Examen Integrador – Módulo I

# 1. Análisis Exploratorio Espacial de Datos (ESDA)

Para este análisis se utilizó la base de datos Chi-SDOH, correspondiente a las zonas censales de Chicago, que contiene variables sociales, económicas y de salud. Se trabajó con dos variables numéricas:

- vcrimert: Tasa de crimen violento por zona censal.  
- EP\_DISABL: Porcentaje de personas con discapacidad en cada zona.

Se utilizó el software GeoDa para calcular el Índice de Moran Bivariado con dos matrices espaciales: rook y queen.

## 1.1 Resultados del Índice de Moran Bivariado

Con la matriz de pesos 'queen', el valor del índice de Moran fue de -0.030. Este valor indica que no existe autocorrelación espacial significativa entre la tasa de crimen violento y el porcentaje de personas con discapacidad. El valor está muy cerca de cero, lo que refleja una distribución aleatoria.

## Conclusión del Moran I:

No se encontró autocorrelación espacial significativa entre las dos variables analizadas. Esto indica que las zonas con alto crimen violento no tienden a estar rodeadas por zonas con altos porcentajes de discapacidad, y viceversa.

# 2. Regresión Espacial

Se aplicaron tres modelos para evaluar la relación entre las variables: modelo OLS, modelo SAR (Spatial Lag), y modelo SEM (Spatial Error).

## 2.1 Regresión OLS

La regresión lineal ordinaria (OLS) arrojó un coeficiente de pendiente de -0.001 con un p-valor de 0.168, indicando que no hay una relación significativa entre las variables. El R² fue de 0.002, por lo que el modelo explica muy poca variabilidad.

## 2.2 Modelo SAR (Spatial Lag)

En el modelo SAR se obtuvo un coeficiente espacial (rho) de -3.217 con p-valor de 0.000. Aunque el parámetro espacial fue significativo, el modelo presentó problemas graves: R² igual a 0 y valores de AIC y log-likelihood no disponibles. El coeficiente de la variable independiente tampoco fue significativo (p = 0.249).

## 2.3 Modelo SEM (Spatial Error)

En el modelo SEM, el coeficiente espacial (lambda) fue de 1.238 con p-valor de 0.000, lo que indica una fuerte dependencia espacial en los errores. La variable independiente no fue significativa (p = 0.159), pero el modelo es estadísticamente estable y presentó un R² de 0.987. El test de Breusch-Pagan no detectó heterocedasticidad (p = 0.085).

## Conclusión:

El modelo SEM fue el más adecuado, ya que mostró mejores resultados estadísticos, mayor estabilidad y un coeficiente espacial significativo. Aunque no se halló una relación directa entre crimen y discapacidad, sí se identificó una fuerte dependencia espacial.

# 3. Machine Learning

## 3.1 Aprendizaje Supervisado

Un ejemplo real es el uso de modelos de predicción de fraude en tarjetas de crédito. Estos modelos se entrenan con datos etiquetados (fraude o no fraude) y aprenden a predecir nuevas transacciones sospechosas con base en patrones históricos.

## 3.2 Aprendizaje No Supervisado

Un caso común es la segmentación de clientes en marketing usando clustering. Los algoritmos como K-means agrupan clientes según similitudes sin usar etiquetas previas, permitiendo descubrir patrones ocultos en los datos.

## 3.3 Aprendizaje por Refuerzo

Se usa en robótica o videojuegos, donde un agente aprende a tomar decisiones con base en recompensas y penalizaciones. Un ejemplo real es AlphaGo, que aprendió a jugar y ganar a humanos jugando miles de partidas simuladas.

# 4. Consumo de Agua en Zonas de Cultivo – Caso Driscoll’s

## 4.1 Fuente de Datos

Para estimar el consumo de agua en las zonas de cultivo de Driscoll’s, se propone usar datos de sensores climáticos remotos como los proporcionados por NASA POWER (https://power.larc.nasa.gov/), específicamente variables como evapotranspiración, humedad del suelo y precipitación. Estos datos se complementan con imágenes satelitales (como Landsat o Sentinel-2) para estimar la cobertura agrícola activa.

Ventajas: cobertura amplia, datos históricos y acceso gratuito.  
Limitaciones: resolución espacial baja (1 km o más), requiere validación local con sensores in situ.

## 4.2 Análisis del Consumo de Agua

Se puede combinar la superficie sembrada (obtenida de mapas geoespaciales) con datos de evapotranspiración estimada para calcular el volumen de agua usado por hectárea. Esto se visualiza con mapas de calor sobre las zonas de cultivo y gráficos de consumo mensual.

## 4.3 Propuesta de Algoritmo de Predicción

Para predecir el uso futuro de agua, se puede utilizar un modelo de Machine Learning supervisado. El procedimiento sería el siguiente:

1. Recolectar datos históricos de consumo de agua por zona (usando estimaciones de ET).  
2. Añadir variables climáticas: temperatura, humedad, precipitación, viento.  
3. Entrenar un modelo como Random Forest o XGBoost para predecir el consumo futuro.  
4. Validar el modelo con datos reales y ajustar hiperparámetros.  
5. Usar el modelo para simular distintos escenarios de cambio climático o expansión agrícola.

## 4.4 Escenario Ideal con Recursos Ilimitados

Si se contara con recursos ilimitados, se instalarían sensores de humedad y caudal en cada zona de cultivo. Además, se recolectarían datos satelitales de alta resolución diaria. Esto permitiría entrenar modelos LSTM (redes neuronales recurrentes) con series de tiempo para anticipar con precisión el uso de agua. Los resultados se presentarían en dashboards interactivos para agricultores y técnicos, con alertas de uso excesivo y recomendaciones basadas en predicción.