generar conocimiento que permita la mejora en la toma de decisiones con respecto a la población de la Ciudad Autónoma de Buenos Aires (CABA). Para esto, generamos preguntas de interés social, sanitario, económico y demográfico, las cuales tratamos de resolver valiéndonos de datos generados por la encuesta anual de hogares de CABA y tomando como unidad de estudio a las personas. Como objetivo específicos nos propusimos: analizar la probabilidad de que una persona utilice el sistema de salud privado o público en función a características demográficas, laborales, educativas y sanitarias; estudiar la influencia de características demográficas, laborales, educativas y sanitarias sobre el ingreso económico de una persona y agrupar personas de acuerdo a características socio-económicas, educativas, demográficas y sanitarias para identificar patrones y segmentaciones relevantes.

Origen de los datos

El dataset seleccionado fue generado por el Gobierno de la Ciudad de Buenos Aires a partir de encuestas realizadas en hogares de la Ciudad de Buenos Aires durante el año 2019. La población de estudio en la que se basó, está constituida por los hogares particulares de CABA y las personas integrantes de cada unidad de hogar. El data set original contiene 14319 registros y 31 variables relacionadas a vivienda y hábitat, hogares (tamaño y composición de los hogares), características demográficas básicas de la población (composición y estructura, situación conyugal, migración), trabajo y empleo (situación de la población en el mercado de trabajo, características ocupacionales), ingresos (laborales y no laborales), educación (asistencia al sistema de educación formal, nivel educativo de la población), salud (tipo de cobertura, consultas al médico, estado general de salud) y fecundidad. El Data Set se obtuvo del sitio oficial: <https://data.buenosaires.gob.ar/dataset/encuesta-anual-hogares/resource/3a45c563-396d-42de-ba93-8a93729e0723>

Parte 1. Exploración y procesamiento de datos

En la primera parte de nuestro análisis realizamos exploración y procesamiento de los datos. Observamos que tres variables presentaron gran proporción de valores nulos (74% en la variable Hijos nacidos vivos y 7% en Nivel máximo educativo). La mayoría de estos nulos fueron tratados a partir de imputación y creación de nuevas categorías. Para facilitar la extensibilidad del análisis y la aplicación de los algoritmos decidimos generar nuevas variables a partir de otras preexistentes (va. generadas: ingresos totales en USD, categoría de ingresos, atención en sistema sanitario público). Las variables categóricas fueron transformadas a numéricas dependiendo de la cantidad de niveles (va. binarias fueron transformadas con GetDummies y las va. multinivel con LabelEncoder). Observamos correlaciones no lineales entre algunas de las variables cuantitativas y asociaciones significativas entre algunas de las variables categóricas.

Parte 2. Aprendizaje supervisado

En la segunda parte de nuestro análisis empleamos algoritmos de aprendizaje supervisado para poder resolver los primeros dos objetivos planteados (predecir el uso del sistema de salud público y los ingresos de las personas). Para el primer objetivo utilizamos los algoritmos de clasificación KNN y Random Forest. Debido al desbalance de clases presente en nuestros datos reajustamos ambos algoritmos utilizando la técnica SMOTE. Random Forest con la técnica SMOTE fue el algoritmo que mejor performace tuvo, con una precisión general del 86%. Observamos que los ingresos y la edad fueron variables que tuvieron gran importancia en este modelo. Para el segundo objetivo aplicamos en primera instancia Regresión Lineal Múltiple utilizando la variable de ingresos cuantitativa. Observamos que los modelos para esta variable explicaron poco porcentaje de la variabilidad de los datos (entre un 30 y un 40%). Los coeficientes del modelo muestran que, por ejemplo, la edad y la educación influyen positivamente en los ingresos de las personas. En segunda instancia utilizamos algoritmos de clasificación KNN y Random Forest sobre la variable de ingresos cualitativa. El mejor modelo fue el del algoritmo Random Forest, el cual tuvo una precisión general del 56%. Las variables que aportaron en mayor medida a este modelo fueron nuevamente la edad y la educación.

Parte 3. Aprendizaje no supervisado

En la tercera y última parte de nuestro análisis empleamos algoritmos de aprendizaje no supervisado con el objetivo de poder agrupar personas de acuerdo a sus características socioeconómicas y demográficas. Empleamos K-Means y DBCAN como algoritmos. Para K-Means estimamos un valor k optimo de 9 clusters. En el caso de DBSCAN, se reconoció 12 clusters pero gran parte de los registros fueron clasificados como ruido. La baja precisión observada en ambos algoritmos puede ser debida a la gran densidad de nuestros datos. Esto puede ocasionar que los clusters estén muy superpuestos entre sí y que por lo tanto la pertenencia de un registro a un cluster o no, no sea tan precisa.

Conclusiones

Si bien algunos de los algoritmos de aprendizaje supervisado no fueron muy precisos, nos permitieron distinguir qué factores son importantes a la hora de predecir el uso del sistema sanitario público o el ingreso de las personas. Para el caso del sistema de salud observamos que la mayoría de las personas usan obras sociales o prepagas y que los factores más importantes que modulan el uso o no del sistema público fueron los ingresos, la edad y la escolaridad. En el caso de los ingresos, observamos que los factores más importantes que parecen influir sobre los ingresos que percibe una persona son la edad y la escolaridad. Por otro lado, la gran densidad de nuestros datos  y la presencia de ruido influyeron en la calidad del agrupamiento de los clusters en los algoritmos de aprendizaje no supervisado, impidiendo considerar a estas herramientas como alternativas efectivas para comprender nuestros datos. Nuestro trabajo, en términos generales, logra discretizar a la educación como factor importante a tener en cuenta para poder asegurar buenos ingresos y mejor acceso a la salud. La información generada puede servir como soporte para la toma de decisiones futuras del gobierno y para asegurar la mejora en la calidad de vida de las personas.