INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

Entrega Final: Predicción de las emisiones de CO2

Presentado a:
Raul Ramos Pollan

Presentado por:
Michael Stiven Zapata Giraldo
1017172358
Estudiante

Daniel Andres Vergara de Leon
1002388181
Estudiante

Wilmer Arley De Jesus Ceballos Hoyos
1000290030
Estudiante

Universidad De Antioquia

Facultad de Ingeniería

Medellín

2023

INTRODUCCIÓN

La predicción precisa de emisiones de dióxido de carbono (CO_2) es esencial para la toma de decisiones informada en diversos contextos, desde la gestión ambiental hasta la planificación energética. En este proyecto, abordamos el desafío de prever las emisiones de CO_2 para el año 2022 en los diferentes lugares de Rwanda (497), considerando sus condiciones específicas, que van desde zonas rurales y urbanas hasta plantas de energía. Para lograr esto, utilizamos un conjunto de datos recopilado de Kaggle, que consta de 79,023 muestras ordenadas por año, número de semanas del año, densidad, latitud y longitud.

Métricas de Desempeño:

Como métrica de machine learning, empleamos el Root Mean Squared Error (RMSE), que es coherente con la propia competencia y ofrece una medida de la precisión de las predicciones. Para evaluar el rendimiento desde una perspectiva comercial, consideramos la disminución en el nivel de emisiones como un indicador clave. La reducción efectiva de emisiones podría traducirse en beneficios económicos, como incentivos ecológicos, o en la preparación para evitar multas por emisiones elevadas.

Desempeño Deseable en Producción:
Se establece como criterio que el modelo debe tener un desempeño tal que las predicciones de emisiones no se desvíen en más del 5%. Este límite se elige cuidadosamente, ya que un error mayor podría resultar en costos significativos asociados con multas por incumplimiento y los gastos operativos y de actualización del modelo no serían compensados. En otras palabras, el modelo debe ser lo suficientemente preciso como para evitar consecuencias financieras negativas y fomentar prácticas sostenibles.

EXPLORACIÓN DESCRIPTIVA DEL DATASET

El objetivo de este desafío es crear modelos de aprendizaje automático que utilicen datos de emisiones de fuente abierta (de observaciones del satélite Sentinel-5P) para predecir las emisiones de carbono.

Se seleccionaron aproximadamente 497 ubicaciones únicas de múltiples áreas de Ruanda, distribuidas entre tierras agrícolas, ciudades y plantas de energía. Los datos de esta competición están divididos por tiempo; Los años 2019 - 2021 están incluidos en los datos de entrenamiento y su tarea es predecir los datos de emisiones de CO2 desde 2022 hasta noviembre.

Se extrajeron siete características principales semanalmente de Sentinel-5P desde enero de 2019 hasta noviembre de 2022. Cada característica (dióxido de azufre, monóxido de carbono, etc.) contiene subcaracterísticas como column_number_density, que es la densidad de la columna vertical a nivel del suelo, calculada utilizando la técnica DOAS. Puede leer más sobre cada característica en los enlaces a continuación, incluido cómo se miden y las definiciones de las variables. Se le proporcionan los valores de estas características en el conjunto de prueba y su objetivo de predecir las emisiones de CO2 utilizando información de tiempo además de estas características.

Dióxido de azufre - COPERNICUS/S5P/NRTI/L3_SO2

Monóxido de carbono - COPERNICUS/S5P/NRTI/L3_CO

Dióxido de nitrógeno - COPERNICUS/S5P/NRTI/L3_NO2

Formaldehído - COPERNICUS/S5P/NRTI/L3_HCHO

Índice de aerosoles UV - COPERNICUS/S5P/NRTI/L3_AER_AI

Ozono - COPERNICUS/S5P/NRTI/L3_O3

Nube - COPERNICUS/S5P/OFFL/L3_CLOUD

ITERACIONES DE DESARROLLO

- **Preprocesado de datos:**
  El primer paso de esta fase implicó la adquisición de datos, los cuales fueron descargados desde Kaggle y posteriormente añadidos a un Google Drive público. Esta acción fue deliberada para asegurar un acceso continuo a los datos, independientemente de la disponibilidad futura de la competencia en Kaggle. Esta estrategia no solo garantizó la accesibilidad a los datos en el futuro, sino que también eliminó la necesidad de utilizar un usuario y contraseña específicos, simplificando considerablemente el manejo y acceso a la información.

