Clasificación supervisada de las zonas de riesgo en Barcelona

Anna Moreno Gayubas
19 marzo 2024

Resumen

Este estudio se centra en el desarrollo de un modelo de clasificación supervisada de las zonas de riesgo en Barcelona, utilizando Machine Learning. El objetivo es crear un modelo que clasifique en 3 categorías riesgo alto, medio o bajo.

Características generales

El dataset es un conjunto de datos privado. Este conjunto de datos se obtienen variables categóricas: lugar concreto (calle, parque, etc), barrio, distrito, fecha y hora. Las variables numéricas són: jóvenes hombres o mujeres o personas no binarias de menos de 21 años en situación de calle. Las variables binárias, tienen en cuenta dinámicas de impacto en el espacio público como: pernocta, consumo de diferentes tipos de drogas, acumulación de volumen de personas o proximidad a un recurso de personas en situación de calle. Estos datos se extraen de los registros de los profesionales que trabajan en la ciudad de Barcelona. Concretamente, el equipo que trabaja con menores y jóvenes de hasta 21 años que viven en la calle.
Los datos obtenidos son del año 2020 al 2024. Constan de variables numéricas sobre el número de personas (menores de 21 años) detectadas en un espacio de Barcelona, el barrio y distrito específico de la ciudad. Así como las dinámicas observadas. El dataset obtenido consta de 21344 filas y 29 columnas.

Objetivo general

El principal propósito de este proyecto es la creación de una herramienta destinada a la clasificación de las áreas de riesgo en el área de Barcelona. En este contexto, se pretende aplicar técnicas de Machine Learning que ayuden a comprender si se puede ayudar a identificar patrones o tendencias en el conjunto de datos, facilitando la tarea de clasificación a los profesionales que trabajan con dichos datos. En este estudio se aborda el problema de la clasificación utilizando técnicas de clasificación supervisada multiclase por tal de agilizar las tareas de clasificación. Para ello se establece la variable target: semáforo. Es una variable ya dada en el dataset, que clasifica los diferentes lugares según su riesgo como rojo, amarillo o verde.

Objetivo específico

Primero, se realizará un análisis exploratorio de los datos con tal de profundizar en el comportamiento del conjunto de datos. Luego se procederá a la limpieza de los datos y se hará uso de diferentes modelos de machine learning para entrenar modelos que posteriormente ayuden a la predicción de la clasificación. Además, se explorará qué variables van a ayudar a predecir mejor el modelo, y aquellas variables ruidosas.

Evaluación

Finalmente, se evaluará el modelo utilizando métricas como accuracy,precisión, el recall, el F1-score. También se utilizará la validación cruzada estratificada para asegurar que cada pliegue tenga una distribución de clases similar a la del conjunto de datos completo. Con estas dos evaluaciones se decidirá qué modelo es más eficiente.

Conclusiones

En este estudio se aborda el problema de la clasificación del riesgo teniendo en cuenta el impacto de dinámicas disruptivas registradas por los agentes que hacen uso del espacio público. Crear una herramienta que impida el sesgo de cada trabajador a la hora de registrar el riesgo de impacto en la variable target es fundamental para poder optimizar y analizar los datos obtenidos. Aplicando la herramienta, el profesional registrará los datos numéricos y empíricos y el modelo almacenará la variable riesgo reduciendo el sesgo de la información recogida. Así se realizará un mejor análisis y se identificará de manera precoz los impactos en el espacio público.