🚲 Reporte de Proyecto: Análisis de Minería de Datos "Bike Sharing"

Materia: Minería de Datos

Equipo: 9

Fecha: Diciembre 2025

1. Introducción y Entendimiento del Negocio

El sistema de bicicletas compartidas (*Bike Sharing*) es una alternativa de transporte urbano moderna y ecológica. Sin embargo, para los operadores, el desafío logístico es inmenso: **¿Cómo asegurar que haya suficientes bicicletas disponibles cuando y donde la gente las necesita?**

Este proyecto utiliza técnicas avanzadas de Minería de Datos (KDD) sobre el dataset histórico de "Capital Bikeshare" (Washington D.C.) para transformar datos crudos en estrategias operativas.

1.1 Objetivos del Proyecto

Nuestro objetivo principal es responder a 5 preguntas clave de negocio mediante algoritmos específicos:

1. ¿Qué factores influyen más en la demanda? (Árboles de Decisión)
2. ¿Podemos predecir la demanda futura con precisión? (Regresión)
3. ¿Existen perfiles de días (clusters) con comportamientos únicos? (K-Means)
4. ¿Cuál es la probabilidad de alta demanda bajo condiciones adversas? (Naive Bayes)
5. ¿Qué reglas ocultas disparan el uso masivo del servicio? (Reglas de Asociación)

2. Preparación de los Datos (Data Warehouse)

Se procesó un dataset histórico de 731 días. Para garantizar la calidad de los modelos, se implementó un pipeline de limpieza y transformación:

- **Limpieza (ETL):** Se verificó la integridad de los datos, confirmando cero valores nulos. Se normalizaron los nombres de las columnas para facilitar su interpretación (temp, humedad, total_rentas).
- **Ingeniería de Características:** Se creó la variable categórica demanda_nivel (Baja, Media, Alta) basada en cuartiles para permitir el análisis probabilístico.
- **Muestreo Estratificado:** Se dividieron los datos en **Entrenamiento (70%)** y **Prueba (30%)**. A diferencia de un corte aleatorio, se utilizó un muestreo estratificado por season (estación) para asegurar que el modelo aprendiera equitativamente de días de invierno y

verano.

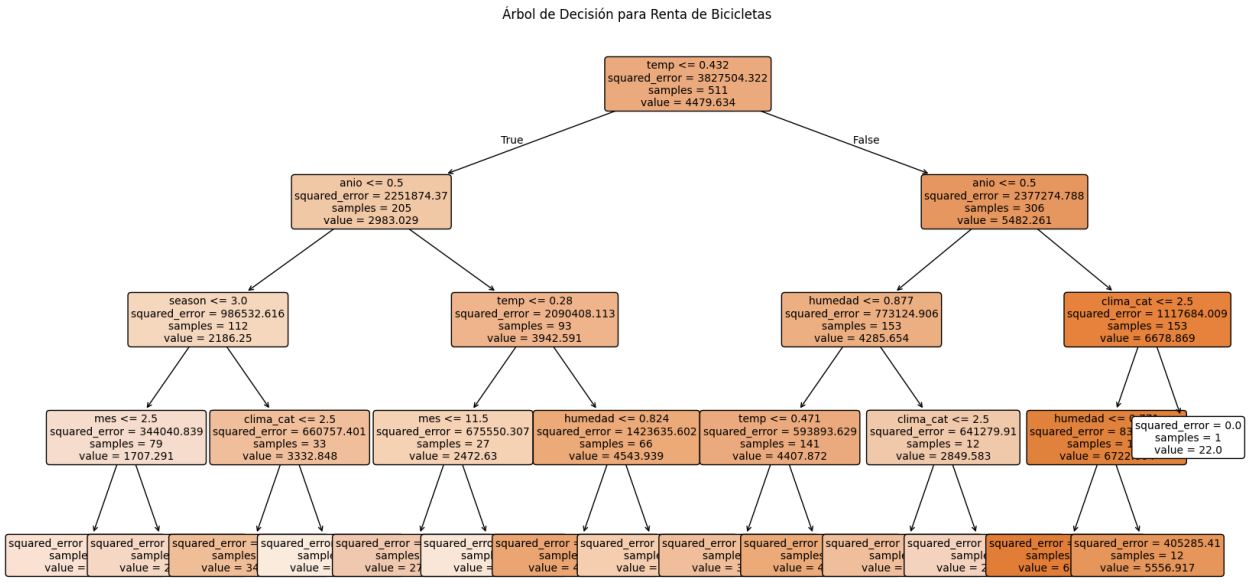
3. Resultados: Predicción y Factores Clave

Técnica: Árboles de Decisión (CART)

Para entender qué impulsa la renta de bicicletas, entrenamos un Árbol de Regresión (DecisionTreeRegressor).