  Seguidamente, implementamos una estrategia de manejo de valores faltantes para abordar la presencia de datos ausentes. Esta estrategia implica el reemplazo de los valores faltantes, representados como NaN, utilizando distintas técnicas como la imputación de la media o promedio, la moda (valor más frecuente en una columna), asignación de un valor por defecto, eliminación de registros incompletos o incluso interpolación de datos. En nuestro caso específico, optamos por una solución que implicó rellenar las celdas con valores NaN utilizando ceros. Esta decisión se tomó considerando la naturaleza de nuestros datos y las implicaciones en el procesamiento posterior. El reemplazo con ceros fue una estrategia viable y adecuada para nuestra situación particular.

- **Modelos supervisados:**

  En nuestro enfoque, hemos empleado una variedad de métodos para abordar la tarea de predicción de nuestro objetivo. Hemos explorado métodos clásicos como la regresión lineal y árboles de decisión, los cuales se centran en establecer relaciones entre las características de entrada y la variable objetivo que deseamos predecir.

  Además, para capturar patrones complejos y secuenciales, hemos incorporado técnicas más avanzadas, como Long Short-Term Memory (LSTM), que forma parte de las redes neuronales. Estos modelos se

destacan por su capacidad para capturar dependencias a largo plazo en los datos, siendo especialmente eficaces en situaciones donde la secuencia y el contexto son relevantes para la predicción.

Adicionalmente, hemos empleado algoritmos de ensamblaje como gradient boosting y ElasticNet. Estos métodos están diseñados para combinar múltiples modelos más simples para formar un modelo más robusto y preciso, maximizando así nuestra capacidad para predecir con precisión el objetivo deseado.

- **Modelos no supervisados:**

  No utilizamos estos tipos de métodos, puesto que, consultando en la web, la falta de supervisión hace más complejo entrenar, evaluar y mejorar modelos no supervisados en comparación con modelos supervisados. Hay que tener cuidado al implementarlos y analizar sus resultados.

- **Resultados, métricas y curvas de aprendizaje:**

  Resultados: Los resultados obtenidos revelaron diferencias significativas en la precisión entre los distintos métodos utilizados para predecir las emisiones de carbono. Tanto la regresión lineal como la Regresión con Bosques Aleatorios arrojaron errores cuadráticos de 142.28 y 27.92, respectivamente. Aunque estos métodos son útiles, mostraron ciertas limitaciones en la precisión de sus predicciones.

  Por otro lado, la implementación del método Long Short-Term Memory (LSTM) produjo un resultado excepcional, con un error cuadrático reducido drásticamente a 0.046. Este rendimiento sobresaliente destaca la capacidad de LSTM para capturar patrones complejos y secuenciales en los datos, superando ampliamente a los métodos tradicionales de regresión.

  Además, se observaron otros resultados interesantes: el uso de Gradient Boosting generó un error de 33.03, mientras que ElasticNet obtuvo un error de 126.79. Estos resultados muestran una variación en la precisión entre estos métodos y resaltan la superioridad de LSTM en términos de precisión para este conjunto de datos específico de emisiones de carbono.

El error por debajo de 1 en las predicciones de LSTM subraya su precisión notable y confirma su eficacia para modelar relaciones complejas en comparación con los demás métodos empleados.

