APRENDIZAJE AUTOMATICO

Trabajo Práctico

Objetivo:

Utilizando técnicas de machine learning, este proyecto tiene como objetivo desarrollar un sistema capaz de clasificar el Índice de Entorno de Calidad de Vida basándose en las siguientes variables: espacios verdes, transporte público, industrias, la existencia de cavas, y el nivel de riesgo de inundación. La combinación de estas variables da como resultado la clasificación del Índice de Entorno.

Contexto del Problema:

El Índice de Entorno de Calidad de Vida es una medida fundamental para evaluar el bienestar de la población en áreas urbanas. Mejorar la calidad de vida no solo beneficia a sus habitantes directos, sino que también contribuye al desarrollo socioeconómico y ambiental más amplio de la zona.

Interrogantes de estudio:

Algunas de las preguntas que intenta responder este trabajo son:

1. ¿Con qué precisión puede un modelo de aprendizaje automático clasificar el Índice de Entorno de Calidad de Vida utilizando las variables seleccionadas?
2. ¿Cuáles variables tienen el impacto más significativo en la clasificación?
3. ¿Puede el modelo identificar áreas urbanas específicas que necesitan mejoras para aumentar la calidad de vida?

Descripción del Índice de Entorno:

Es un índice en el que se incorporan indicadores, del ambiente que rodea a la población, vinculados a servicios públicos, eventos climáticos o intervenciones humanas, que aportan, positiva o negativamente, a la calidad de vida. Estos indicadores son: espacios verdes públicos, cavas, transporte público, industrias y riesgo por inundación.

Relevancia:

Este índice apunta a relevar las características de aquella porción del ambiente que interactúa significativamente con la población en su lugar de residencia, incidiendo en su calidad de vida. Para ello evalúa una serie de elementos (de orden social, productivo, cultural y físico natural) presentes en el territorio.

Alcance (qué mide el índice):

El Índice mide el nivel de condiciones (vinculadas al entorno) de bienestar general que ofrece el Estado a la población, en un radio censal determinado.

Descripción de las variables que componen el índice:

- ✔ Indicador de Espacio verde público: porcentaje de población, en un radio censal determinado, que vive en el área de influencia de espacios verdes públicos (EVP), apto para la realización de actividades recreativas, deportivas y culturales.

- ✔ Indicador de Cavas: porcentaje de población, en un radio censal determinado, que reside en el área de influencia cavas. Según la

distancia de la población a la cava se consideran distinto el grado de influencia de esta última.

- ✔ Indicador de Transporte público: porcentaje de población, en un radio censal determinado, que vive en el área de influencia del recorrido del transporte público de pasajeros (línea de colectivo urbano o interurbano y tren). Se adoptan distintas áreas de influencia para cada medio de transporte.

- ✔ Indicador de Industrias: porcentaje de población que, en un radio censal dado, reside en el área de influencia de industrias con impacto ambiental significativo. Considera que las industrias afectan a distinto porcentaje de la población según la distancia a la que se encuentre, y dichas distancias dependen de la categoría de las industrias.

- ✔ Indicador de Riesgo por inundación: porcentaje de población, en un radio censal dado, expuesta a un riesgo relevante por inundación para una recurrencia de 10 años. Se entiende por relevante cuando puede causar daños a la población más vulnerable, niños y ancianos.

Tipo de presentación de resultados:

- Muy bajo
- Bajo
- Medio
- Alto
- Muy alto

Área de cobertura:

Cuenca Hídrica Matanza-Riachuelo

Fuente de datos:

INDEC, Población, Censo Nacional de Población, Hogares y Viviendas.
Datos del censo: información disponible en formato digital por INDEC.

Desarrollo del proyecto

Para el desarrollo del proyecto se utilizarán las siguientes técnicas:

Exploración y pre procesamiento de Datos:

Se cargarán los datos desde un archivo Excel utilizando la biblioteca Pandas. Posteriormente, se seleccionarán las columnas relevantes para el análisis. Se llevará a cabo una exploración inicial de los datos para comprender su estructura y características. Durante esta fase, se identificarán y manejarán tanto los valores faltantes como los outliers, asegurando que los datos estén listos para el análisis.

División de Datos y Normalización:

Los datos se dividirán en conjuntos de entrenamiento y prueba para facilitar la evaluación del modelo. Además, se aplicará la normalización a los datos para asegurar que todas las características tengan la misma escala. Esto garantizará un rendimiento óptimo del modelo y evitará cualquier sesgo causado por diferencias en las magnitudes de las características.

Visualizaciones y Análisis:

Se incluirán visualizaciones adicionales, como mapas de calor o gráficos de dispersión, para respaldar las conclusiones y destacar patrones interesantes en los datos. Esto ayudará a comunicar de manera efectiva los resultados del análisis y facilitará la interpretación de los hallazgos.

Selección y Entrenamiento del Modelo:

Se seleccionará un algoritmo de clasificación adecuado para el problema en cuestión. En este caso, se empleará un clasificador de Random Forest debido a su capacidad para manejar datos complejos y su robustez frente a overfitting. El modelo se entrenará utilizando los datos de entrenamiento, ajustando sus parámetros según sea necesario para lograr un rendimiento óptimo.

Evaluación del Modelo y Generación de Resultados:

Una vez entrenado el modelo, se realizarán predicciones utilizando el conjunto de prueba. Se evaluará el rendimiento del modelo utilizando métricas apropiadas, como la precisión y el informe de clasificación. Se generará un análisis detallado de los resultados obtenidos, extrayendo conclusiones significativas y destacando cualquier hallazgo relevante.

Conclusiones:

Se extraerán conclusiones basadas en el rendimiento del modelo y los resultados obtenidos durante el análisis. Se proporcionarán recomendaciones para futuras investigaciones o mejoras en el modelo, identificando áreas potenciales para optimizar el rendimiento o explorar en profundidad.