



FACULTAD
DE CIENCIAS
ECONÓMICAS



Universidad
Nacional
de Córdoba

Unidad 4. Herramientas para la visualización de datos

Docente: Mauricio Morán

Septiembre - Octubre 2023



Programa tentativo de la Unidad 4:

- **Clase 1:** Introducción y conceptos fundamentales. Tipos de datos. Tipos de gráficos. Buenas prácticas para la visualización de datos. Principales paquetes Python. Introducción a Matplotlib.
- **Clase 2:** Visualización de datos numéricos: datos, mapas, gráficos, uso de glyphs, gráficos apilados. Matplotlib-Pandas y Seaborn.
- **Clase 3:** Visualización de datos no numéricos: gráficos, redes, gráficos de componente-principal, árboles, escalado multidimensional. Presentación de scikit-learn: escalado - estandarización - normalización, PCA, árboles de decisión para clasificación y regresión.
- **Clase 4:** Reportes, dashboards y otros recursos. Comunicación de datos. Business Intelligence. Storytelling. Dashboards con Plotly.
- **Tutoria:** Desarrollo de actividad práctica.

Clase 3: Visualización de datos no numéricos.

Análisis de componentes principales

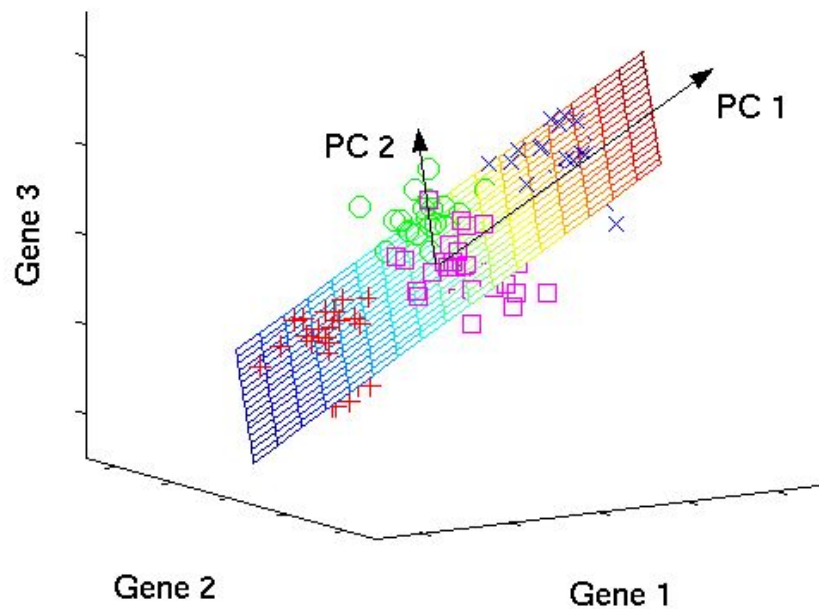


Principal component analysis (PCA)

Técnica estadística para el análisis multivariado y para la reducción de la dimensionalidad

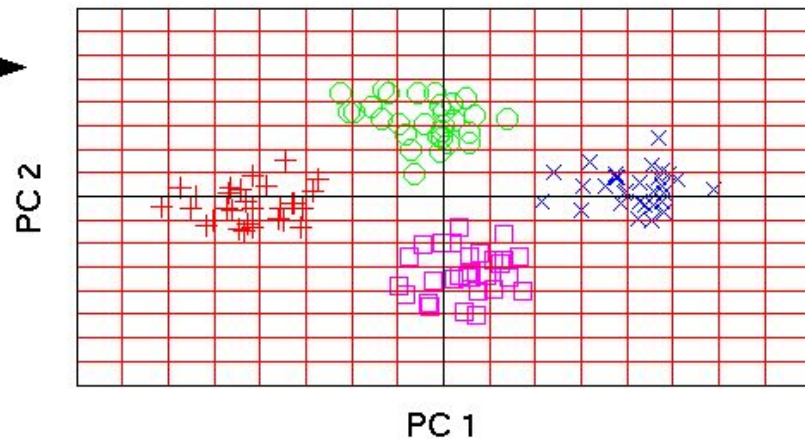
- Simplifica la complejidad en datasets grandes al reducir el número de variables.
- Transforma las variables originales en nuevas variables no correlacionadas llamadas **componentes principales** (PC).
- Las PC's son ortogonales (perpendiculares) entre sí.
- Se las ordena en función de la proporción de la variabilidad explicada (varianza).