Métricas: En todos los modelos la métrica usada fue root mean squared error

Regresión lineal:

```
[ ]  #Verificamos con el Root Mean Squared Error
     mse = mean_squared_error(data_1[['emission']], data_1[['Emission_predict']])
     rmse = np.sqrt(mse)
     print("Root Mean Squared Error")
     print(rmse)

     Root Mean Squared Error
     142.2755970461999
```

Regresión con bosques aleatorios

```
[ ]  #Verificamos con el Root Mean Squared Error
     mse = mean_squared_error(y_test, y_pred)
     rmse = np.sqrt(mse)
     print("Root Mean Squared Error")
     print(rmse)

     Root Mean Squared Error
     27.919856619277812
```

XGBoost:

```
     XGBoost - Root Mean Squared Error: 33.033947562557024
     XGBoost - Evaluation with Real Data:
```

ElasticNet:

```
ElasticNet - Root Mean Squared Error: 126.79215304204597
ElasticNet
```

Long Short-Term Memory::

```
#Verificamos con el Root Mean Squared Error
mse = mean_squared_error(y_true_inverse, predictions_inverse)
rmse = np.sqrt(mse)
print("Root Mean Squared Error")
print(rmse)

1976/1976 [==============================] - 7s 3ms/step
Shape of y_true_inverse: (63208, 1)
Shape of predictions_inverse: (63208, 1)
Root Mean Squared Error
0.046756518829530316
```

Durante las etapas iniciales del proyecto, se exploró la exclusión de ciertas columnas de datos de las variables independientes, incluyendo el año, número de semana, longitud y latitud. Sin embargo, este enfoque no produjo resultados satisfactorios, evidenciado por un error significativo en el modelo de regresión con bosques aleatorios, el cual se situaba alrededor de 130 aproximadamente.

Tras una serie de pruebas, se observó que la inclusión específica de dos columnas, longitud y latitud, mejoraba notablemente el rendimiento de los modelos. Esta revelación impulsó la calibración de los modelos existentes, logrando reducir el error en el modelo de regresión con bosques aleatorios a aproximadamente 28, como se detalla en etapas anteriores.

A medida que se avanzaba, se comprendió la importancia crucial de estas dos columnas y se mantuvo su inclusión en los nuevos modelos. No obstante, surgió la necesidad de capturar patrones más complejos y secuenciales presentes en los datos. Fue entonces cuando se exploró y adoptó el modelo de Long Short-Term Memory (LSTM).

Este cambio de enfoque hacia LSTM se basó en la búsqueda de un modelo capaz de identificar y aprovechar las complejidades secuenciales inherentes en los datos. Esta decisión marcó un punto crucial en el proyecto al dirigirnos hacia un enfoque más avanzado y especializado en la captura de patrones secuenciales, impulsando así la exploración de este modelo en particular.

RETOS Y CONSIDERACIONES DE DESPLIEGUE

1. **Preprocesamiento de Datos:**
   Limpieza y Manejo de Datos Ausentes: La estrategia utilizada para abordar valores faltantes consistió en rellenar celdas con valores NaN utilizando ceros. Esta decisión se basó en la naturaleza de los datos y las implicaciones en el procesamiento posterior. Sin embargo, es crucial considerar otras estrategias de manejo de valores faltantes, como la imputación de la media, la moda o la interpolación, para evaluar su impacto en el desempeño del modelo.

2. **Selección y Optimización de Modelos:**
   Experimentación Continua: La evaluación de modelos incluyó métodos clásicos como regresión lineal y árboles de decisión, así como enfoques más avanzados como Long Short-Term Memory (LSTM), gradient boosting y ElasticNet. La experimentación debe ser continua para explorar nuevos modelos, ajustar hiperparámetros y combinar métodos con el objetivo de mejorar la precisión.

3. **Dimensionamiento Computacional:**
   Infraestructura y Recursos: El despliegue efectivo de modelos complejos puede requerir recursos computacionales específicos, como GPUs o clusters de alto rendimiento. Garantizar la disponibilidad de la infraestructura necesaria y su capacidad para manejar el volumen de datos es esencial.

4. **Monitoreo y Recalibración:**
   Adaptabilidad a Cambios: La implementación de un sistema de monitoreo continuo es crucial para evaluar el rendimiento del modelo en producción. Se debe establecer un proceso para la recalibración periódica de los modelos, considerando cambios en los datos y evitando la degradación del rendimiento con el tiempo.

