# Proyecto final de introducción a ciencia de datos

# Análisis del sistema de transporte universitario, ruta CU-CU2

Marín Nieva Josué Salvador

Riveroll Martínez Danna Patricia

Materia: Introducción a la Ciencia de Datos

Profesor: Jaime Alejandro Romero Sierra

Fecha de entrega: 25 de noviembre de 2024

# Introducción

El objetivo principal del proyecto es analizar y comprender las dificultades del sistema de transporte universitario en la ruta CU-CU2 para optimizar su funcionamiento, reducir los tiempos de espera y mejorar la experiencia del usuario, con un enfoque en incrementar la capacidad de transportación y reducir el impacto ambiental al aprovechar mejor los recursos disponibles.

El sistema de transporte universitario (STU) es una herramienta esencial para los estudiantes, ya que conecta Ciudad Universitaria (CU) y Ciudad Universitaria 2 (CU2). Actualmente, este sistema enfrenta problemas como la saturación de usuarios, largas filas y demoras en los traslados, lo que afecta el desempeño académico al causar llegadas tardías.

Además, el uso ineficiente de las unidades no solo impacta a los usuarios, sino también al medio ambiente, ya que los vehículos recorren la ruta sin aprovechar su capacidad máxima. Resolver esta problemática no solo beneficiará la puntualidad y calidad del servicio, sino también contribuirá a disminuir la huella de carbono, alineándose con objetivos de sostenibilidad.

La base de datos fue construida a partir de encuestas aplicadas a los usuarios del sistema de transporte universitario, enfocándose en recopilar información clave sobre su experiencia y patrones de uso. Características principales:

Origen: Encuestas realizadas a estudiantes que utilizan el STU.

Cantidad de datos: Aproximadamente 3000 datos

Columnas principales:

Frecuencia de uso: ¿Cuántos días utilizan el STU a la semana?

Horarios de uso: ¿En qué horarios utilizan el STU?

Modalidad de uso: ¿Utilizan el STU de ida y vuelta?

Satisfacción: ¿Cuál es su nivel de satisfacción con el servicio?

Tiempo de espera: ¿Cuánto tiempo esperan para abordar una unidad?

La estructura de estos datos permite identificar horarios pico, medir niveles de satisfacción, y determinar patrones de uso para sugerir mejoras basadas en evidencia.

# Metodología

## Proceso de limpieza de datos

Se identificaron las columnas con valores mayormente nulos o columnas irrelevantes para el análisis fueron eliminadas, para las columnas con valores nulos parciales se utilizaron estrategias como imputación con la media mediana o moda dependiendo del tipo de dato; se identificaron 142 filas duplicadas que fueron eliminadas para evitar redundancia y sesgos dentro del análisis.

En las columnas "¿A qué hora tomas el STU de CU a CU2?" y "¿A qué hora tomas el STU de CU2 a CU?" se separaron los valores concatenados por punto y coma y se ajustaron a un formato estándar; las variables relacionadas con frecuencia y satisfacción que se encontraban inicialmente en un formato de texto fueron convertidas a enteros para facilitar su análisis, finalmente se homogenizaron los formatos de fechas, tiempo de espera y días.

Para el manejo de outliers se utilizaron complots y métodos estadísticos para detectar valores fuera de rango y se evaluó cada caso para decidir si los outliers reflejaban errores de entrada o requerían mantenerse.

Se simplificaron nombres de las columnas con caracteres inusuales o largos para mejorar la claridad legibilidad y manejo.

## Análisis exploratorio de datos (EDA)

### Descripción general de los datos

El dataset contenía 3,063 filas y 17 columnas, después se realizó un regex el cual nos daba un total de 6,339 filas y 10 columnas de las cuales con procesos ETL se redujeron a 3,426 filas x 10 columnas, lo que nos da un total de 34,260 datos.

COLUMNAS

1. Días: Valor tipo entero (int) el cual nos indica el número de días que aborda el STU a la semana, en un intervalo de uno a cinco [1,5].
2. Ida/Vuelta: Valor tipo entero (int) esta columna nos indica que "Si", "Solo de ida", "Solo de vuelta", representados por valores numéricos para su análisis [1,2,3].
3. Hora Cu-Cu2: Valor tipo entero (int) nos muestra las horas en las que abordan el STU de Cu a Cu2, es tipo entero para llevar a cabo su análisis.
4. Hora Cu2-Cu: Valor tipo entero(int) en este apartado nos indica de igual manera la hora que abordan el STU, pero en esta ocasión de Cu2 a Cu. De igual manera el valor es tipo entero para su análisis.

5. Satisfacción: Valor tipo entero (int), en esta columna nos indica que tan satisfecho esta con el Transporte universitario en un intervalo de uno a cinco [1,5] indicando que tan satisfecho esta.
6. Espera Cu-Cu2: Valor de tipo entero (int), en esta sección nos indica los tiempos de espera para abordar el STU de Cu a Cu2.
7. Espera Cu2-Cu: Valor tipo (int), contiene los tiempos de espera para abordar de Cu2 a Cu
8. Frecuencia parado: Valor tipo entero (int), en este apartado esta la frecuencia en la que llegas a estar parado el cual tiene un intervalo que va desde uno a cinco [1,5],

## Estadísticas generales

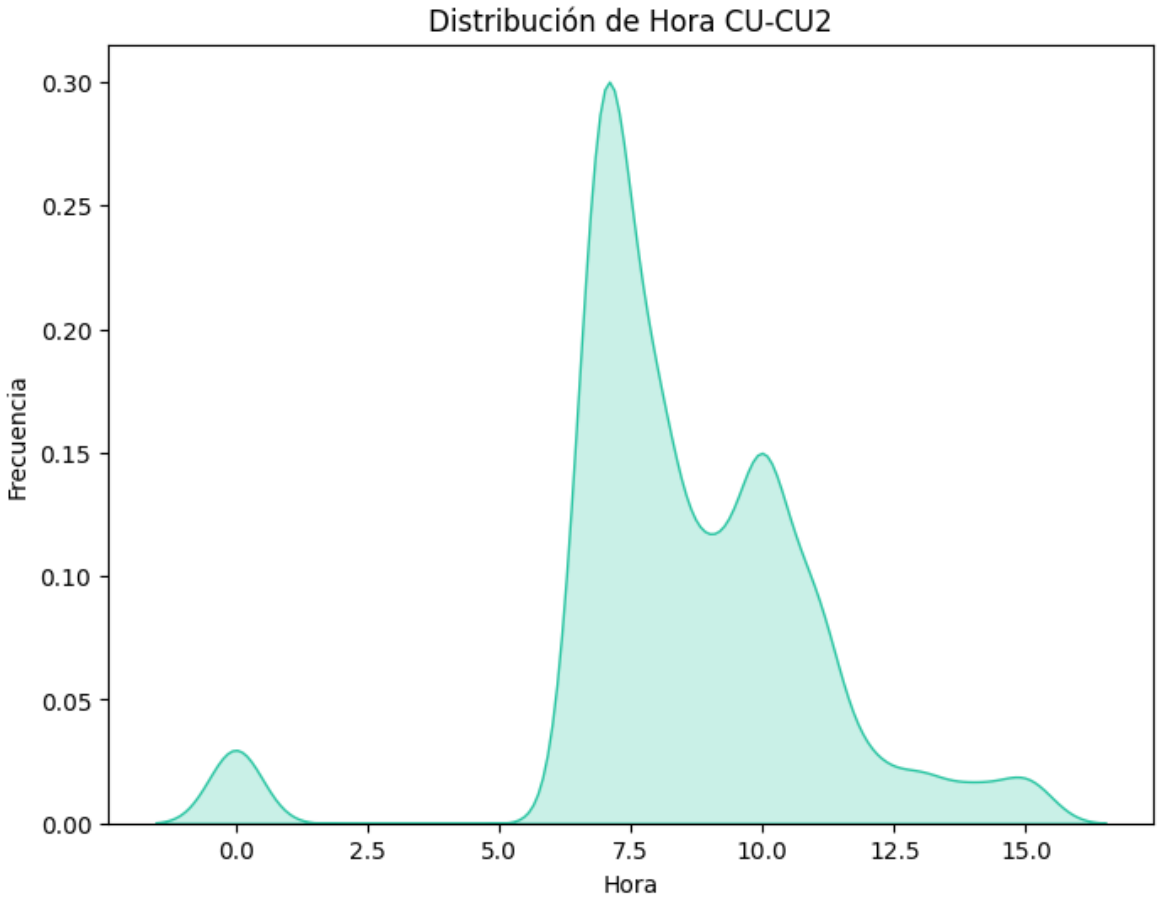
| | Dias | Ida/Vuelta | Hora Cu-Cu2 | Hora Cu2-Cu |
|---|---|---|---|---|
| count | 3426.000000 | 3426.000000 | 3426.000000 | 3426.000000 |
| mean | 4.773205 | 0.113543 | 8.500000 | 15.169294 |
| std | 0.588151 | 0.429847 | 2.573787 | 2.595437 |
| min | 1.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 5.000000 | 0.000000 | 7.000000 | 14.000000 |
| 50% | 5.000000 | 0.000000 | 8.000000 | 15.000000 |
| 75% | 5.000000 | 0.000000 | 10.000000 | 17.000000 |
| max | 5.000000 | 2.000000 | 15.000000 | 19.000000 |

| | Satisfaccion | Espera Cu-Cu2 | Espera Cu2-Cu | Frecuencia parado |
|---|---|---|---|---|
| count | 3426.000000 | 3426.000000 | 3426.000000 | 3426.000000 |
| mean | 2.265616 | 29.503503 | 43.098365 | 4.004962 |
| std | 0.808764 | 15.046366 | 14.356624 | 0.906572 |
| min | 1.000000 | 0.000000 | 0.000000 | 1.000000 |
| 25% | 2.000000 | 15.000000 | 30.000000 | 3.000000 |
| 50% | 2.000000 | 30.000000 | 45.000000 | 4.000000 |
| 75% | 3.000000 | 44.000000 | 60.000000 | 5.000000 |
| max | 5.000000 | 60.000000 | 60.000000 | 5.000000 |

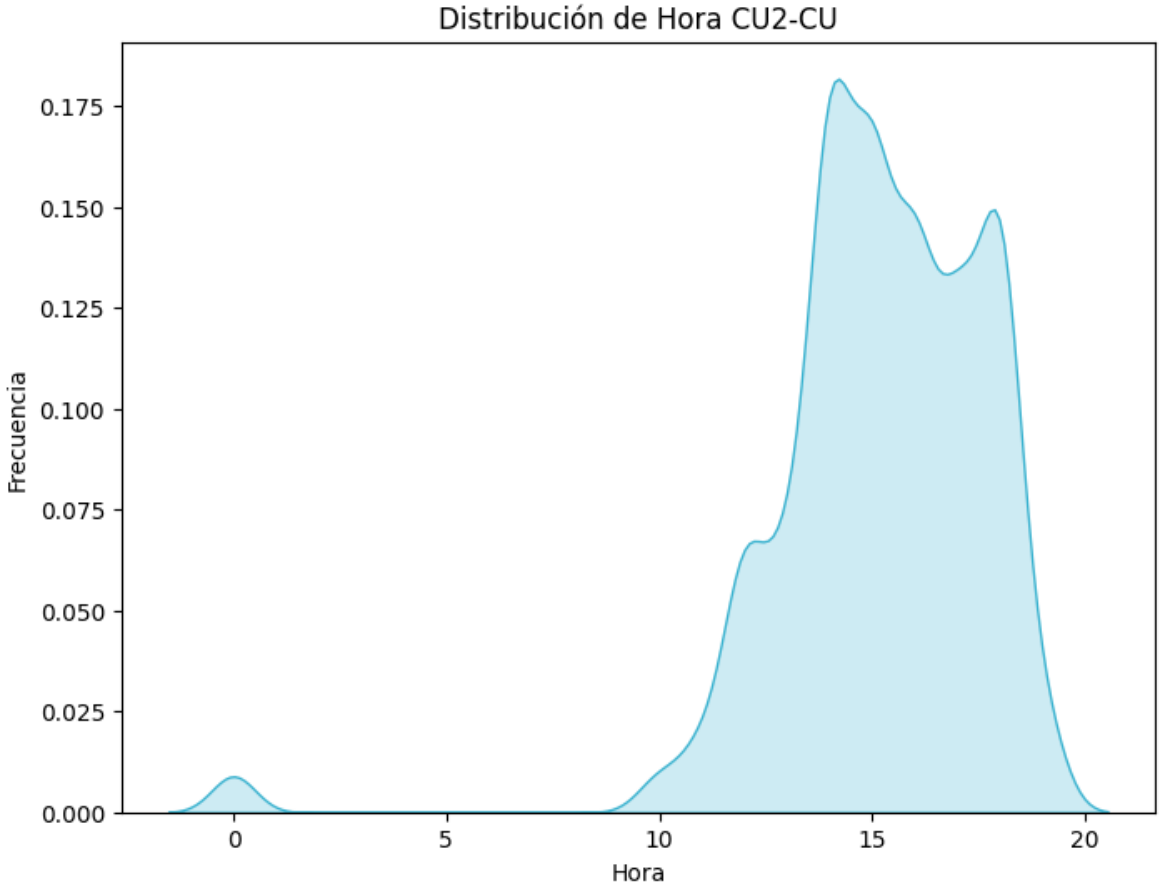
## Visualización y distribución de variables individuales

### *Numéricas*

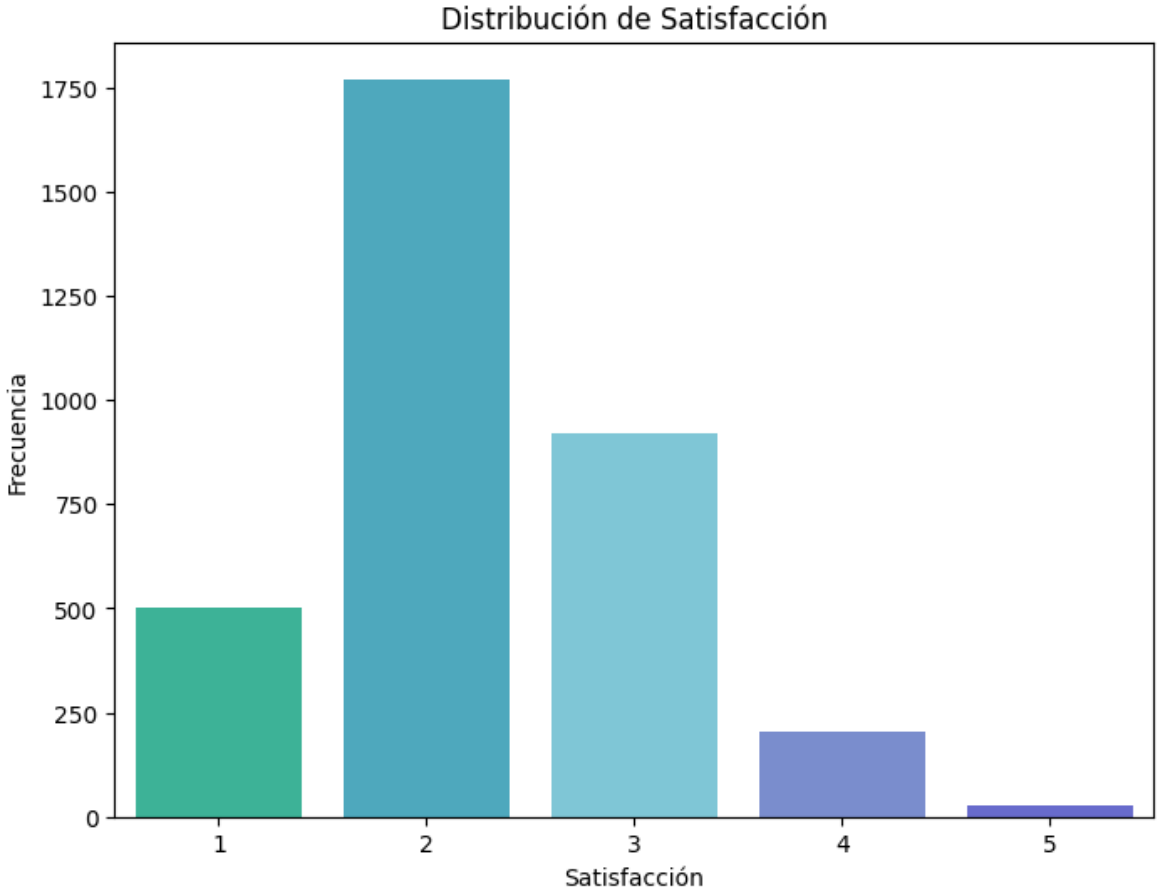
- La distribución de la variable 'Hora Cu-Cu2' muestra dos picos significativos en el horario de abordo del STU. El primer pico, de mayor intensidad, se encuentra entre las 6:00 y 8:00 a.m., lo que sugiere que una gran cantidad de estudiantes utilizan el servicio temprano en la mañana, probablemente para llegar a clases que inician a las primeras horas del día. Un segundo pico, de menor magnitud, se observa entre las 9:00 y 11:00 a.m., lo que podría indicar una segunda ola de estudiantes abordando el transporte, quizás para clases posteriores o actividades académicas en la mañana. Esta distribución refleja una posible concentración de clases en estos horarios, lo que genera picos en la demanda del transporte.



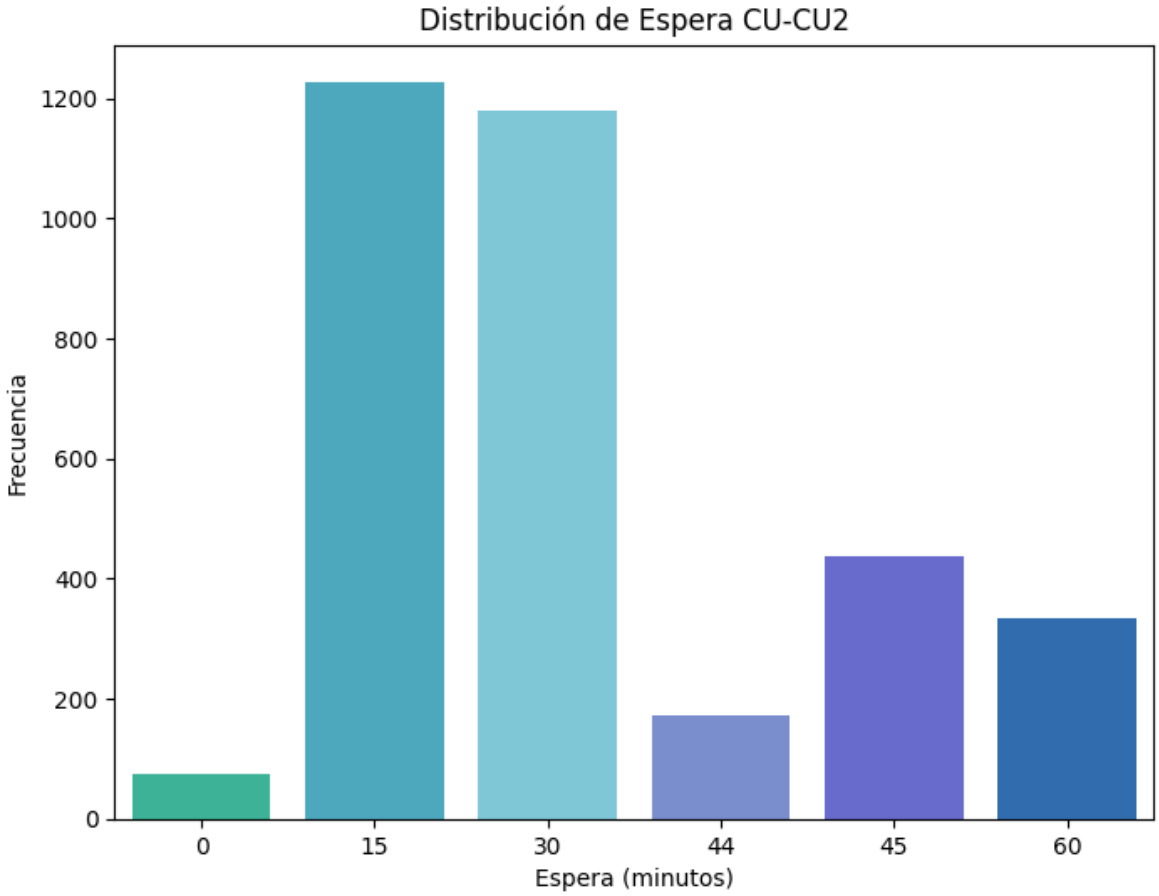
- La distribución de la variable 'Hora Cu-Cu2' muestra un patrón interesante con un pico significativo entre las 13:00 y 15:00 horas, lo que indica que muchos estudiantes utilizan el servicio durante este intervalo, posiblemente coincidiendo con el horario de salida de clases o actividades de medio día. Posteriormente, se observa una disminución gradual en la cantidad de usuarios hasta alrededor de las 17:00 horas. Sin embargo, hay un aumento notable nuevamente a las 18:00 horas. Este comportamiento sugiere dos momentos clave de alta demanda en la tarde.



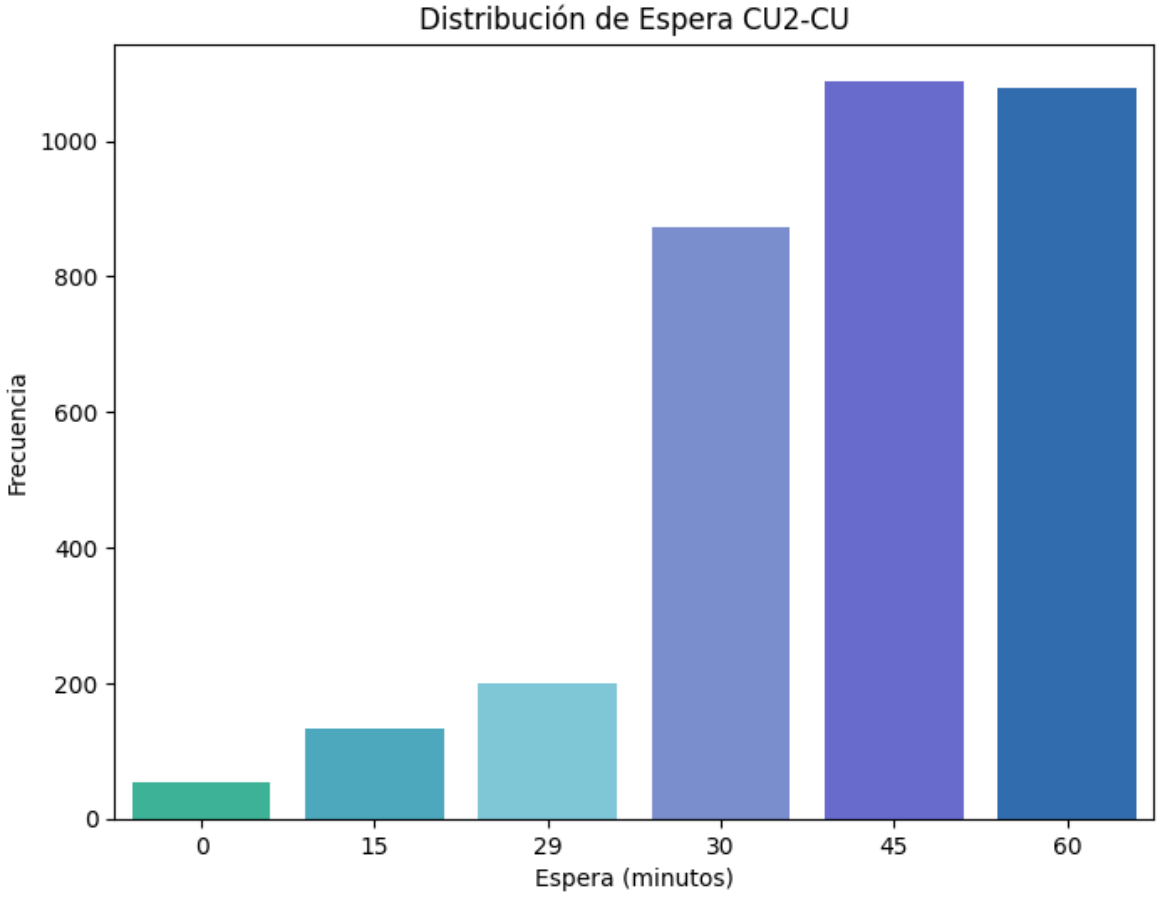
- La distribución de la variable 'Satisfacción' revela que el valor predominante es 2, seguido por el valor 3. Esto indica que una gran parte de los estudiantes tienen una percepción baja o moderadamente baja del servicio de Transporte Universitario. La predominancia del valor 2 sugiere que existen áreas significativas de insatisfacción entre los usuarios, mientras que la menor frecuencia de valores más altos como 4 y 5 refuerza esta tendencia. Este hallazgo resalta la necesidad de investigar las causas de insatisfacción para implementar mejoras en el servicio.



- La distribución de la variable 'Espera Cu-Cu2' muestra una predominancia de tiempos de espera de 15 y 30 minutos. Esto sugiere que la mayoría de los estudiantes experimentan tiempos de espera moderados, tanto al abordar el STU desde CU hacia CU2. Los demás valores de espera presentan una menor frecuencia y siguen un patrón similar, con tiempos de espera menos comunes. Esta concentración en minutos podría estar relacionada con horarios específicos de alta demanda, como las horas pico, lo que podría indicar que el servicio es relativamente puntual en la mayoría de los casos, pero aún hay una proporción considerable de estudiantes que esperan entre 15 y 30 minutos para abordar.

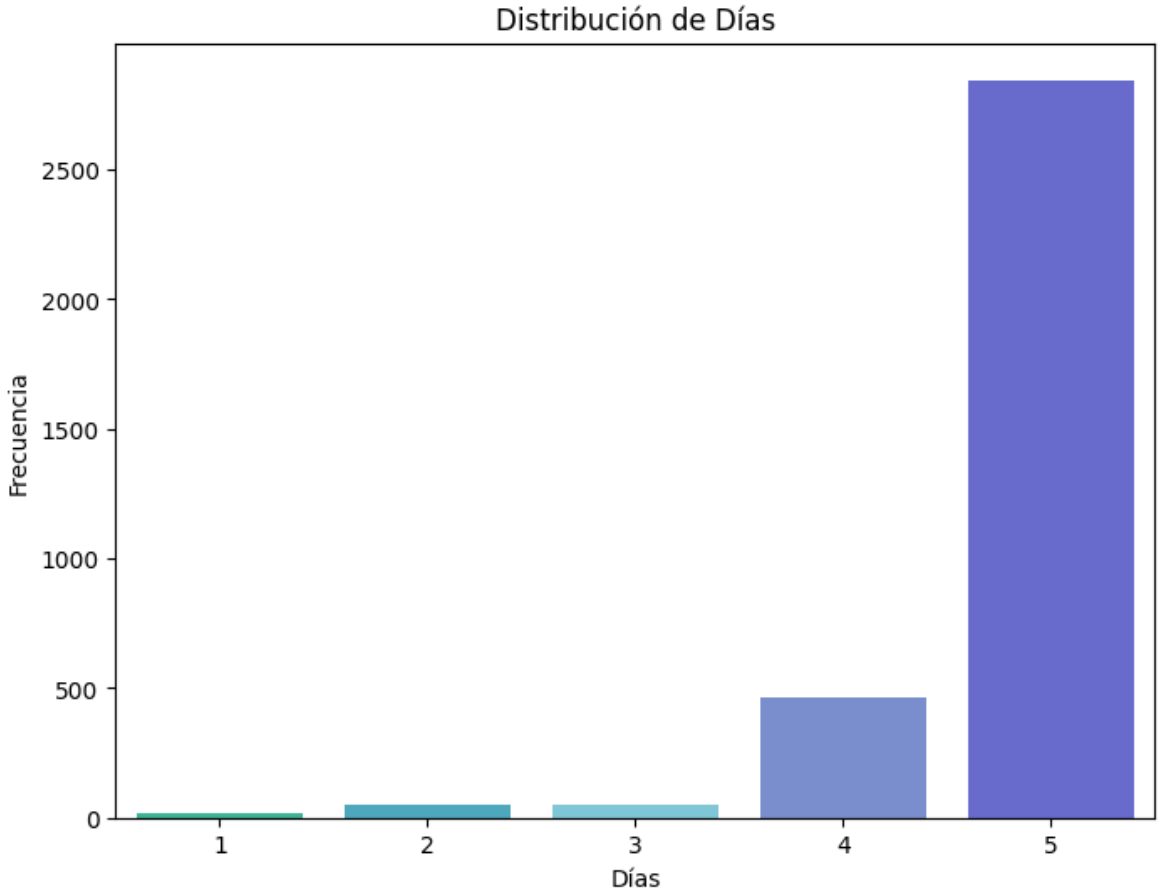


- La distribución de la variable 'Espera Cu2-Cu' muestra una mayor predominancia de los valores de espera de 45 y 60 minutos, seguidos por 30 minutos. Esto indica que un número significativo de estudiantes experimenta tiempos de espera más largos al abordar el STU de Cu2 a Cu. Esto sugiere que podría haber áreas de mejora en la en la disponibilidad de transporte.

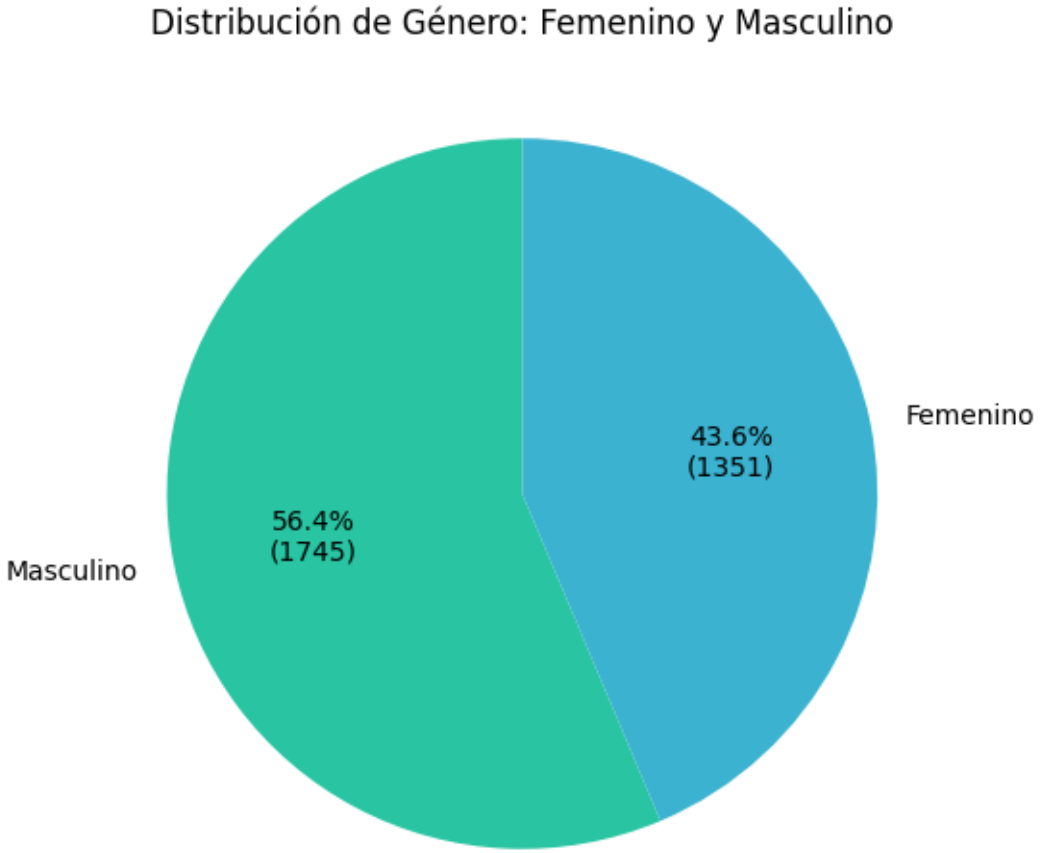


*Categóricas*

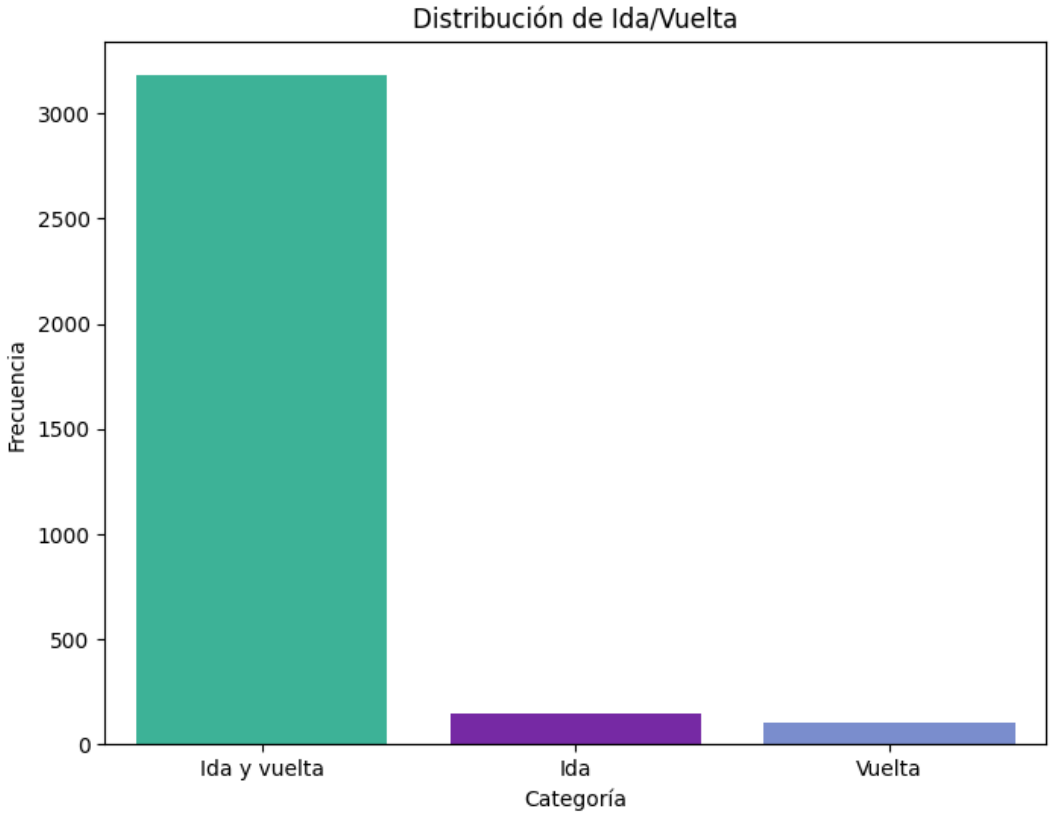
- En la distribución de la variable 'Dias' predomina de manera significativa el valor 5, lo que indica que la mayoría de los estudiantes utilizan el STU los cinco días de la semana, la diferencia con los demás valores es pronunciada, esto podría reflejar la demanda del servicio.



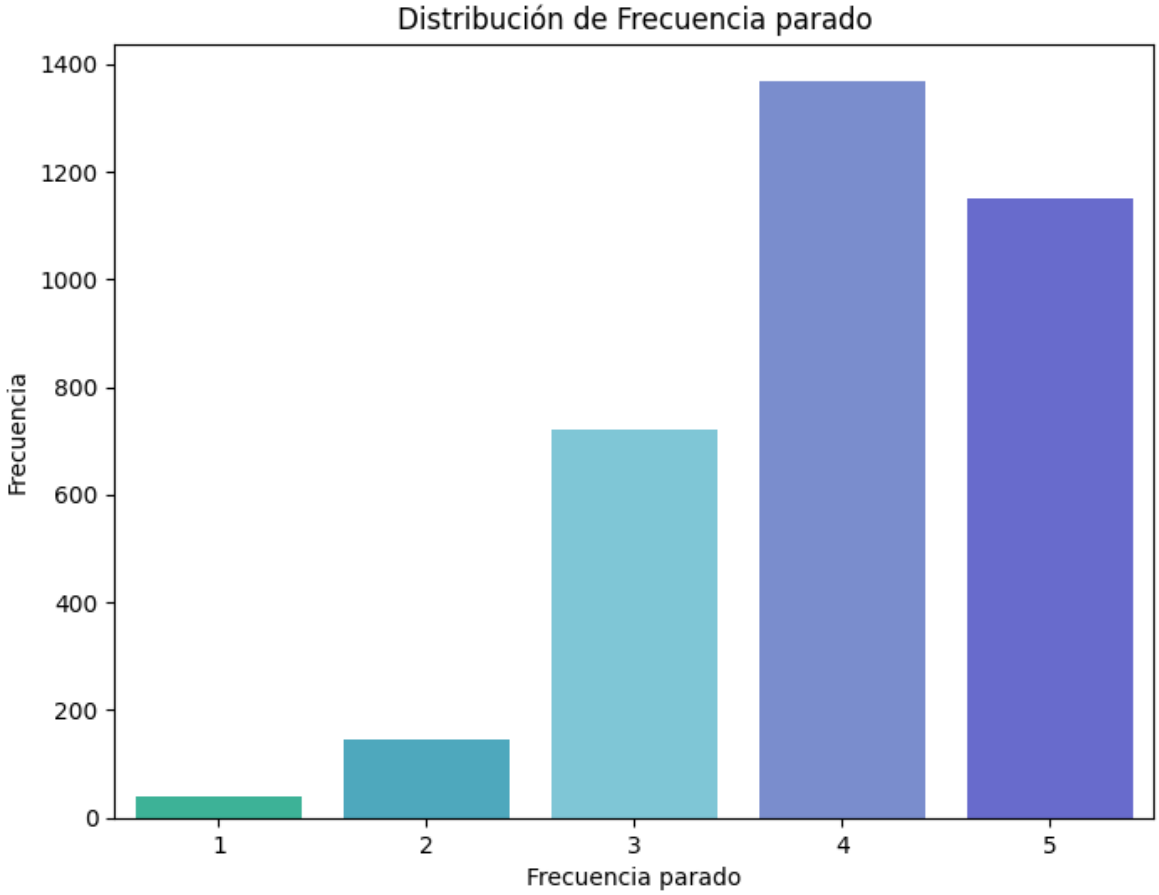
- La distribución de la variable 'Género' muestra una ligera predominancia de estudiantes masculinos, con un 56.4% de la muestra, frente al 43.6% de estudiantes femeninos. Esta distribución sugiere una representación relativamente equilibrada entre ambos géneros, aunque los estudiantes masculinos tienen una ligera ventaja numérica



- La distribución de la variable 'Ida/Vuelta' muestra una predominancia significativa de la opción Ida y Vuelta, lo que indica que la mayoría de los estudiantes utilizan el servicio de transporte tanto para ir como para regresar. Esto sugiere que una gran parte de los usuarios del STU tiene un horario regular que requiere ambos trayectos, lo que refleja una alta demanda para este tipo de servicio.



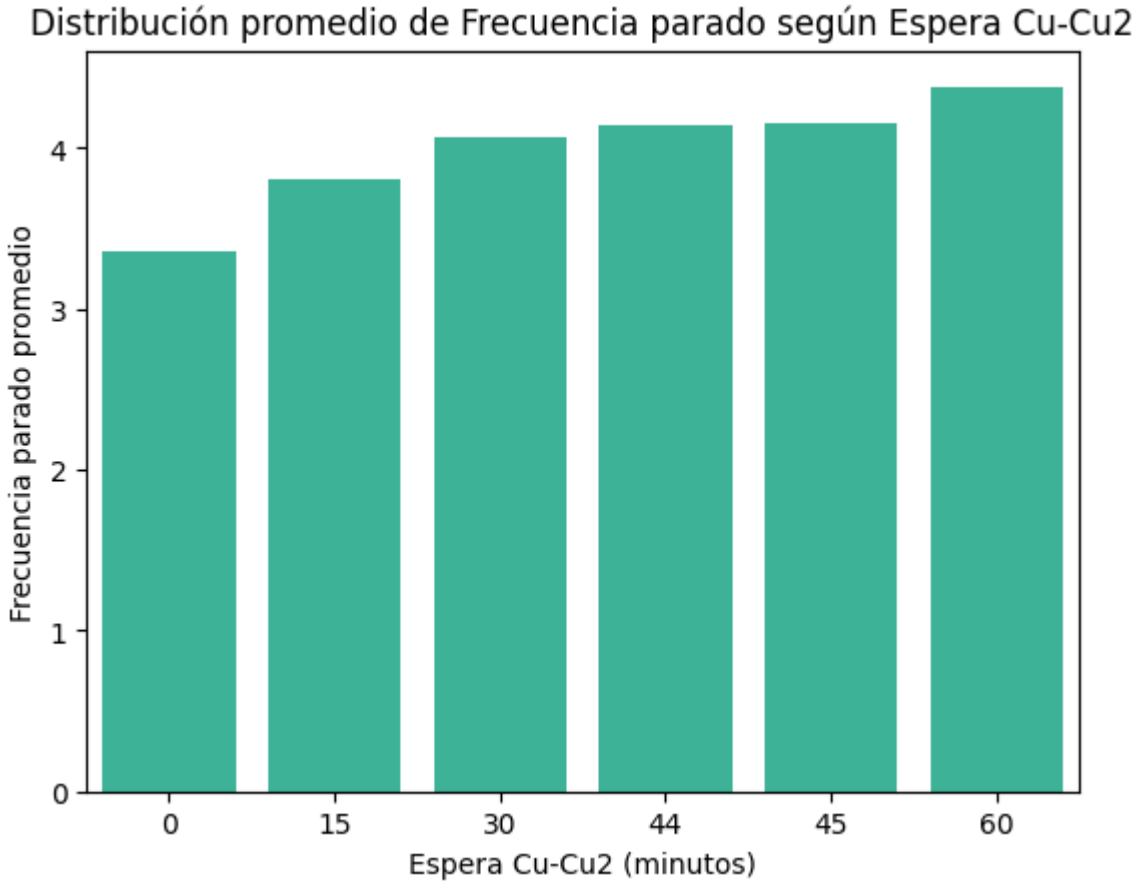
- La distribución de la variable ‘Frecuencia parado’ muestra que el valor predominante es 4, seguido de 5 lo que sugiere que una parte significativa de ellos enfrenta problemas de espacio y sobrecarga en los vehículos



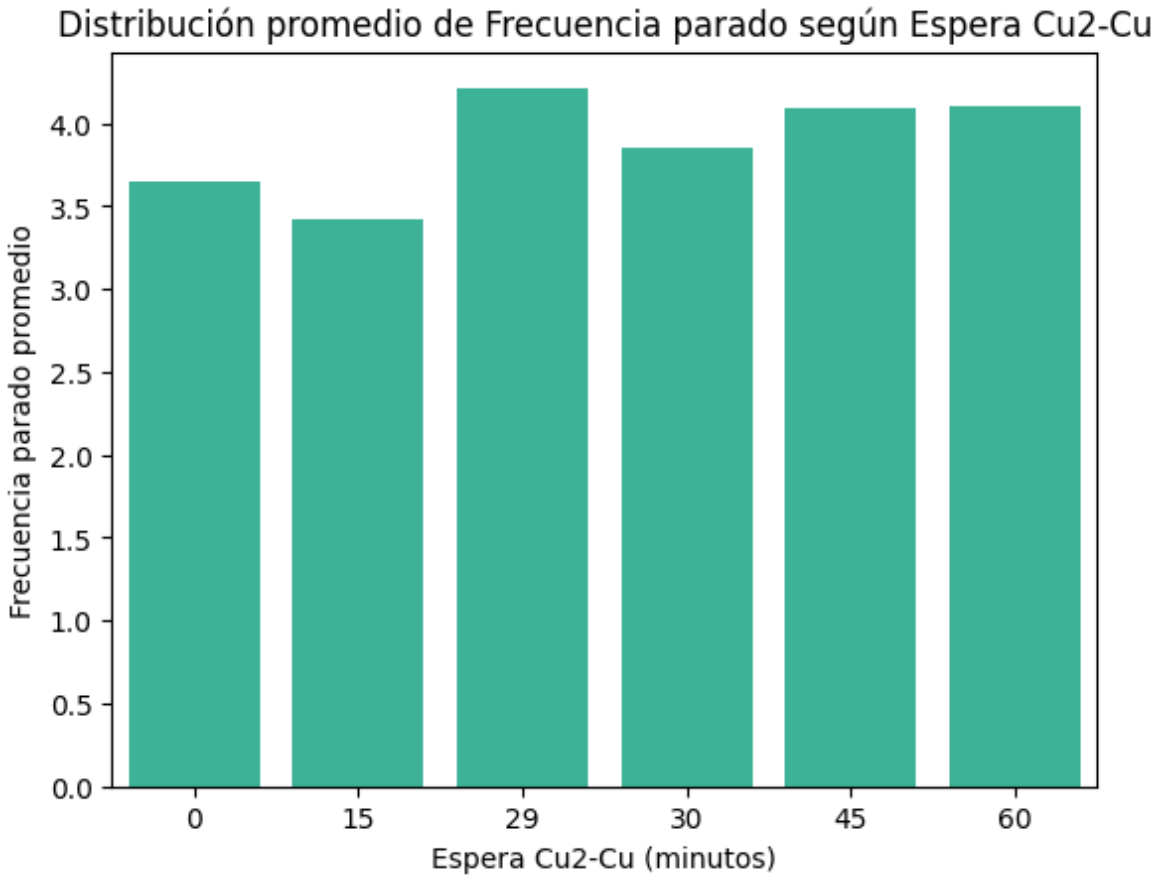
## Correlación entre variables

La matriz de correlación muestra relaciones débiles y en su mayoría no significativas entre las variables del modelo. Las correlaciones más altas se encuentran entre ‘Espera Cu-Cu2’ y ‘Frecuencia parado’ (0.22) y entre ‘Espera Cu2-Cu’ y ‘Frecuencia parado’ (0.14), lo que indica que, aunque existe una leve relación, no son factores determinantes para predecir con precisión la frecuencia con que los usuarios viajan de pie. Las otras variables, como ‘Dias’, ‘Hora Cu-Cu2’ y ‘Satisfacción’, muestran correlaciones muy bajas o nulas con ‘Frecuencia parado’, sugiriendo que su influencia en esta variable es mínima. En general, la falta de fuertes correlaciones entre las variables predictores y la variable objetivo puede explicar parte del desempeño limitado del modelo.

- Al analizar la relación entre las variables 'Espera Cu-Cu2' y 'Frecuencia parado', se observa un ligero aumento en la frecuencia de estar parado a medida que incrementa el tiempo de espera para abordar el transporte. Este patrón sugiere que los estudiantes que esperan más tiempo tienden a encontrar el transporte más lleno, lo que podría estar relacionado con la capacidad limitada o la demanda elevada durante ciertos horarios. Aunque la tendencia es moderada, podría indicar la necesidad de optimizar la capacidad del servicio en horarios de mayor espera para mejorar la experiencia de los usuarios.



- Al analizar la relación entre las variables 'Espera Cu2-Cu' y 'Frecuencia parado', no se identifica un patrón significativo. Esto indica que el tiempo de espera para abordar el transporte en el trayecto de Cu2 a Cu no parece estar directamente relacionado con la frecuencia con la que los estudiantes deben viajar de pie. Este resultado sugiere que otros factores, como la hora del día o la demanda general del servicio, podrían tener mayor influencia en la frecuencia de estar parado durante este trayecto.




## Análisis de outliers

### *Identificación de outliers*

Para detectar valores atípicos en las variables del conjunto de datos, se utilizaron visualizaciones como boxplots, que permiten identificar puntos que exceden 1.5 veces el rango Inter cuartil (IQR) por encima del tercer cuartil o por debajo del primero. Este enfoque fue aplicado a variables como los tiempos de espera (Espera Cu-Cu2 y Espera Cu2-Cu) y la frecuencia con la que los usuarios viajan parados (Frecuencia parado). Los outliers identificados se concentraron en valores extremos de espera y en categorías menos comunes de frecuencia.

### *Tratamiento de outliers*

Decidimos no eliminar los valores atípicos detectados ya que estos valores representan situaciones reales dentro del contexto del STU. Por ejemplo, tiempos de espera largos o frecuencias altas de viajar parado podrían corresponder a condiciones específicas que afectan a ciertos estudiantes y son relevantes para el análisis. Eliminar estos valores habría distorsionado la realidad del fenómeno estudiado y reducido la representatividad del modelo

## Análisis de valores faltantes

El análisis inicial identificó datos faltantes en las variables 'Espera Cu-Cu2', 'Espera Cu2-Cu', 'Frecuencia parado', y 'Satisfacción', con proporciones bajas. Decidimos rellenar los valores faltantes con el promedio de cada variable correspondiente. Este enfoque garantiza que los datos imputados reflejen la tendencia general de la población y evita introducir sesgos significativos.

- Para las variables de tiempos de espera (Espera Cu-Cu2 y Espera Cu2-Cu), el uso del promedio permitió mantener una aproximación consistente con el comportamiento observado.
- En variables categóricas transformadas a valores numéricos, como Frecuencia parado y Satisfacción, la imputación con el promedio ayudó a completar los datos sin afectar la distribución general.

Se eligió la imputación con el promedio debido a la baja proporción de datos faltantes y la necesidad de preservar el tamaño del conjunto de datos para garantizar un modelo robusto. Además, al utilizar el promedio, se mantuvo una aproximación representativa de la tendencia central sin alterar significativamente la estructura estadística de las variables.

## Observaciones y hallazgos importantes

El análisis exploratorio reveló patrones interesantes en las variables. En 'Dias', se observó que la mayoría de los usuarios utilizan el servicio durante los 5 días de la semana, lo que refleja una alta dependencia del transporte universitario. En cuanto a los tiempos de espera (Espera 'Cu-Cu2' y 'Espera Cu2-Cu'), los valores más frecuentes son 15, 30, 45 y 60 minutos, lo que sugiere la existencia de horarios definidos o picos de demanda. La 'Frecuencia parado' mostró una saturación significativa del servicio, con las categorías 4 y 5 como las más frecuentes, indicando que muchos usuarios deben viajar de pie regularmente.

En cuanto a relaciones entre variables, la correlación más destacada fue la leve relación positiva entre 'Espera Cu-Cu2' y 'Frecuencia parado' (0.22), lo que indica que tiempos de espera más largos están asociados con una mayor probabilidad de viajar parado. Sin embargo, la mayoría de las variables, como 'Satisfaccion' y 'Dias', mostraron correlaciones débiles o nulas con la variable objetivo, lo que sugiere que su influencia directa en la predicción de la frecuencia parado es limitada.

En el tratamiento de valores atípicos, se identificaron outliers en los tiempos de espera y en la frecuencia parado. Estos valores no fueron eliminados porque representan situaciones reales dentro del contexto del transporte, como tiempos de espera prolongados o una alta saturación del servicio. Mantener estos datos garantiza un análisis más representativo y permite que el modelo capture patrones críticos relacionados con condiciones extremas.

Por último, los datos faltantes en variables como 'Espera Cu-Cu2', 'Espera Cu2-Cu', 'Frecuencia parado', y 'Satisfacción' se imputaron utilizando el promedio de cada variable. Esta estrategia permitió preservar el tamaño del conjunto de datos y mantener consistencia en las distribuciones. Sin embargo, dado que la mayoría de las variables tienen una relación débil con la variable objetivo, se anticipa que el modelo puede enfrentar desafíos en la predicción precisa, especialmente en clases menos frecuentes, como se observó en los resultados del modelo de árbol de decisión.

# Modelo de machine learning

## Descripción del modelo

El modelo elegido es un árbol de decisiones, implementado para predecir la frecuencia con la que los usuarios viajan de pie en el STU clasificado en 5 categorías de la variable 'Frecuencia parado

## Justificación del modelo

El árbol de decisiones fue seleccionado debido a que genera reglas claras que permiten entender como las variables predictoras influyen las decisiones, trabaja bien con variables categóricas y numéricas, es eficiente para el conjunto de datos y es fácil hacer ajustes en los parámetros como profundidad o tamaño de nodos para optimizar el rendimiento

## Implementación y entrenamiento

1. División de los datos:
- Se dividimos el conjunto en 70% para entrenamiento y 30% para prueba con train_test_split
- Las variables predictoras incluidas fueron: 'Dias', 'Ida/Vuelta', 'Hora Cu-Cu2', 'Hora Cu2-Cu', 'Satisfaccion', 'Espera Cu-Cu2', 'Espera Cu2-Cu'
- La variable objetivo fue 'Frecuencia parado'
2. Entrenamiento de modelo
- El modelo fue entrenado usando DecisionTreeClassifier con un parámetro random_state=42 para garantizar la reproducibilidad
- Durante el entrenamiento, el modelo ajustó reglas basadas en las relaciones entre las variables predictoras y la variable objetivo

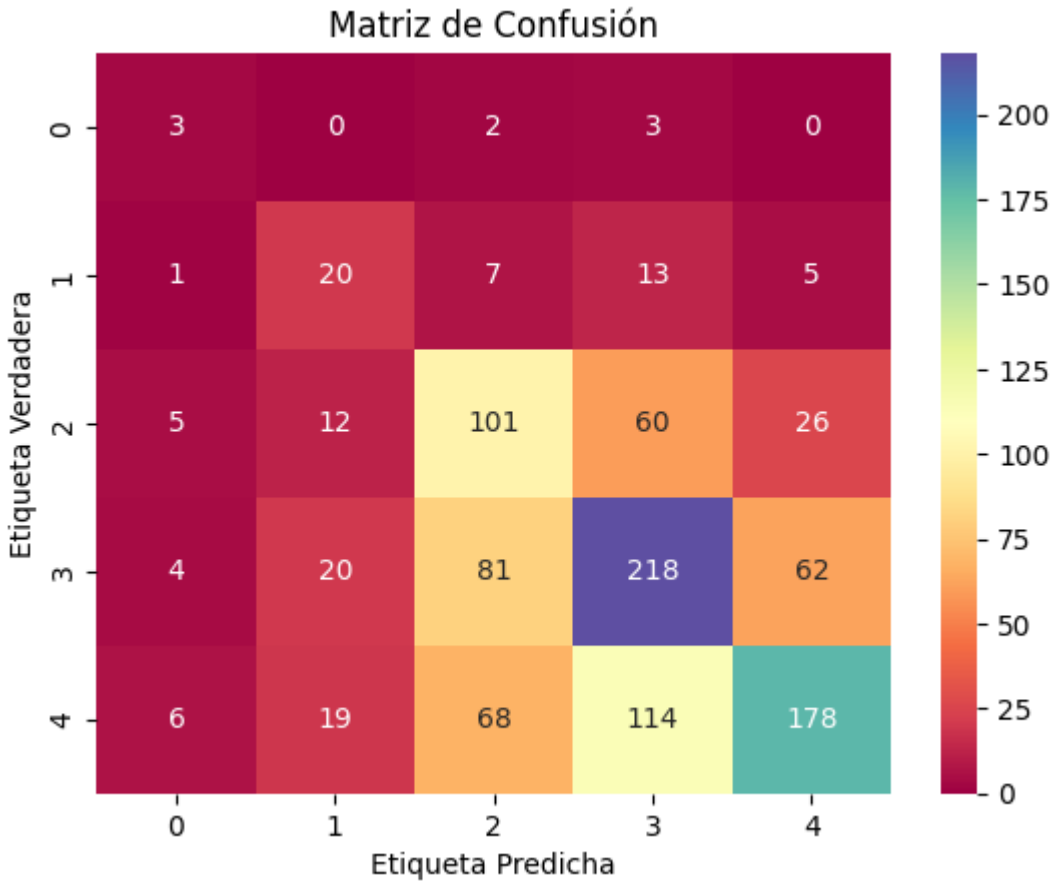
3. Evaluación del modelo

- Métrica principal: precisión (Accuracy).
- Matriz de confusión: Proporcionó información detallada sobre el desempeño para cada categoría de la variable objetivo.
- Tasas de error específicas

## Resultados

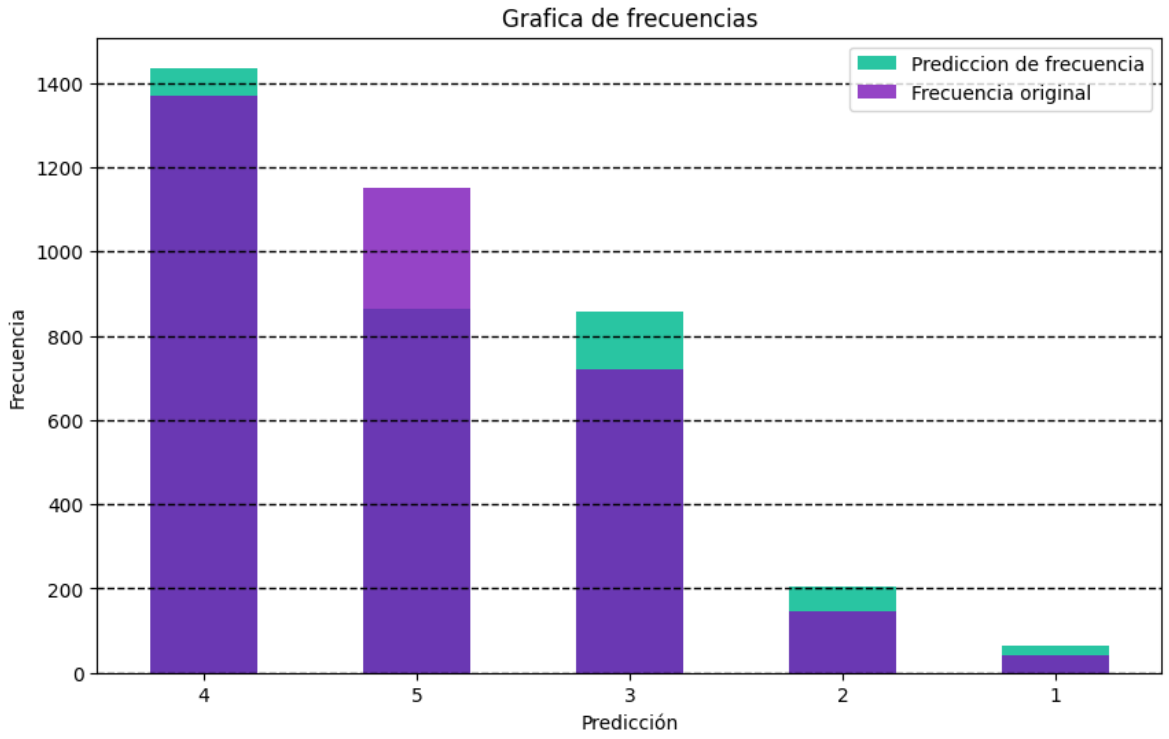
- Precisión del modelo: 51% lo que indica poco mas de la mitad de las predicciones coinciden con las etiquetas reales

- Matriz de confusión: la diagonal principal refleja las predicciones correctas, las categorías 4 y 5 tienen una mayor cantidad de aciertos lo que puede indicar que el modelo las identifica con mayor facilidad, los errores en categorías 1 y 2 pueden deberse a un desequilibrio de los datos o a limites difusos entre las clases.




- Falsos positivos: 0.00, prácticamente inexistentes para algunas clases
- Falsos negativos: 0.05, moderados lo que sugiere que el modelo pierde algunas etiquetas verdaderas

- La gráfica de barras comparativa entre las predicciones del modelo y los valores reales muestra que la clase 1 es la mejor representada, con una predicción similar a la distribución real, seguida de las clases 4 y 2, que también tienen un buen desempeño. Sin embargo, las clases 3 y 5 presentan mayores discrepancias, siendo la clase 5 la que tiene la mayor diferencia entre las predicciones y los valores reales. Esto sugiere que el modelo tiene dificultades para clasificar correctamente estas clases, en especial la clase 5, que podría beneficiarse de un ajuste de parámetros o de técnicas para equilibrar las clases y mejorar la precisión.



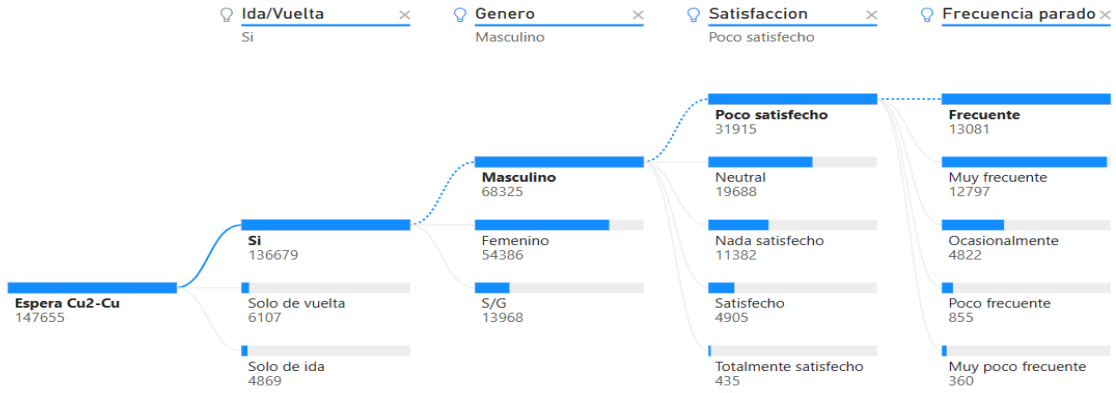

## Interpretación de los resultados

El modelo presenta un desempeño moderado con una precisión de 0.51, la precisión es aceptable, pero hay algunas áreas en las que se podría mejorar:

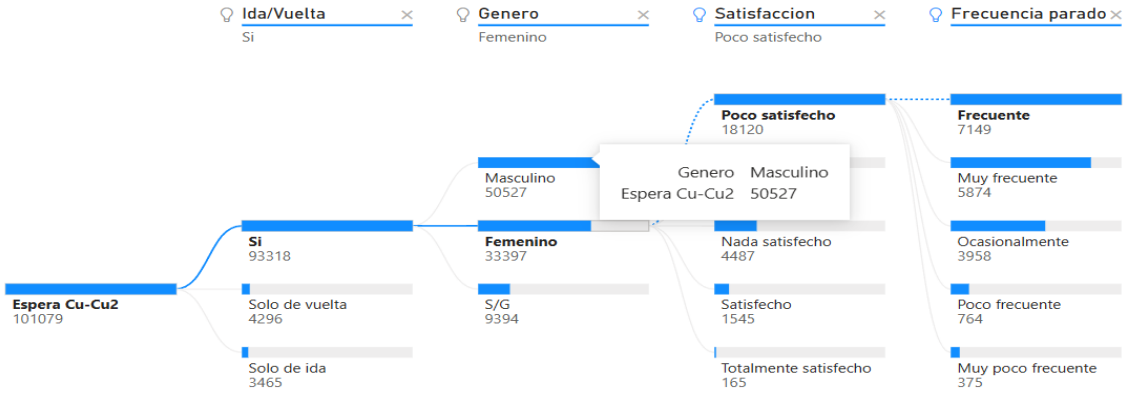
- Ajustar los parámetros como profundidad o número mínimo de muestras por hojas para evitar el sobreajuste
- Probar con modelos complementarios como random forest o gradient boosting para comparar los resultados
- Evaluar que variables tienen mayor peso en la toma de decisiones y considerar reducir aquellas menos relevantes

# Dashboard

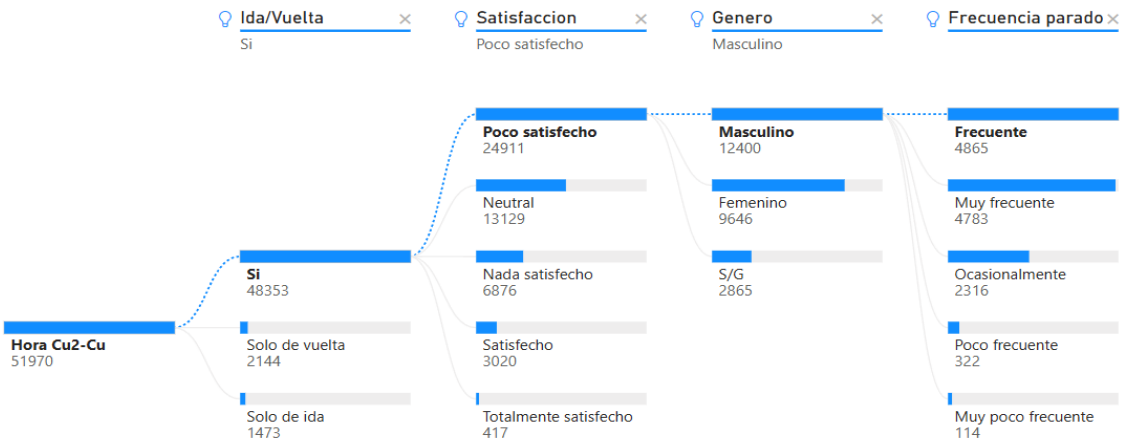
El objetivo del dashboard en este proyecto es tener una visualización más avanzada de los datos e información para contar una historia (storytelling) con los datos analizados por medio de Dashboards en PowerBI , comunicar datos de manera clara y efectiva.



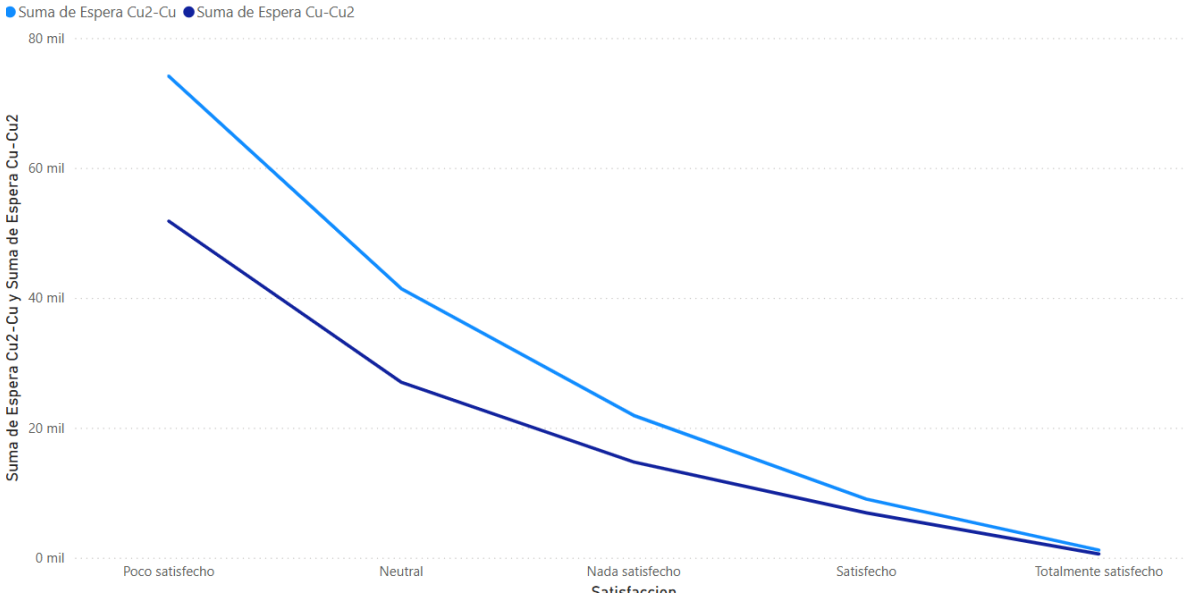

En este primer dashboard  nos muestra que en la categoría de Espera Cu2-Cu, que los estudiantes masculinos que abordan el STU de ida y de vuelta, tienen una poca satisfacción ya que es frecuente que vallan parados.




En este Dashboard, nos muestra que en la Espera Cu-Cu2, las estudiantes femeninas están poca satisfechas ya que es muy frecuente que vallan paradas.




En estos dashboards, como podemos visualizar cada categoría relevante (inicio) tiene ramificaciones dependiendo a lo que se quiera encontrar, en este caso nuestro objetivo es mostrar las frecuencias en las que los estudiantes van parado




En este lado, tenemos la relación en entre las esperas de Cu-Cu2 y Cu2-Cu, conforme a la satisfacción de los estudiantes (poco satisfecho, neutral, nada satisfecho, satisfecho, totalmente satisfecho) en donde predomina ”Poco satisfecho”, con este dashboard podemos decir que los estudiantes están inconformes con el Sistema de transporte universitario de la BUAP en la Ruta de Cu-Cu2 y viceversa.

Estas graficas (Dashboards), nos ayudan a entender lo que está pasando en la logística del STU, por ende, podemos dar posibles soluciones a la logística del Transporte universitario.

# Conclusiones y futuras líneas de trabajo

## Conclusiones

El análisis exploratorio y el modelo de árbol de decisión permitieron obtener información valiosa sobre el comportamiento de los usuarios del transporte universitario. Se identificó que los usuarios tienden a utilizar el servicio durante toda la semana y que los tiempos de espera más comunes están distribuidos en intervalos definidos (15, 30, 45 y 60 minutos). Asimismo, se observó una alta frecuencia de viajes realizados de pie, lo que sugiere un problema de saturación en el servicio. Aunque se detectaron correlaciones débiles entre las variables predictoras y la frecuencia parado, la leve relación entre tiempos de espera y saturación resalta la importancia de optimizar los tiempos de espera para mejorar la experiencia del usuario.

El modelo de árbol de decisión alcanzó un desempeño limitado con una precisión del 51%, mostrando una mayor efectividad al predecir categorías más frecuentes, como los niveles 1 y 4 de frecuencia parado, mientras que tuvo dificultades en las clases menos representadas. Esto refleja que la variabilidad inherente y las correlaciones débiles en los datos afectan la capacidad predictiva del modelo, lo que sugiere oportunidades para ajustar o mejorar la selección de características.

## Posibles mejoras

Para mejorar el modelo y los datos, sería útil recopilar más información sobre factores que influyen directamente en la experiencia de los usuarios, como la capacidad del transporte, número de usuarios por horario y la distancia recorrida. Además, aplicar técnicas avanzadas de ingeniería de características podría ayudar a capturar relaciones no evidentes entre las variables. En términos de modelado, explorar algoritmos más complejos, como bosques aleatorios o redes neuronales, podría mejorar la precisión, especialmente en categorías menos frecuentes.

## Futuras líneas de tiempo

Un área interesante para investigaciones futuras sería analizar los horarios específicos de mayor saturación y los factores externos que afectan los tiempos de espera, como el clima o la disponibilidad de unidades. También sería valioso considerar un enfoque de predicción multinivel que incorpore datos categóricos (como género o satisfacción) con datos continuos

(como tiempos de espera), para obtener un modelo más robusto. Finalmente, realizar análisis de sensibilidad para evaluar el impacto de las imputaciones o valores atípicos en el rendimiento del modelo podría proporcionar mayor claridad sobre su efecto en las predicciones.

Como posible solución al problema de logistica del transporte universitario en el caso de regreso (Cu2-Cu), es que la problemática de que aquellos transportes que solo dejan estudiantes sin recoger a los estudiantes que van para CU, se le puede dar una aplicación y es que en un intervalo de 30 minutos salga un camión de Cu a Cu2 ya sea con estudiantes que van para Cu2 o sin estudiantes. Y que en 15 minutos después salga uno que ya haya recogido a los estudiantes de Cu2 con dirección a Cu de esta forma cada 15 minutos hay un transporte disponible a los estudiantes en ambos casos sin tener que hacer esperar mucho tiempo y con esto reducimos las emisiones innecesarias de carbono y los recursos extras que se usan.

# Video

https://youtu.be/yo7WsVDqj98?si=3jOqYVneW1q07F8b

# Referencias


- Aggarwal, C. C. (2015). Outlier Analysis. Springer.
- Few, S. (2006). Information Dashboard Design: The Effective Visual Communication of Data. Analytics Press
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (2000). Understanding Robust and Exploratory Data Analysis. Wiley.
- Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data.
- Scikit-learn. (n.d.). Decision Trees. Retrieved from https://scikit-learn.org/stable/modules/tree.html
- Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.
- Riveroll Martínez, D. P., & Marín Nieva, J. S. (2024). Base de datos de uso del transporte universitario. Datos no publicados.