original data space



PCA

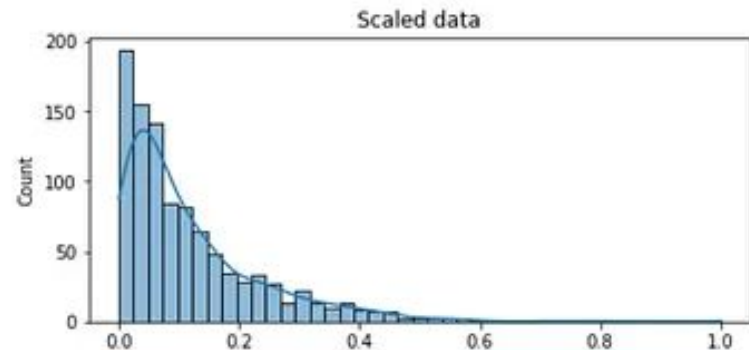
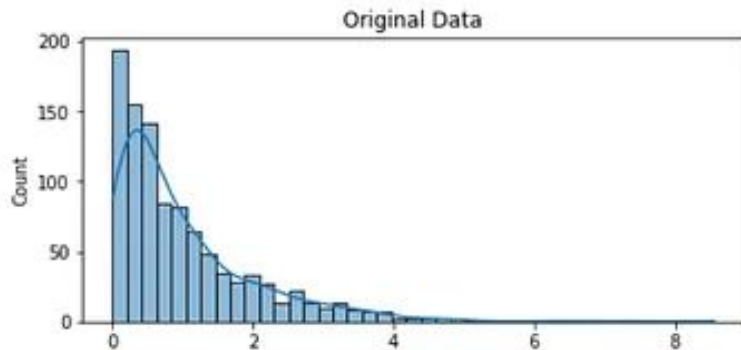
component space




Escalado / Estandarización / Normalización

Pre-procesado:

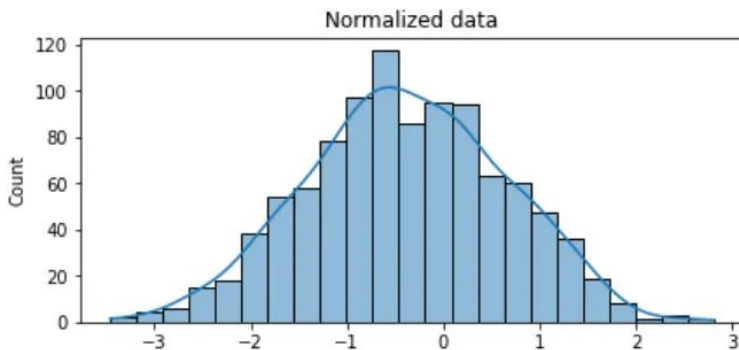
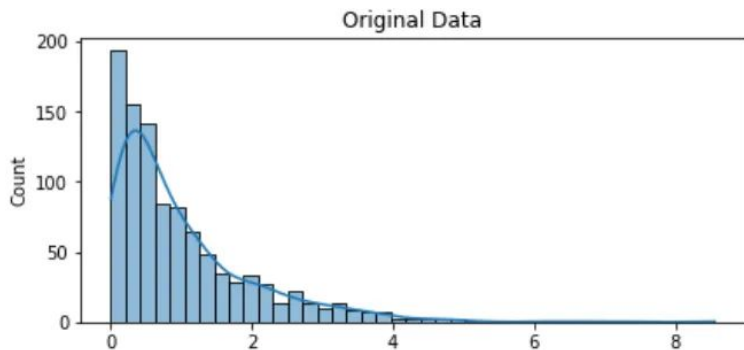
Escalado (Scaling): Transforma los datos para que estén dentro de un rango específico, es decir cambia el rango de los datos. Es útil para cuando las variables poseen escalas muy diferentes. Es útil para algoritmos basados en distancias, como KNN y PCA. Ejemplos: MinMaxScaler, MaxAbsScaler.