5. **Interpretabilidad de Modelos:**
   Explicabilidad de Predicciones: La interpretación de los modelos es esencial para la toma de decisiones y la corrección de posibles sesgos. Considerar modelos más interpretables, como árboles de decisión, puede facilitar la auditoría y asegurar la transparencia en las predicciones.

6. **Aspectos Éticos:**
   Evitar Sesgos y Garantizar Justicia: Asegurar que los modelos no repliquen sesgos existentes en los datos históricos es crucial. Se deben implementar medidas para garantizar la justicia y transparencia en los resultados, especialmente cuando las predicciones afectan decisiones sobre personas o comunidades.

Desafíos Técnicos y de Despliegue:
Integración en Sistemas en Tiempo Real: Desplegar modelos predictivos en sistemas del mundo real implica desafíos técnicos y de gestión. Asegurar la integración efectiva de los modelos, considerando la latencia y la capacidad de respuesta en tiempo real, es esencial más allá de la precisión teórica de las predicciones.

Este análisis continuo y adaptación a desafíos y cambios es fundamental para el éxito a largo plazo del despliegue de modelos predictivos en un entorno operativo.

CONCLUSIONES

- El proyecto se enfocó en predecir las emisiones de dióxido de carbono (CO_2) en diferentes ubicaciones de Ruanda utilizando datos recopilados del satélite Sentinel-5P. Exploramos una variedad de modelos de aprendizaje automático, desde enfoques clásicos hasta técnicas avanzadas, con el objetivo de encontrar el modelo más preciso para esta tarea.
- Durante la iteración inicial, se enfrentaron desafíos al descubrir la relevancia de ciertas características, como la longitud y la latitud, que tuvieron un impacto significativo en la precisión de los modelos. Esta comprensión nos llevó a recalibrar los modelos existentes, mejorando drásticamente su desempeño y reduciendo los errores de predicción.
- La implementación del modelo Long Short-Term Memory (LSTM) demostró ser un punto de inflexión crucial. Este modelo, capaz de capturar patrones complejos y secuenciales en los datos, superó notablemente a otros métodos en términos de precisión, reduciendo el error a niveles mínimos.
- Además, se identificaron desafíos cruciales en el despliegue efectivo de modelos, desde la adaptabilidad a cambios en los datos hasta la interpretación y explicabilidad de los resultados para tomar decisiones informadas.
- En resumen, este proyecto resalta la importancia de seleccionar cuidadosamente las características relevantes, explorar técnicas avanzadas como LSTM y enfrentar desafíos críticos relacionados con el despliegue y la interpretación de modelos. Este enfoque multidimensional en la

construcción, evaluación y despliegue de modelos nos ha brindado valiosas lecciones y perspectivas para futuros proyectos de predicción de emisiones de CO2 y otros desafíos de aprendizaje automático.

- La aplicación de un modelo de regresión lineal demostró ser una herramienta valiosa para la predicción de datos en nuestro estudio. La combinación de la simplicidad conceptual de la regresión lineal y su capacidad para proporcionar resultados interpretables lo posiciona como un método efectivo en la caja de herramientas de la modelización predictiva. Sin embargo, es fundamental reconocer las limitaciones inherentes a asumir relaciones lineales y considerar la posibilidad de explorar modelos más complejos en contextos donde las relaciones no lineales son prominentes. Este trabajo sienta las bases para futuras investigaciones y aplicaciones en el ámbito de la predicción de datos mediante técnicas de aprendizaje supervisado.
- XGBoost es una poderosa herramienta en el campo del aprendizaje automático y es adecuado para una variedad de problemas. Sin embargo, como con cualquier algoritmo, es crucial comprender su funcionamiento, ajustar adecuadamente los hiperparámetros y evaluar su rendimiento de manera rigurosa.
- ElasticNet es una herramienta versátil y efectiva en la práctica de la regresión, proporcionando un equilibrio entre la regularización L1 y L2. La elección de este método dependerá de la naturaleza específica del problema y la importancia de la selección de características en el contexto de la tarea de predicción.