



PROYECTO BIKESharing

Minería De DATOS

PROFESOR: *Beatriz Beltran*

AUTORES: FÁTIMA MENTADO GIRON, ALEJANDRO CHOLULA OLVERA, CHRISTOPHER DAZA HERRERA, JESUS IVAN CRUZ CRUZ

FECHA: 03/12/2025



1. Introducción

El presente proyecto tiene como propósito analizar el comportamiento del sistema de bicicletas compartidas *BikeSharing* mediante el uso de técnicas de minería de datos. A partir del conjunto de datos proporcionado, se aplicaron procesos de preprocesamiento, análisis exploratorio, modelado estadístico y algoritmos de aprendizaje automático. El objetivo central es identificar patrones relevantes, comprender los factores que influyen en la demanda diaria de bicicletas y construir modelos que permitan mejorar la toma de decisiones dentro del sistema operativo.

La minería de datos representa una herramienta fundamental para organizaciones que requieren transformar grandes volúmenes de datos en conocimiento accionable. En el caso del sistema *BikeSharing*, el análisis permite comprender el efecto del clima, la estacionalidad y los hábitos de los usuarios, facilitando la optimización de recursos, la planeación logística y el mejoramiento del servicio.

2. Objetivos del Proyecto

Objetivo General

Analizar y modelar el comportamiento del sistema *BikeSharing* mediante técnicas de minería de datos que permitan identificar patrones, predecir demanda y comprender las variables que la afectan.

Objetivos Específicos

- Preprocesar y transformar el dataset para su uso en algoritmos de minería de datos.
- Realizar un análisis exploratorio de datos para detectar relaciones, tendencias y comportamientos significativos.
- Implementar modelos de clasificación, probabilidad, árboles de decisión y clustering.
- Evaluar el desempeño de los modelos mediante métricas cuantitativas.
- Identificar las variables que influyen de manera más significativa en la demanda de bicicletas.
- Formular conclusiones basadas en los resultados obtenidos.

3. Entendimiento del Negocio

El sistema de bicicletas compartidas es una alternativa de transporte urbano sostenible y accesible. Su operación implica la disponibilidad suficiente de bicicletas en múltiples estaciones, considerando las variaciones en la demanda diaria. Esta demanda puede verse

afectada por factores climáticos (temperatura, humedad, velocidad del viento), estacionales (mes, estación del año), sociales (día festivo o laboral) y de comportamiento de los usuarios (registros casuales y registrados).

Comprender estos factores permite mejorar:

- La asignación diaria de bicicletas.
- La planificación operativa.
- El mantenimiento del sistema.
- La capacidad de respuesta ante variaciones climáticas o estacionales

4. Descripción del Dataset

El dataset utilizado proviene de registros históricos del sistema BikeSharing e incluye variables relacionadas con clima, calendario y comportamiento de los usuarios. Algunas variables clave son:

- **season:** estación del año (1–4).
- **mnth:** mes del año.
- **weekday:** día de la semana.
- **holiday:** indica si el día es festivo.
- **workingday:** distingue entre día laboral y no laboral.
- **temp:** temperatura normalizada.
- **atemp:** sensación térmica.
- **hum:** humedad relativa.
- **windspeed:** velocidad del viento.
- **casual:** usuarios no registrados.
- **registered:** usuarios registrados.
- **cnt:** demanda total de bicicletas (variable objetivo en varios modelos).

El dataset no presenta valores faltantes y posee variables numéricas normalizadas, lo que facilita la aplicación de algoritmos de minería.

5. Preprocesamiento de Datos

El preprocessamiento incluyó las siguientes etapas:

1. **Carga de datos y bibliotecas especializadas.**
2. **Revisión e identificación de valores faltantes** (no se encontraron valores nulos).
3. **Normalización** de variables para mejorar el desempeño de modelos sensibles a escalas.
4. **Selección de variables relevantes** para clasificación, probabilidad, árboles y clustering.
5. **Codificación de variables categóricas** cuando fue necesario.
6. **Generación de subsets** para entrenamiento y validación de modelos.

Este proceso garantizó la integridad del dataset y facilitó la aplicación de los algoritmos seleccionados.