Estandarización (Standardization): Es un tipo de escalado que transforma los datos para que tengan una media de 0 y una desviación estándar de 1. Asume que las variables siguen una distribución normal. Ejemplos: StandarScaler.

Normalización (Normalization): Este procedimiento cambia la distribución de los datos para que puedan ser descritas como una distribución normal (gaussiana de media cero y desviación estándar 1). Ejemplos: PowerTransformer.

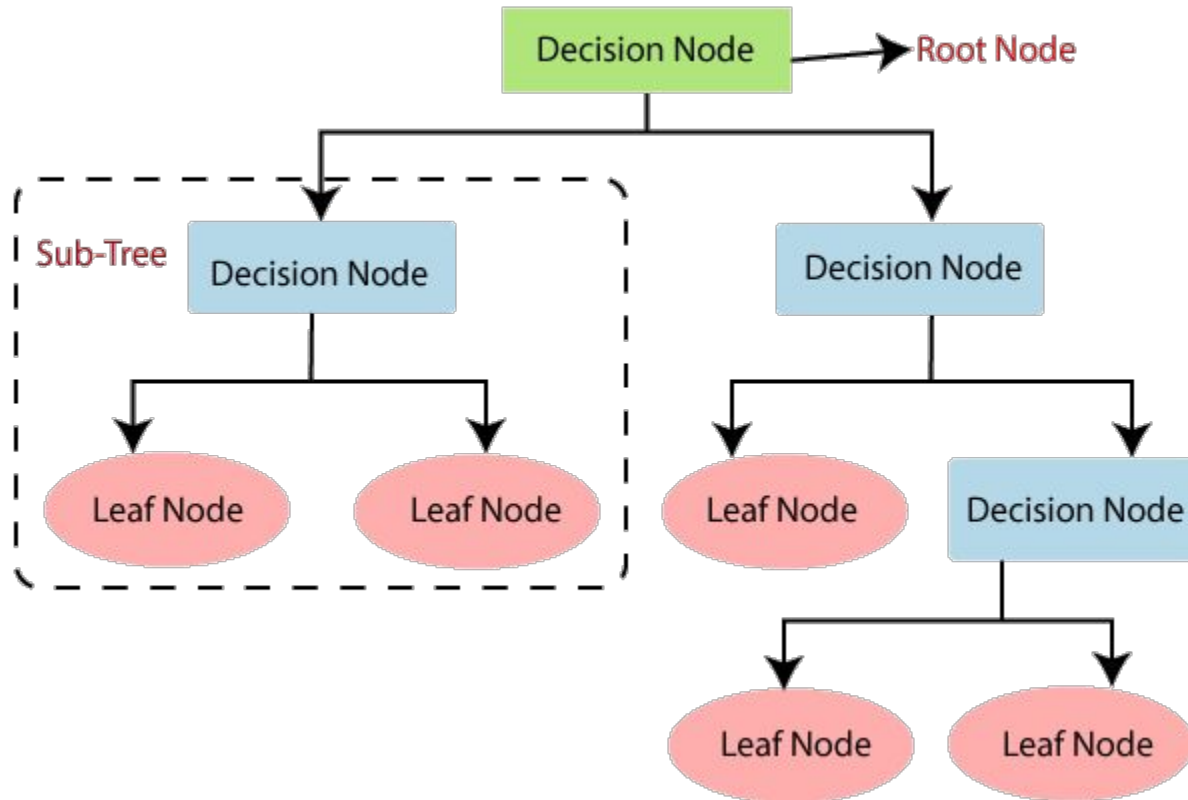


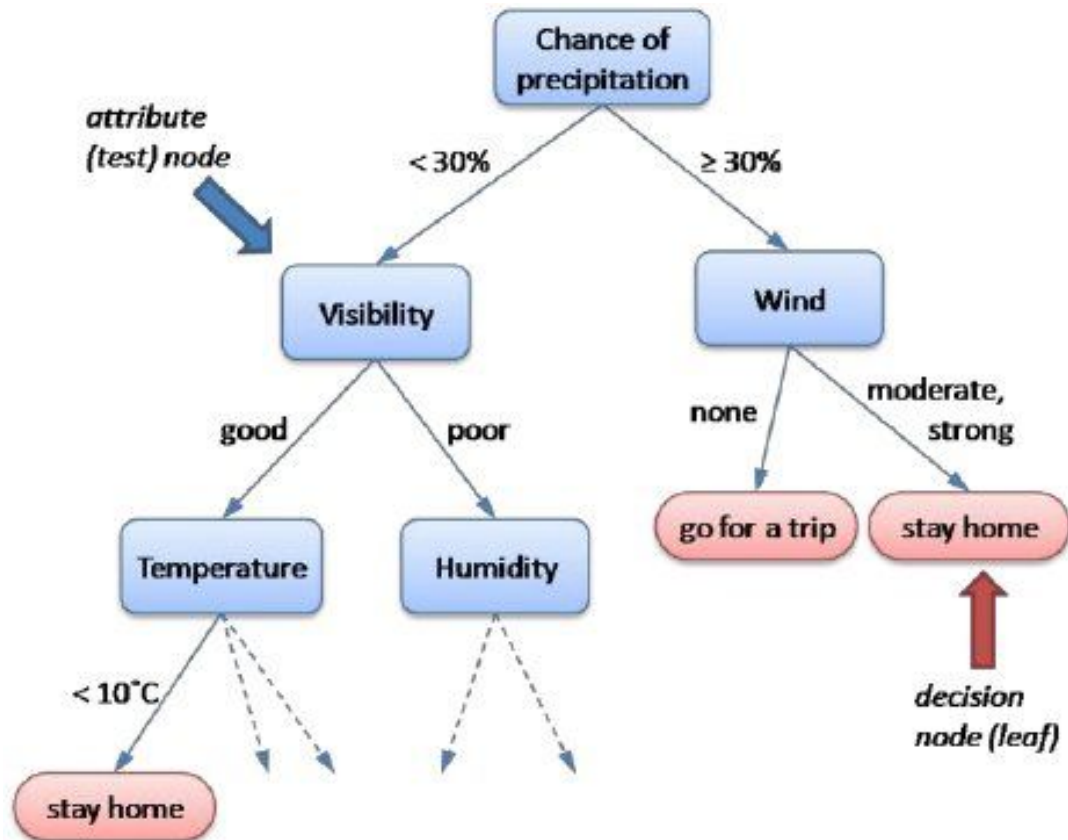
Árboles de decisión



Árboles de decisión

- Algoritmo de aprendizaje supervisado, que se utiliza tanto para tareas de clasificación como de regresión.
- Estructura jerárquica basada en la toma de decisiones, buscando puntos de división entre los datos:
 - Nodo raíz: representa el punto de inicio de la toma de decisiones.
 - Nodos intermedios: representan decisiones basadas en atributos específicos. Cada nodo intermedio tiene ramas que conducen a otros nodos o subnodos.
 - Nodos hoja: representan resultados finales o clasificaciones. No tienen ramas salientes y contienen la respuesta final o la etiqueta de clasificación.
- Las decisiones se toman en base a criterios tales como impureza gini o ganancia de información.
- Requieren poca preparación de datos y son fáciles de interpretar.
- Tienen rápidamente al sobreajuste y pueden ser costosos de entrenar.





iA programar!

