APLICANDO APRENDIZAJE AUTOMÁTICO SUPERVISADO PARA UN DESAFÍO REAL: PREDICCIÓN DE LA HUELLA DE CARBONO

SELECCIÓN DEL PROBLEMA

- El problema debe ser de naturaleza supervisada.
    - El dataset contiene registros históricos de las huellas de carbono asociadas a diversas actividades, incluyendo, pero no limitado a, consumos energéticos, tipos de transporte, y producción industrial. Cada registro en el dataset tiene una etiqueta correspondiente que indica la cantidad de carbono emitido, lo que nos permite aplicar técnicas de aprendizaje supervisado. Utilizando estos datos, podemos entrenar un modelo predictivo para estimar la huella de carbono en base a diferentes características de entrada. Esta tarea se alinea perfectamente con el enfoque de aprendizaje supervisado, donde el objetivo es predecir una variable objetivo (en este caso, la huella de carbono) a partir de un conjunto de variables de entrada.
- El problema debe ser relevante para el mundo real.
    - Actualmente ocupo el cargo de director de Tecnología en KLAXEN SAS, una empresa dedicada a la producción de productos de limpieza y desinfección que priorizan la sostenibilidad ambiental. Nuestra misión corporativa se centra en el desarrollo de soluciones biodegradables que no requieran enjuague, sean fáciles de transportar y contribuyan positivamente a la reducción de la huella de carbono típicamente asociada con nuestro sector. Adjunto encontrarás nuestra presentación corporativa llamada,"*Brochure Klaxen Línea Klaxinn Tabs Detclork*" , la cual profundiza en nuestra motivación para elegir este dataset específico. Creo firmemente que nuestra evaluación no debe limitarse al proceso de fabricación; es esencial que nuestras prácticas diarias reflejen los valores que promovemos y vendemos.
- El problema debe ser factible de resolver con aprendizaje automático
    - Es factible, los modelos de aprendizaje automático, como la regresión, pueden manejar múltiples variables de entrada y descubrir patrones no lineales y relaciones complejas entre ellas. Esto los hace convenientes para predecir la huella de carbono, donde la cantidad de emisiones puede depender de una amplia gama de factores, incluyendo el tipo de actividad, la eficiencia energética, las fuentes de energía utilizadas, y más. Además, la disponibilidad de datos históricos etiquetados en el dataset proporciona una base sólida para entrenar y validar modelos predictivos, lo que demuestra la factibilidad de resolver este problema mediante técnicas de aprendizaje automático.

DESARROLLO DE LA PROPUESTA DE SOLUCIÓN

- Descripción del fenómeno/proceso modelado y el problema a abordar.
    - Fenómeno: Este proyecto se centra en analizar y optimizar la huella de carbono asociada a las actividades diarias de nuestros empleados, como actualmente no contamos con esta información, para efectos de aprendizaje, utilizaremos el dataset de Kaggle sobre emisiones de carbono.
    - Modelado: Desarrollaremos modelos de aprendizaje automático supervisado con el fin de estimar potenciales huellas de carbono derivadas de las actividades diarias de nuestros empleados, utilizando el dataset de Kaggle como fundamento. Esto nos permitirá identificar áreas clave donde podemos optimizar nuestras prácticas y reducir nuestro impacto ambiental.
    - Problema que abordar: Basándonos en los resultados obtenidos del análisis de datos, implementaremos estrategias dirigidas a minimizar la huella de carbono individual y colectiva dentro de la empresa. Esto incluirá la promoción de alternativas de transporte más verdes, la optimización del uso de recursos en el lugar de trabajo, y la sensibilización sobre la importancia de adoptar hábitos sostenibles.
- Proceso de obtención/generación del conjunto de datos.
    - Selección de la Fuente de Datos: Se eligió Kaggle como fuente debido a su calidad de los datasets disponibles. Fue seleccionado un conjunto de datos específico que ofrece información detallada sobre las emisiones de carbono asociadas a diversas actividades y procesos. Link del dataset: https://www.kaggle.com/code/olgaeleftherakou/carbon-footprint-eda-rf-regression/notebook

- **Exploración Inicial:** Realizamos una exploración inicial del dataset para entender su estructura, incluyendo las variables disponibles, el rango de datos cubiertos, y la relevancia de estos datos para modelar la huella de carbono.
- **Análisis Exploratorio de Datos (EDA):** Un análisis exploratorio de datos profundo fue ejecutado para examinar la calidad de los datos, identificar valores atípicos, y entender las correlaciones entre diferentes variables. Este paso es crucial para asegurar que el conjunto de datos esté listo para el modelado predictivo.
- **Preparación de Datos:** Basándonos en el EDA, se realizó la limpieza y preprocesamiento del dataset. Esto incluyó el manejo de valores faltantes, la normalización de datos, y la ingeniería de características para optimizar el rendimiento del modelo de aprendizaje automático.

- Descripción del problema de aprendizaje automático.
  - El problema central de aprendizaje automático que tratamos en este proyecto se centra en la predicción y análisis de la huella de carbono asociada a diversas actividades y procesos, utilizando un conjunto de datos de Kaggle. Este problema tiene un enfoque de aprendizaje supervisado, donde el objetivo es desarrollar un modelo predictivo capaz de estimar las emisiones de carbono basándose en una serie de variables o características.
    - **Predicción de la Huella de Carbono:** Utilizar las variables disponibles en el conjunto de datos para predecir la cantidad de emisiones de carbono generadas por diferentes actividades. Esto implica la construcción de un modelo que pueda entender y cuantificar la relación entre las características de las actividades (como el tipo de actividad, intensidad energética, entre otros) y sus correspondientes emisiones de carbono.
    - **Identificación de Factores Significativos:** Determinar cuáles son las variables más influyentes en la generación de emisiones de carbono.
    - **Regresión de Bosques Aleatorios:** Para abordar este problema, se seleccionó la técnica de regresión de bosques aleatorios debido a su robustez y eficacia en el manejo de datos complejos y no lineales. Esta metodología permite construir un modelo basado en múltiples árboles de decisión que trabajan conjuntamente para mejorar la precisión de las predicciones.
    - **Evaluación del Modelo:** La efectividad del modelo se evaluará mediante métricas específicas de rendimiento para modelos de regresión, como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Estas métricas ayudarán a determinar la precisión de las predicciones del modelo y su capacidad para explicar la variabilidad de las emisiones de carbono.

