Asignatura	Datos del alumno	Fecha
Aprendizaje Automático y Minería de Datos	Apellidos: Jiménez Acosta	
	Nombre: Ronaldo	

Trabajo: Lectura de datos y análisis descriptivo

Introducción

El objetivo de este trabajo es comprender la estructura y características del Bike Sharing Dataset, extraído de la UCI Machine Learning Repository (ID=275). Este dataset contiene datos sobre el alquiler de bicicletas durante los años 2011 y 2012 en el sistema Capital Bikeshare. La variable respuesta es cnt (cantidad total de bicicletas alquiladas), y el resto de variables (como temp, hr, season, etc.) actúan como predictores que influyen en la demanda de bicicletas.

Metodología

Para el análisis se ha utilizado Python y el paquete ucimlrepo para importar directamente el dataset, sin necesidad de descargar manualmente el CSV. Se siguieron los siguientes pasos:

1. **Importación y Exploración Inicial:**

Se cargan los datos y se combinan las variables predictoras y la respuesta en un único DataFrame. Se revisa la estructura, el tipo de cada variable y se visualizan las primeras filas y estadísticas descriptivas.

2. **División en Conjunto de Modelización y Validación:**

Se separa el DataFrame en dos subconjuntos, destinando el 70% para modelización y el 30% para validación, lo que es clave para la futura evaluación de modelos predictivos.

3. **Tratamiento de Missing Values:**

Se verifica la presencia de datos faltantes. En este caso, el dataset no presenta missing values, pero se incluye este paso para garantizar la calidad del análisis.

4. **Análisis de Correlaciones:**

Se calcula la matriz de correlación entre las variables. Dado que la columna de fechas (dteday) es de tipo string, se excluye del análisis para evitar errores. Se presta especial atención a la relación entre los predictores y la variable respuesta (cnt).

5. **Análisis de Distribuciones y Visualización:**

Se generan gráficos para explorar la distribución de cnt, la relación entre temp y cnt, y se analiza cómo varía cnt según la estación del año (season). Estos gráficos permiten identificar tendencias y posibles relaciones significativas.

Código del Proyecto

Comparto el repositorio del proyecto, el cual está en GitHub:
https://github.com/JimcostDev/Mis_Apuntes_Unir/tree/master/04_Curso/Machine_Learning/lab-1

Resultados y Discusión

Exploración inicial:

El DataFrame resultante contiene 17,379 registros y 14 columnas. Las estadísticas descriptivas muestran una amplia variabilidad en cnt, lo cual indica diferencias significativas en la demanda de bicicletas a lo largo del tiempo.

División del dataset:

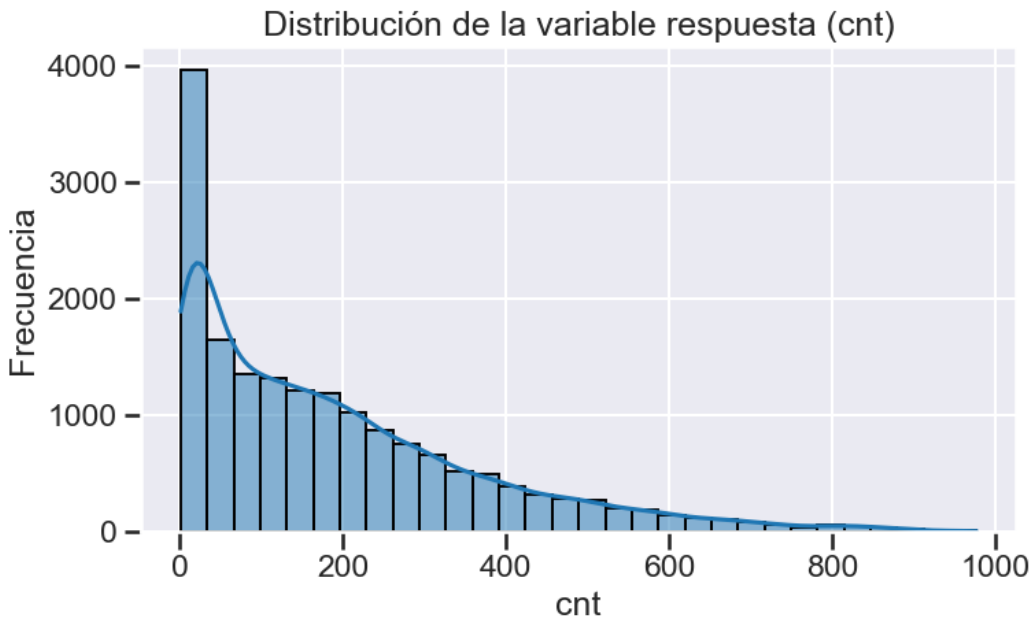
La separación en entrenamiento y validación generó conjuntos de tamaños adecuados para futuros modelos predictivos (aproximadamente 70% y 30% respectivamente).