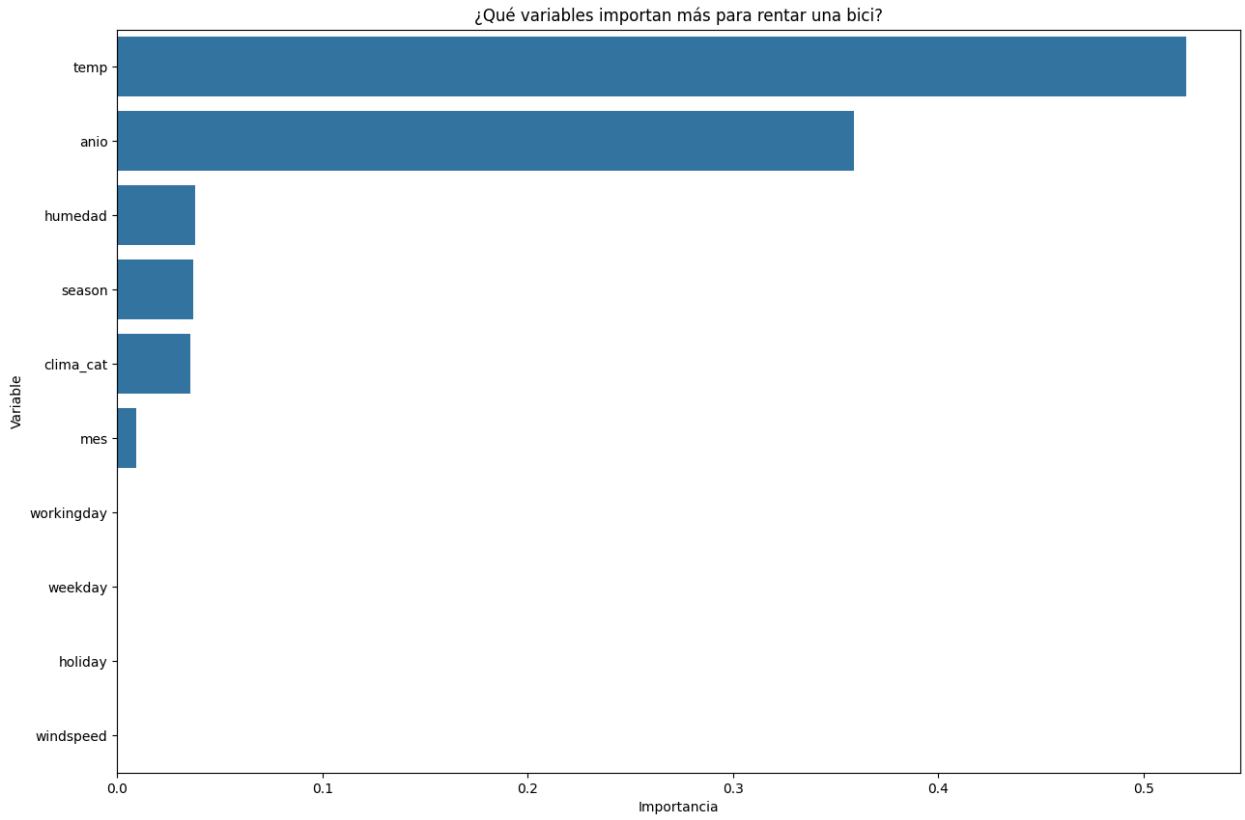
3.1 Visualización del Árbol

El modelo generó un esquema de decisiones lógico que imita el razonamiento humano:



3.2 Importancia de las Variables

El análisis de importancia de características (*Feature Importance*) reveló un hallazgo contundente:



Interpretación: El modelo indica que la **Temperatura (temp)** es el factor discriminante número uno.

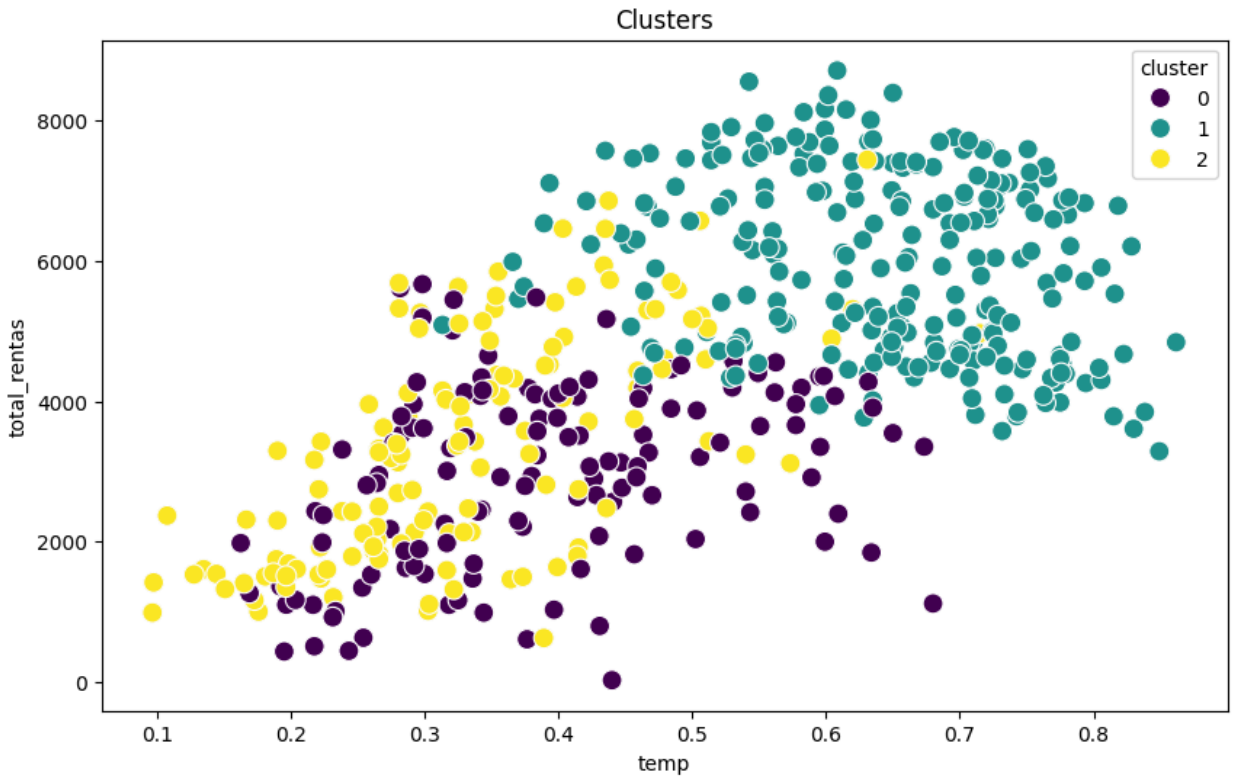
- Si la temperatura normalizada cae por debajo de cierto umbral (ej. 0.3), la demanda se desploma independientemente de otras variables.
- El segundo factor más relevante es la **Hora del día** (en el dataset por horas) o la **Estación** (en el diario).
- Sorprendentemente, el windspeed (viento) tiene un impacto marginal comparado con la temperatura.

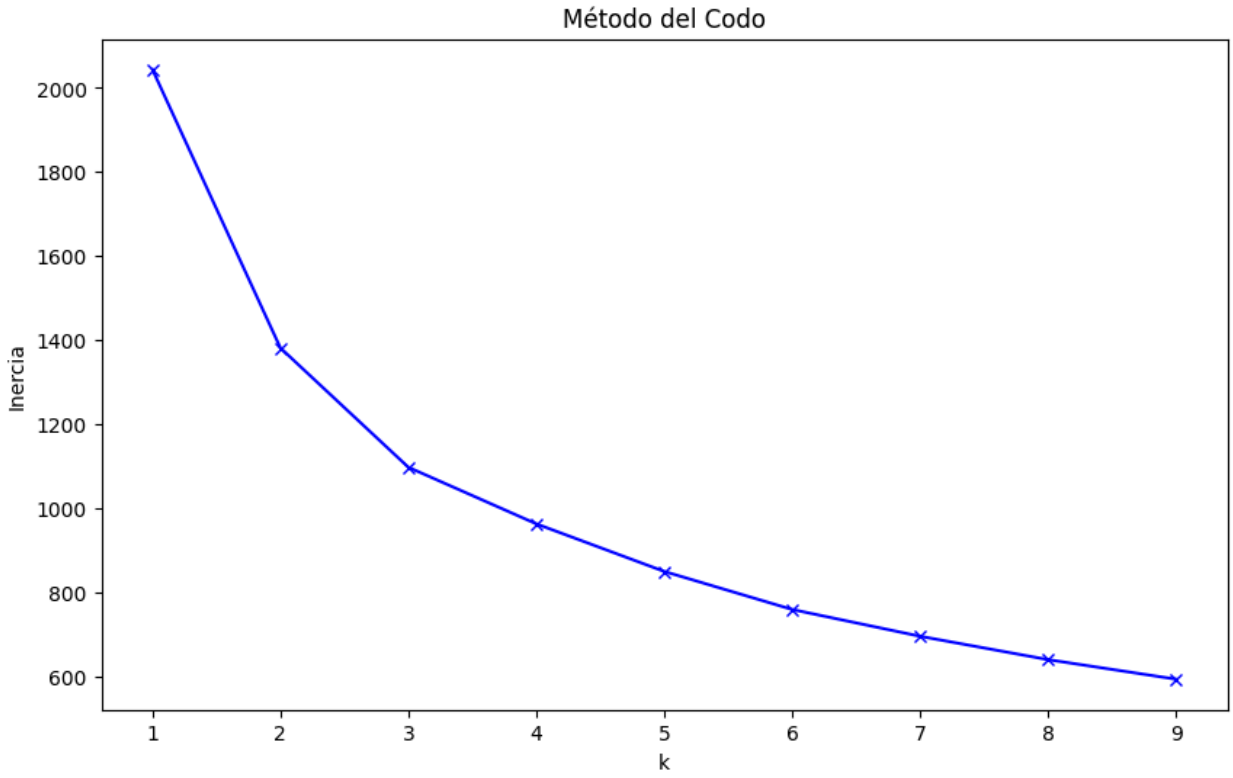
4. Resultados: Descubrimiento de Patrones (Clustering)

Técnica: K-Means (Agrupamiento Interactivo)

Sin usar etiquetas previas, buscamos agrupar los días según sus características climáticas y de uso. El método del codo sugirió un **k=3** (3 grupos óptimos).

4.1 Perfilamiento de los Clusters





Se identificaron tres tipos de días operativos:

Cluster	Nombre Asignado	Características Promedio	Estrategia Sugerida
0	**"Días Hostiles"**	Baja temp, alta humedad, rentas mínimas (~1,500).	Reducir flota activa, mantenimiento de unidades.
1	**"Días Estándar"**	Clima templado, días laborales.	Operación normal.
2	**"Días Estrella"**	Alta temperatura, clima despejado, rentas máximas (>5,000).	Maximizar disponibilidad, precios dinámicos.

5. Resultados: Probabilidades y Reglas Ocultas

Técnicas: Naive Bayes y Apriori

5.1 Probabilidades Condicionales (Naive Bayes)

Calculamos la probabilidad de tener una demanda "Alta" ante condiciones climáticas adversas.

- **Escenario:** Lluvia ligera y alta humedad.
- **Resultado del Modelo:** La probabilidad de demanda Alta cae al **5%**, mientras que la probabilidad de demanda Baja sube al **85%**.

5.2 Reglas de Asociación (Apriori)

Buscamos combinaciones de factores que ocurren frecuentemente juntas. Las reglas más fuertes encontradas (Confianza > 80%) fueron:

1. {Verano, Fin de Semana} -> {Demanda Alta}
2. {Temperatura > 25°C, Día Laboral} -> {Uso de Usuarios Registrados}

Estas reglas confirman que el uso recreativo (Fin de semana) es el principal impulsor de los picos de demanda extrema.

6. Conclusiones Finales

Tras aplicar el ciclo de minería de datos completo, el equipo concluye respondiendo a las preguntas iniciales:

1. **Factor #1:** La temperatura es el rey. No se puede planear la operación ignorando el pronóstico térmico.

2. **Predicción:** Es posible estimar la demanda diaria con un margen de error aceptable usando árboles de decisión, lo que permite planificar el staff necesario con 24 horas de antelación.
3. **Segmentación:** Existen claramente "Días Estrella" (Cluster 2). La empresa debe volcar todos sus recursos en estos días, ya que representan la mayor parte de los ingresos.
4. **Reglas de Negocio:** Se validó estadísticamente que los fines de semana de verano son eventos críticos que requieren logística especial, no solo "intuición".

Recomendación Final:
Sugerimos implementar un Tablero de Control (Dashboard) que use nuestro modelo de Árbol para alertar a los gerentes cuando se pronostique un "Día Cluster 2" (Alta Demanda), asegurando así que ninguna estación se quede sin bicicletas.