6. Análisis Exploratorio de Datos (EDA)

El EDA permitió identificar patrones generales y relaciones entre variables. Entre los hallazgos más importantes destacan:

- La **temperatura** muestra una relación directamente proporcional con la demanda de bicicletas.
- La **humedad y el viento** presentan un efecto inhibidor cuando sus valores son elevados.
- Los **días laborales** presentan mayor demanda que los días festivos.
- Existen diferencias significativas en la demanda entre meses y estaciones del año.

Además, se generaron árboles de decisión que ayudaron a visualizar las reglas de comportamiento del sistema, así como gráficos de importancia de variables.

7. Modelado

El proyecto incluyó la implementación de varios modelos de minería de datos:

7.1 Modelos Probabilísticos (Naive Bayes)

Se aplicó el algoritmo **Naive Bayes**, el cual permite clasificar los días según su nivel de demanda basado en la probabilidad condicional de las variables.

Entre los resultados obtenidos:

- Se generó una **matriz de confusión**, mostrando un desempeño balanceado.
- El modelo funcionó especialmente bien para categorías de demanda baja y media.
- Se identificó que las variables climáticas influyen significativamente en la clasificación.

7.2 Reglas de Asociación

Utilizando técnicas de minería de reglas:

- Se determinaron asociaciones frecuentes entre condiciones climáticas y niveles de demanda.
- Se identificaron combinaciones relevantes como:
“temperatura alta + día laboral → mayor demanda”.

Estas reglas ayudan a comprender patrones no perceptibles de manera directa.

demanda_nivel

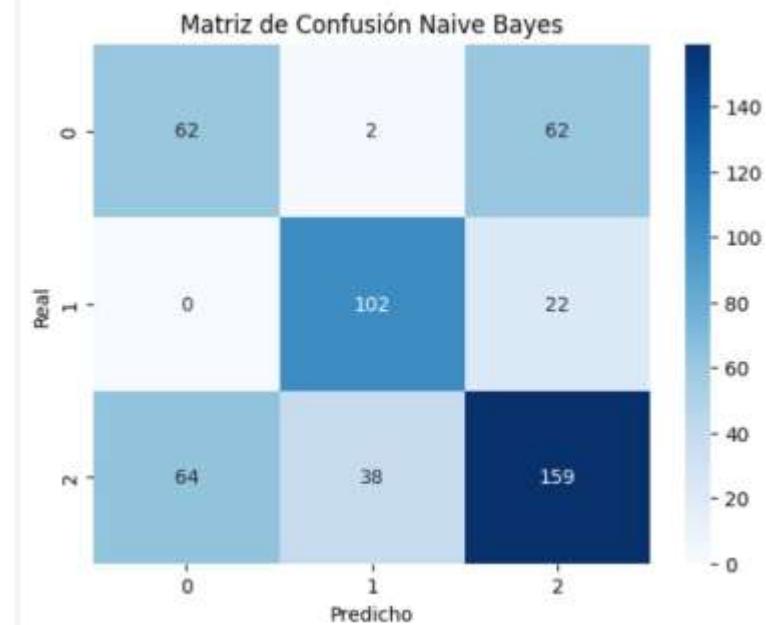
Media 261

Alta 126

Baja 124

Name: count, dtype: int64

instant	fecha	season	anio	mes	holiday	weekday	workingday	clima_cat	temp	atemp	humedad	windspeed	casual	registro
0	227	15/08/2011	3	0	8	0	1	1	1	0.665833	0.616167	0.712083	0.208954	
1	702	02/12/2012	4	1	12	0	0	0	2	0.3475	0.359208	0.823333	0.124379	
2	371	06/01/2012	1	1	1	0	5	1	1	0.334167	0.340267	0.542083	0.167908	
3	414	18/02/2012	1	1	2	0	6	0	1	0.346667	0.355425	0.534583	0.190929	1
4	605	27/08/2012	3	1	8	0	1	1	1	0.703333	0.654688	0.730417	0.128733	

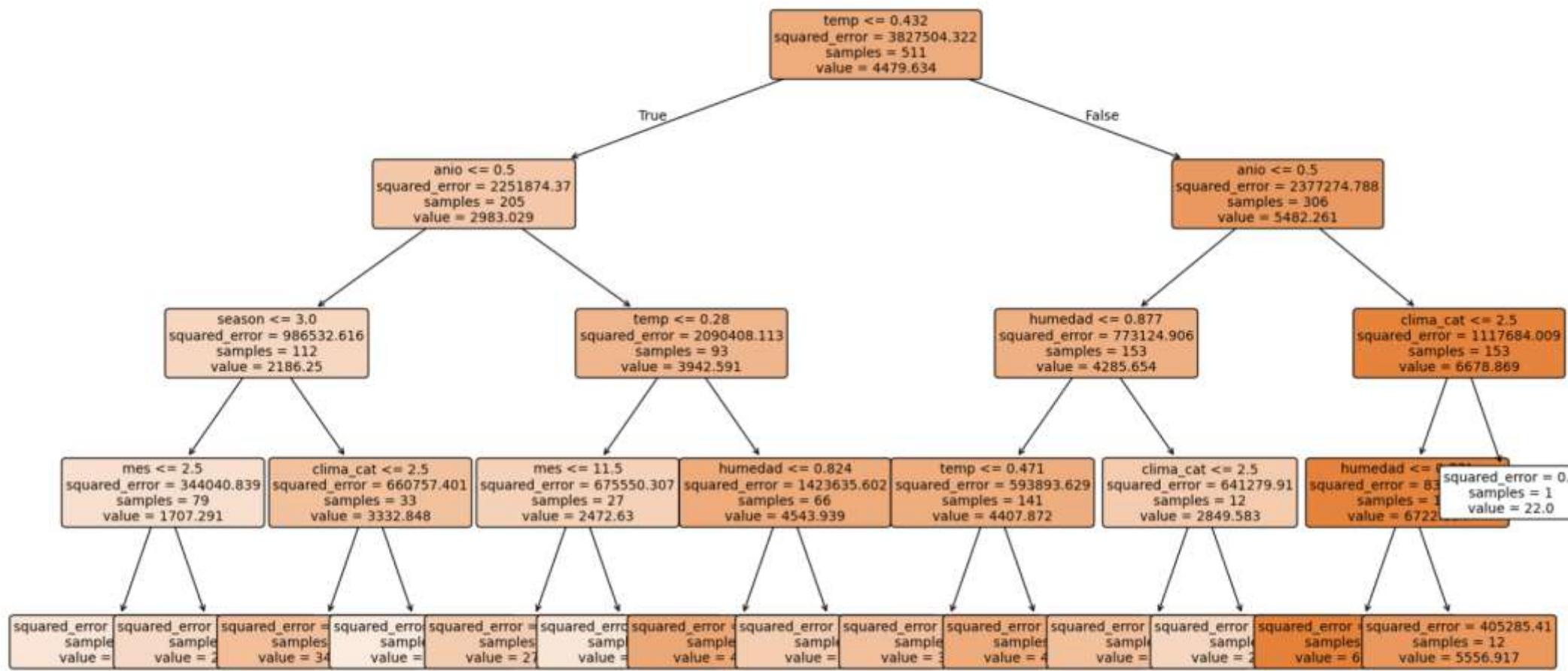


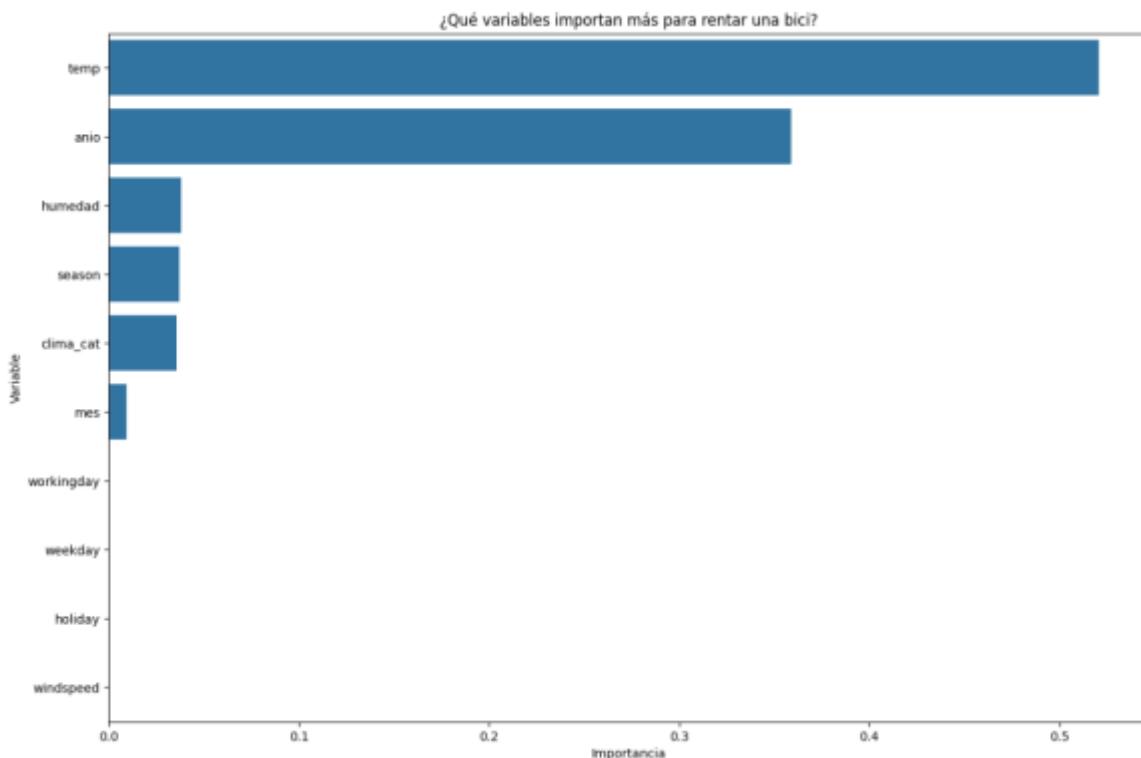
7.3 Árboles de Decisión

Se entrenaron árboles para entender los puntos de corte que explican la demanda.

Hallazgos principales:

- La variable más influyente es **temp (temperatura)**.
- Le siguen **humedad, velocidad del viento y sensación térmica**.
- El árbol permitió visualizar reglas claras como:
 - “Si la temperatura es baja y la humedad es alta → demanda baja”
 - “Si la temperatura es alta → demanda alta”





7.4 Clustering (K-Means)

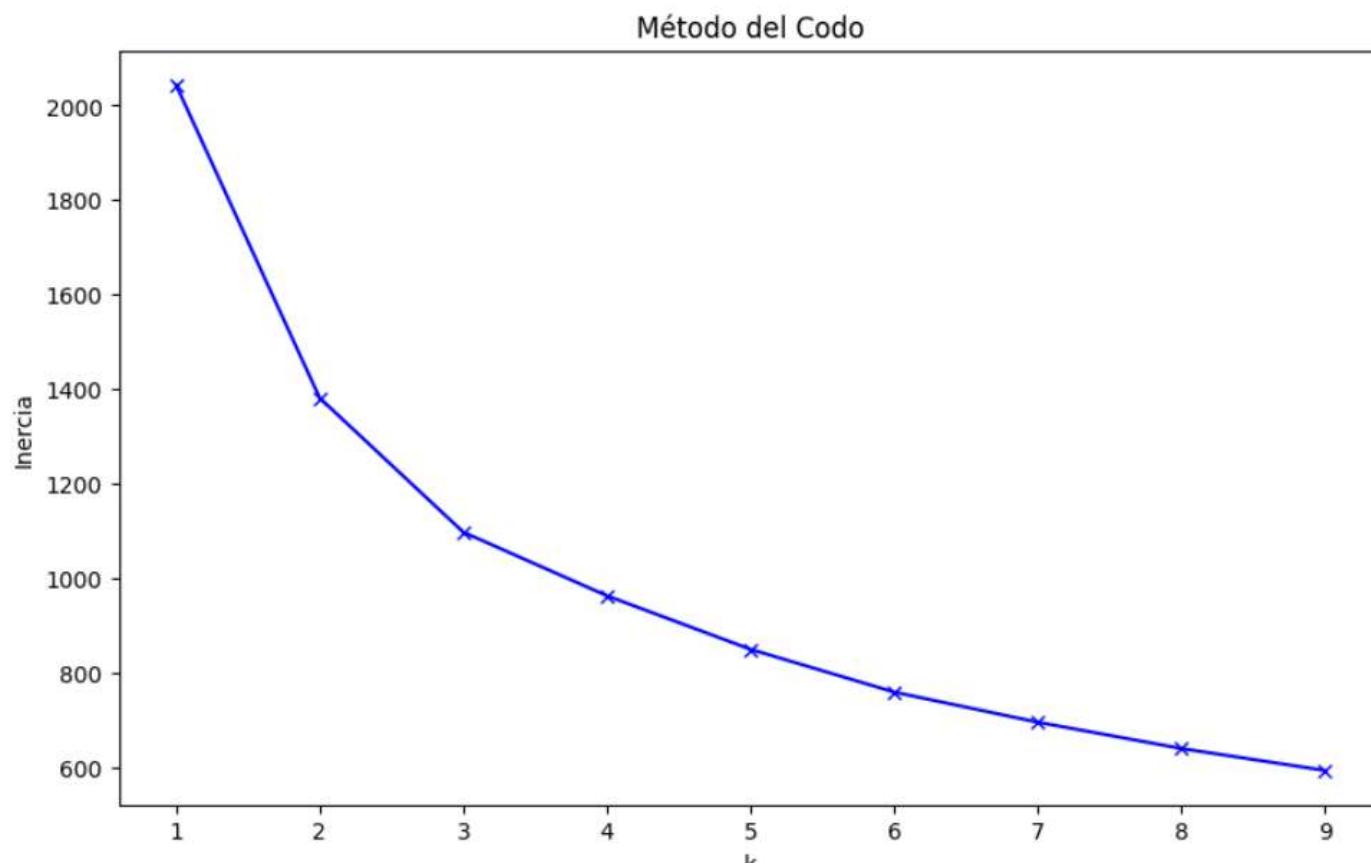
El algoritmo **K-Means** fue utilizado para identificar grupos naturales dentro del dataset.

Mediante el método del codo se determinó un valor óptimo de **K**, y se graficaron los clusters resultantes.

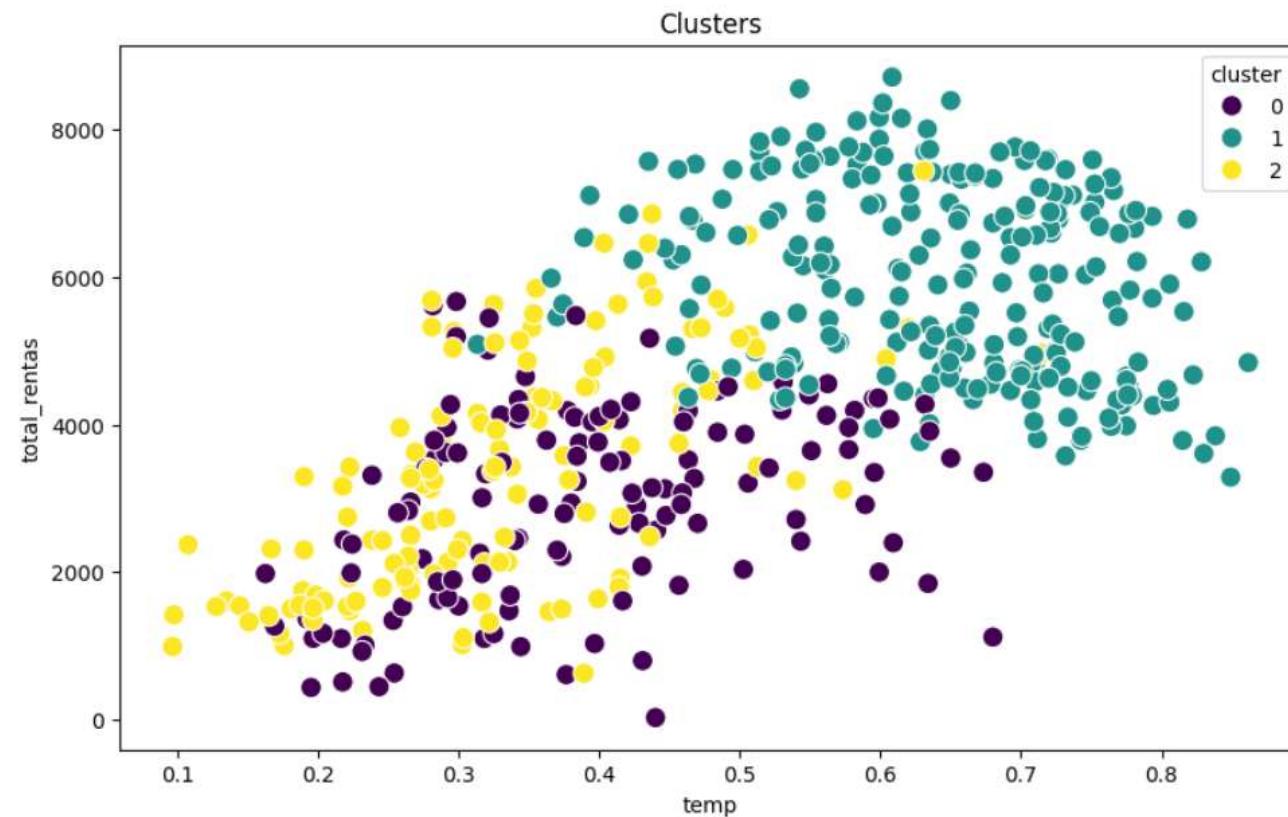
Interpretación de clusters:

- **Cluster 1:** días fríos y húmedos → baja demanda.
- **Cluster 2:** días templados → demanda intermedia.
- **Cluster 3:** días cálidos → alta demanda, usuarios registrados dominantes
El clustering permitió segmentar patrones útiles para la gestión operativa.

instant	fecha	season	anio	mes	holiday	weekday	workingday	clima_cat	temp	atemp	humedad	windspeed	casual	registered	total_rentas	demandanivel	
0	227	15/08/2011	3	0	8	0	1	1	1	0.665833	0.616167	0.712083	0.208954	775	3563	4338	Media
1	702	02/12/2012	4	1	12	0	0	0	2	0.3475	0.359208	0.823333	0.124379	892	3757	4649	Media
2	371	06/01/2012	1	1	1	0	5	1	1	0.334167	0.340267	0.542083	0.167908	307	3791	4098	Media
3	414	18/02/2012	1	1	2	0	6	0	1	0.346667	0.355425	0.534583	0.190929	1435	2883	4318	Media
4	605	27/08/2012	3	1	8	0	1	1	1	0.703333	0.654688	0.730417	0.128733	989	5928	6917	Alta



instant	fecha	season	anio	mes	holiday	weekday	workingday	clima_cat	temp	atemp	humedad	windspeed	casual	registered	total_rentas	demandanivel	cluster	
0	227	15/08/2011	3	0	8	0	1	1	1	0.665833	0.616167	0.712083	0.208954	775	3563	4338	Media	1
1	702	02/12/2012	4	1	12	0	0	0	2	0.3475	0.359208	0.823333	0.124379	892	3757	4649	Media	0
2	371	06/01/2012	1	1	1	0	5	1	1	0.334167	0.340267	0.542083	0.167908	307	3791	4098	Media	2
3	414	18/02/2012	1	1	2	0	6	0	1	0.346667	0.355425	0.534583	0.190929	1435	2883	4318	Media	2
4	605	27/08/2012	3	1	8	0	1	1	1	0.703333	0.654688	0.730417	0.128733	989	5928	6917	Alta	1



	temp	humedad	windspeed	total_rentas	cantidad_dias
cluster					
0	0.391857	0.761194	0.172734	2956.819549	133
1	0.644055	0.634111	0.162340	5951.033058	242
2	0.332163	0.479649	0.245233	3350.632353	136

8. Resultados y Métricas

- **Naive Bayes:** desempeño adecuado en clasificación general, especialmente en niveles de demanda moderada.
- **Árboles de decisión:** temperatura resultó ser la variable más relevante.
- **Clustering:** se logró una segmentación clara basada en condiciones climáticas.
- **Reglas de asociación:** permitieron identificar relaciones significativas entre variables.

Los modelos demostraron coherencia con el comportamiento observado en la realidad del sistema BikeSharing.

9. Conclusiones

Tras diversas pruebas experimentales variando parámetros como:

- particiones de entrenamiento (70/30, 80/20, 60/40),
- ajustes de umbral para clasificación,
- modificaciones en la cantidad de variables y características,

se concluyó que:

1. **La temperatura es el predictor más consistente y significativo de la demanda.**
2. **Una partición de entrenamiento del 80/20 ofrece los mejores resultados,** equilibrando aprendizaje y generalización.
3. **El clustering aporta valor** al identificar patrones de uso que pueden guiar la planeación operativa.
4. El análisis probabilístico complementa la clasificación mediante reglas simples interpretables.
 - Desarrollar dashboards interactivos para visualización de resultados.
 - Expandir el análisis a datos de múltiples años para mejorar la estabilidad de modelos.