Análisis de correlaciones:

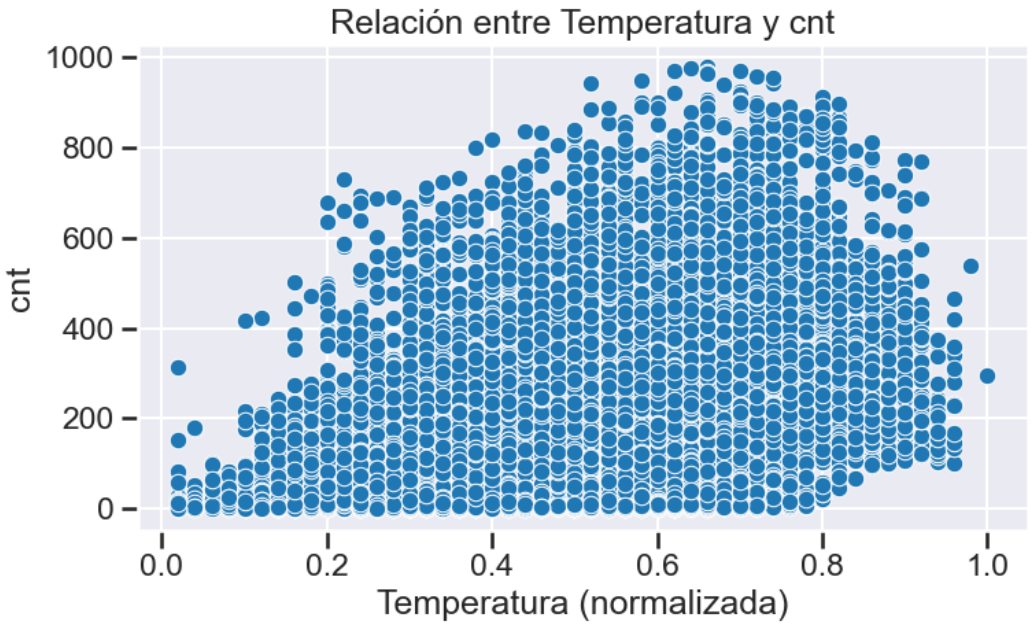
Al calcular la matriz de correlación (excluyendo la columna de fechas), se observa que algunas variables, como la temperatura (temp y atemp) y la hora (hr), presentan correlaciones notables con la variable respuesta cnt. Esto sugiere que a mayor temperatura o en determinadas franjas horarias, la demanda de bicicletas tiende a aumentar. Asimismo, la variable season también muestra diferencias importantes, lo que refuerza la hipótesis de que la estacionalidad influye en el comportamiento de la demanda.

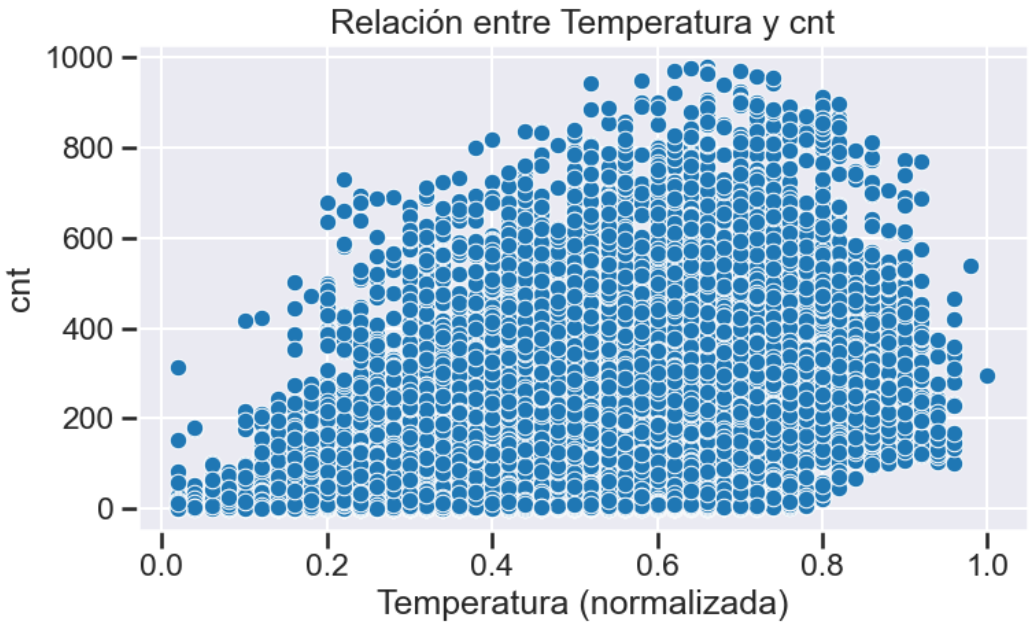
Análisis de distribuciones y gráficos:

- El histograma de cnt revela que su distribución es asimétrica, con una cola larga hacia valores altos, lo cual es común en datos de demanda.
- El gráfico de dispersión entre temp y cnt evidencia una tendencia ascendente, indicando que el aumento en la temperatura (normalizada) se asocia con un incremento en el número de bicicletas alquiladas.
- El boxplot de cnt por season muestra diferencias notables entre estaciones, lo que confirma que el comportamiento de la demanda varía significativamente según la época del año.



Asignatura	Datos del alumno	Fecha
Aprendizaje Automático y Minería de Datos	Apellidos: Jiménez Acosta	
	Nombre: Ronaldo	





Conclusiones

El análisis descriptivo del Bike Sharing Dataset ha permitido identificar características clave y relaciones relevantes entre las variables predictoras y la variable respuesta cnt. La temperatura, la hora y la estacionalidad emergen como factores importantes

que influyen en el número de bicicletas alquiladas. Estos hallazgos son fundamentales para la posterior aplicación de técnicas de modelización predictiva, ya que ayudan a seleccionar y transformar adecuadamente las variables que afectan la demanda.
El código utilizado se encuentra debidamente comentado, y cada paso del análisis ha sido fundamentado con observaciones sobre los resultados obtenidos.

Referencias

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset