MAESTRÍA EN CIENCIA DE DATOS
FUNDAMENTOS CIENCIA DE DATOS
INFORME FINAL

Título del Proyecto:

"Modelo de Predicción del Consumo de Diésel en Unidades de Transporte Público**".**

Objetivo del Proyecto:

- **General:**

  Desarrollar un modelo predictivo que permita estimar con precisión el consumo de diésel en unidades de transporte público, con el fin de optimizar la planificación de compras y asegurar la continuidad del servicio.

- **Específicos:**

  1. Analizar y comprender el comportamiento histórico del consumo de diésel.
  2. Limpiar y transformar los datos recopilados de manera manual para asegurar su calidad.
  3. Generar nuevas variables que mejoren el rendimiento de los modelos predictivos.
  4. Entrenar y evaluar modelos de regresión para estimar el consumo futuro.
  5. Identificar patrones de consumo anómalos que puedan indicar fallos mecánicos en las unidades.

- **Resumen Ejecutivo:**

  El presente proyecto tiene como finalidad desarrollar un modelo de predicción del consumo de diésel en unidades de transporte público. Actualmente, la empresa carece de una estimación precisa del consumo, lo que conlleva a la adquisición ineficiente de combustible, afectando directamente la asignación presupuestaria y la operatividad de las unidades. Para resolver este problema, se llevó a cabo un proceso completo de análisis y tratamiento de datos, seguido de la construcción de variables significativas para mejorar el rendimiento de los modelos. El resultado permitirá mejorar la toma de decisiones estratégicas en la compra de combustible, así como apoyar en el diagnóstico preventivo de fallas mecánicas mediante el análisis del consumo.

- **Valor e Impacto:**

  El modelo contribuirá a una gestión más eficiente del presupuesto destinado al combustible, evitando tanto el exceso como la escasez de diésel. Además, permitirá anticipar fallas mecánicas a partir de patrones anómalos en el consumo, lo cual también optimiza la planificación de mantenimientos. Su implementación impactará directamente en la continuidad del servicio y la sostenibilidad financiera de la empresa.

- **Pregunta Central:**

  ¿Es posible predecir con precisión el consumo de diésel de las unidades de transporte público utilizando datos históricos y variables derivadas de su operación?

- **Hipótesis Inicial:**

  Si se dispone de datos históricos confiables sobre el consumo de diésel y variables operacionales adecuadas, entonces es posible construir un modelo de regresión que prediga con precisión el consumo futuro.

- **Alineación:**

  Este proyecto se alinea con los principios de ciencia de datos aplicada a la optimización de procesos logísticos, en concordancia con los objetivos estratégicos de eficiencia operativa de la empresa de transporte público.

Antecedentes:

- **Situación Actual:**

  Actualmente, la institución encargada del transporte público no cuenta con una herramienta que le permita realizar una predicción precisa del consumo de diésel. Esta falta de visibilidad ha generado inconsistencias en la planificación y adquisición del combustible, comprometiendo tanto el funcionamiento continuo de las unidades como la asignación adecuada del presupuesto institucional.

- **Dolor del Negocio:**

  La ausencia de pronósticos confiables conlleva a dos escenarios recurrentes: la compra excesiva de diésel, lo cual inmoviliza recursos financieros que podrían destinarse a otras áreas estratégicas, o la compra insuficiente, que pone en riesgo la continuidad del servicio. A esto se suma la dificultad para anticipar el desgaste o posibles fallas mecánicas de las unidades, pues no existe una correlación clara entre el consumo anómalo y el estado mecánico de los vehículos.

- **Diagnóstico:**

  Mediante el análisis de la situación actual, se identificó que la obtención de los datos se realiza manualmente, lo que introduce errores de transcripción y registro. Además, no existen mecanismos automáticos de validación o control de calidad de los datos. Esta situación impide generar proyecciones confiables, lo que a su vez afecta la eficiencia operativa.

- **Propósito:**

  El presente proyecto tiene como propósito desarrollar un modelo predictivo que, basado en el comportamiento histórico del consumo de combustible y en variables adicionales como el kilometraje, frecuencia de abastecimiento y características de la unidad, permita pronosticar de manera precisa el consumo de diésel en el corto plazo.

- **Requisitos:**

  1. Un conjunto de datos históricos limpios y estructurados.
  2. Algoritmos de aprendizaje supervisado que permitan generar predicciones precisas.
  3. Variables operativas relevantes, tales como kilometraje, tipo de unidad, frecuencia de carga y fecha de operación.

- **Scope:**

  Este proyecto se enfoca exclusivamente en la predicción del consumo de diésel. Sin embargo, abre la puerta a futuras integraciones con sistemas de mantenimiento predictivo y abastecimiento inteligente. El modelo desarrollado puede servir como base para anticipar averías, planificar mantenimientos de surtidores y mejorar la eficiencia energética general del sistema de transporte.

- **Limitaciones:**

  Entre las principales limitaciones se encuentran:

  1. La recolección manual de datos, que introduce errores y dificulta la estandarización.
  2. La imposibilidad de obtener datos consistentes para todas las unidades, especialmente aquellas que se encuentran averiadas.
  3. La falta de sensores o telemetría que permita la recolección automatizada del kilometraje o estado operativo.

Entendimiento de los Datos:

- **Fuente de Datos:**

  Los datos provienen de registros manuales de abastecimiento de diésel realizados por el personal de las estaciones de carga de combustible. Estos incluyen información de consumo por unidad, kilometraje, tipo de unidad, fecha de abastecimiento, y otros atributos asociados a la operación de los buses.

- **Descripción y Calidad de los Datos:**

  Durante el análisis exploratorio se identificaron múltiples inconsistencias, entre ellas:

  1. Datos faltantes en las columnas de kilometraje, tipo de unidad y hora de abastecimiento.
  2. Ingresos duplicados o erróneos con valores atípicos, como kilometrajes iguales a cero.
  3. Errores en el formato de fechas y tiempos.

  A pesar de estas limitaciones, el conjunto de datos contenía suficiente información para generar una base sólida de análisis, previa una limpieza adecuada

Preparación de los Datos (Data Preparation – CRISP-DM):

- **Procesos de Limpieza:**

  Se aplicaron varios pasos para limpiar el conjunto de datos:

  1. Eliminación de registros duplicados o con datos críticos nulos.
  2. Conversión de tipos de datos, como fechas y horas a formato “datetime”.
  3. Corrección de formatos inconsistentes en el kilometraje y volumen de carga.
  4. Relleno y ajuste de valores faltantes en columnas esenciales, como el tipo de unidad.

- **Featuring Egineering:**

  Se crearon variables adicionales a partir de los datos existentes, tales como:

  1. Día de la semana, hora y mes de abastecimiento, para capturar patrones temporales.
  2. Consumo promedio por unidad y variación de kilometraje, que aportan información contextual relevante.
  3. Categorías por tipo de unidad y estación, transformadas mediante técnicas de codificación categórica (Label Encoding y One-Hot Encoding).

  Estas nuevas características permitieron enriquecer el conjunto de datos y mejorar la capacidad de predicción de los modelos.

- **Ajuste de Formatos:**

  Todos los formatos de fecha, hora, tipo de unidad y categorías nominales fueron estandarizados para facilitar el modelado. Las variables categóricas fueron codificadas, y las numéricas fueron verificadas para evitar valores extremos que afecten la estabilidad del modelo.

Modelado (Modeling – CRISP-DM):

- **Selección de Modelos:**

  Para la fase de modelado se seleccionaron algoritmos de regresión supervisada que permiten predecir variables continuas. Se eligieron cuatro modelos representativos:

  1. Regresión Lineal (Linear Regression): para establecer una línea base simple.
  2. Árbol de Decisión (Decision Tree Regressor): por su capacidad de capturar relaciones no lineales.
  3. Bosques Aleatorios (Random Forest Regressor): una técnica de ensamblado que mejora la precisión general.
  4. Gradiente Potenciado (Gradient Boosting Regressor): por su eficiencia en predicciones más precisas y su capacidad de minimizar errores residuales.

- **Entrenamiento y Validación:**

  Se seleccionaron tres variables como predictores: KILOMETRAJE, GALONES_POR_KM y ANIO (se tuvieron inconvenientes al tratar la columna con la letra “Ñ”), y la variable objetivo fue GALONES, correspondiente al volumen de diésel cargado.

  Antes del entrenamiento, se aplicó escalado estándar sobre las variables predictoras. Este paso fue fundamental, especialmente para algoritmos como la regresión lineal y el gradiente potenciado, los cuales son sensibles a la escala de los datos. El escalado garantiza que todas las características contribuyan de manera proporcional al modelo y evita que variables con rangos mayores dominen el proceso de aprendizaje.

  Los datos fueron divididos en conjuntos de entrenamiento (75%) y prueba (25%) utilizando validación aleatoria con “random_state=42”

- **Métricas de Evaluación:**

  Se utilizaron tres métricas para evaluar el rendimiento de cada modelo:

  MAE (Error Absoluto Medio): indica cuánto se desvía, en promedio, la predicción del valor real.
  RMSE (Raíz del Error Cuadrático Medio): penaliza errores grandes y proporciona una visión más estricta del rendimiento.
  R^2 (Coeficiente de Determinación): mide la proporción de la variabilidad explicada por el modelo.

  Donde se obtuvieron los siguientes resultados:

Modelo	MAE	RMSE	R^2
RandomForestRegressor	50628	8.83×10^6	0.9636
DecisionTreeRegressor	60404	1.27×10^7	0.9245
GradientBoostingRegressor	110749	1.28×10^7	0.9242
LinearRegression	545643	3.89×10^7	0.2923

Evaluación e Interpretación de Resultados (Evaluación – CRISP-DM):

- **Errores Comunes:**

  La regresión lineal presentó una baja capacidad predictiva, evidenciada por su bajo valor de R^2, lo cual indica que no logra captar la complejidad de los datos. Esto puede deberse a relaciones no lineales entre las variables, que no pueden ser modeladas adecuadamente con esta técnica.

  Los modelos basados en árboles demostraron un mejor ajuste, siendo el Random Forest Regressor el que obtuvo el mejor rendimiento general en todas las métricas.

- **Mejoras:**

  Se identificaron posibles mejoras para futuras iteraciones del modelo:

  1. Incorporar más variables predictoras como tipo de unidad, estación de carga, hora del día, etc.
  2. Utilizar técnicas de validación cruzada más robustas.
  3. Implementar un sistema de captura automatizada de datos para mejorar la calidad del dataset.

Conclusiones, Próximos Pasos y Recomendaciones:

El desarrollo del modelo predictivo de consumo de diésel permitió demostrar que es posible obtener proyecciones precisas a corto plazo, lo que representa una ventaja significativa para la planificación operativa del abastecimiento de combustible. El modelo basado en Random Forest obtuvo el mejor desempeño, sugiriendo que es la mejor opción para su implementación inicial.

- **Propuestas de Investigación Adicional:**

    1. Integración del modelo con plataformas IoT que automaticen la recolección de datos desde los surtidores y unidades.
    2. Análisis predictivo para mantenimiento preventivo, utilizando patrones de consumo anómalos como indicadores de posibles fallas mecánicas.
    3. Extensión del modelo a otras fuentes de energía, como electricidad en buses híbridos o eléctricos.