- Visualización y preprocesamiento de los datos.
  - Visualización de Datos:
    - **Exploración Inicial:** Empleamos herramientas de visualización para realizar una exploración inicial del dataset, utilizando gráficos como histogramas, diagramas de dispersión y gráficos de caja para examinar la distribución de las variables y la relación entre ellas. Esto ayudó a identificar patrones, tendencias y posibles anomalías en los datos.
    - **Correlaciones:** Se utilizaron mapas de calor para visualizar las correlaciones entre diferentes variables. Esto fue crucial para entender cómo las distintas características se relacionan entre sí y su potencial impacto en las emisiones de carbono.
  - Preprocesamiento de Datos:
    - **Limpieza de Datos:** La limpieza de datos involucró la identificación y manejo de valores faltantes, eliminando o imputando datos según fuera apropiado para mantener la integridad del conjunto de datos.
    - **Transformación de Variables:** Realizamos transformaciones de variables, incluyendo la normalización de datos y la conversión de variables categóricas en formatos numéricos a través de técnicas como el encoding one-hot, para facilitar su análisis por parte de los algoritmos de aprendizaje automático.
    - **Ingeniería de Características:** La ingeniería de características fue empleada para crear nuevas variables derivadas que pudieran tener una influencia significativa en las emisiones de carbono, basándose en el conocimiento obtenido durante la fase de visualización. Esto incluyó la agrupación de variables relacionadas y la creación de índices compuestos que reflejaran aspectos específicos de las actividades y su impacto ambiental.
    - **Selección de Variables:** Utilizamos técnicas de selección de variables para identificar y conservar solo aquellas características que aportaban información relevante para el modelo, eliminando variables redundantes o poco informativas para simplificar el modelo y mejorar su rendimiento.

CONCLUSIONES

Esta tabla muestra la importancia de cada característica en nuestro modelo de aprendizaje automático, indicando cuánto contribuye cada una al poder predictivo del modelo. La distancia mensual recorrida por vehículos es la más influyente, seguida por la frecuencia de viajes aéreos y el tipo de vehículo. Otras características como la cantidad de ropa nueva comprada mensualmente y la frecuencia semanal de disposición de bolsas de basura también son relevantes, aunque en menor medida. Factores como el género, el tipo de cuerpo, y el tamaño de las bolsas de basura tienen un impacto menor. La eficiencia energética y el modo de transporte son los menos determinantes.

TICKETS DE LA GESTIÓN DEL PROYECTO

#	TICKET	DESCRIPCIÓN
1	Revisión de requisitos y definición de alcances	Teniendo en cuenta los requisitos planteados por el docente, se elabora el presente documento en PDF que soporta la selección del problema y el desarrollo de la propuesta solución
2	EDA (Análisis Exploratorio de Datos)	Realizar un análisis exploratorio de datos del dataset de emisiones de carbono. Incluir la visualización de las primeras filas, revisión de tipos de datos, detección y manejo de valores duplicados y faltantes, y análisis de distribuciones de variables clave mediante gráficos de barras. Finalmente, calcular y visualizar la matriz de correlación para identificar relaciones entre variables. Utilizar librerías como pandas, seaborn y matplotlib para las visualizaciones
3	Preprocesamiento de Datos	Aplicar técnicas de preprocesamiento al dataset de emisiones de carbono, incluyendo la sustitución de valores NaN por un marcador específico, transformación de variables categóricas mediante Label Encoding, y la preparación de los datos para el modelado. Asegurar que los datos estén limpios y listos para su análisis posterior y la construcción de modelos predictivos. Utilizar pandas y scikit-learn para el procesamiento de los datos.
4	Modelado con Random Forest	Desarrollar un modelo predictivo utilizando el algoritmo de Random Forest para estimar la huella de carbono a partir del dataset procesado. Configurar el modelo con parámetros iniciales específicos, incluyendo el número de estimadores y la semilla aleatoria para reproducibilidad. Entrenar el modelo con el conjunto de datos dividido en variables independientes (X) y dependiente (y). Evaluar el rendimiento del modelo utilizando el score OOB como métrica interna de validación
5	Evaluación del Modelo	Evaluar el rendimiento del modelo de Random Forest desarrollado, utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Estas métricas ayudarán a determinar la precisión de las predicciones del modelo y su capacidad para explicar la variabilidad de las emisiones de carbono. Analizar las predicciones frente a los valores reales para identificar áreas de mejora y ajustar el modelo según sea necesario

REFERENCIAS

- https://www.kaggle.com/code/olgaeleftherakou/carbon-footprint-eda-rf-regression#chapter1
- https://www.kaggle.com/code/olgaeleftherakou/carbon-footprint-eda-rf-regression#chapter2
- https://www.kaggle.com/code/olgaeleftherakou/carbon-footprint-eda-rf-regression#chapter3
- https://www.kaggle.com/datasets/dumanmesut/individual-carbon-footprint-calculation?resource=download