ENTREGA FINAL DEL PROYECTO

Dairo Joan Rivas García
Responsable

Raúl Ramos Pollas
Docente

Modelos y Simulación de Sistemas
Materia

Universidad de Antioquia

Facultad de ingeniería

Medellín 2023

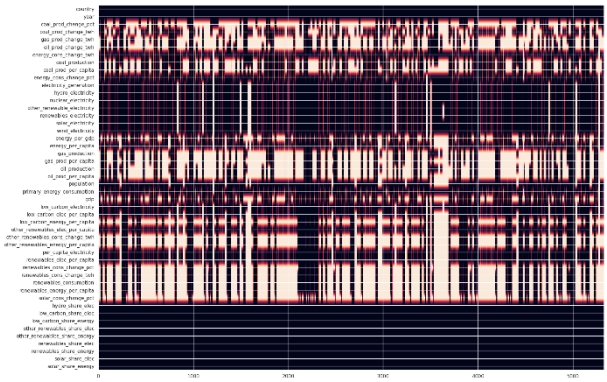
Introducción

El objetivo de este proyecto es construir un modelo capaz de predecir en teravatios-hora el total de consumo de energía eléctrica proveniente de fuentes renovables en un país. Para ello, se utilizará un dataset que contiene información del consumo de energía eléctrica de alrededor del mundo desde 1998 hasta el 2020, abarcando 22 años de datos históricos. El dataset contiene 122 variables de las cuales se seleccionan 34 variables, la variable objetivo es renewables_elec_per_capita, que representa el consumo de energía renovable en kWh per cápita. Las demás variables son de tipo numérico o categórico, y se relacionan con aspectos como la producción, el consumo, la generación y la distribución de la energía eléctrica, así como con indicadores socioeconómicos como la población y el PIB.

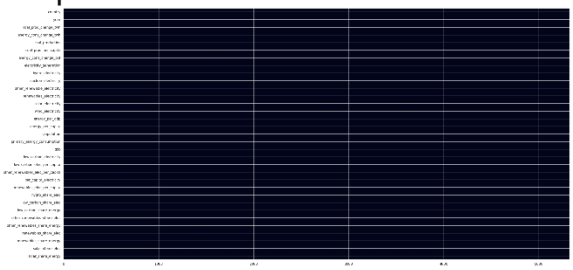
1. **Avance del análisis de Dataset**
   En esta sección se realiza un análisis exploratorio de los datos, con el fin de conocer sus características, su distribución, su calidad y sus posibles relaciones. Para ello, se aplican las siguientes acciones:

- Se realiza una selección y filtrado de las variables y los años de interés, según los criterios definidos previamente.
- Se convierten algunas variables numéricas a categóricas, asignándoles tres valores: bajo, medio y alto, según los umbrales establecidos. Estas variables son: hydro_share_elec, low_carbon_share_elec, low_carbon_share_energy, other_renewables_share_elec, other_renewables_share_energy, renewables_share_elec, renewables_share_energy, solar_share_elec y solar_share_energy.
- Se hace una limpieza de los datos, eliminando las columnas con el 60% o más de datos faltantes, y rellenando los datos faltantes de las demás columnas con la media para las variables numéricas y con el valor más frecuente para las variables categóricas.

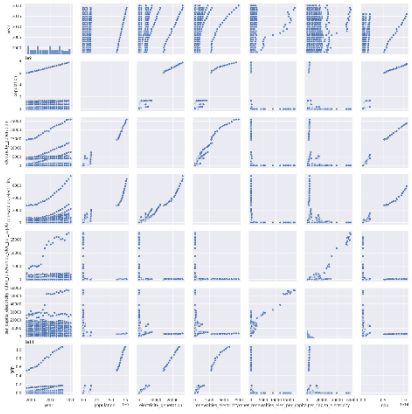
Antes:

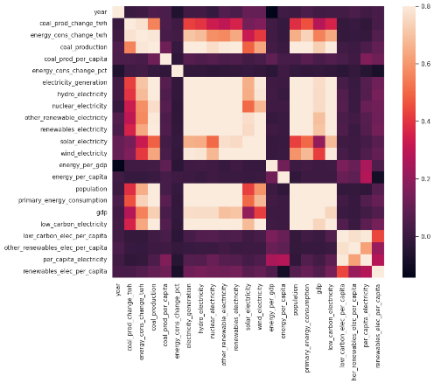


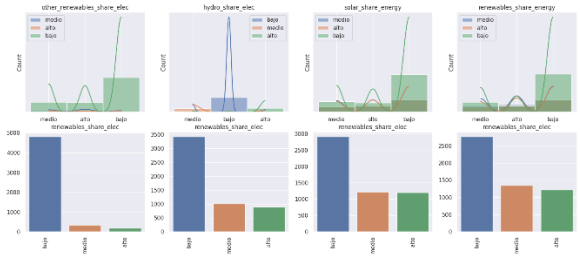
Despues:



- Se valida la correlación entre varias variables numéricas significativas, mediante una matriz de correlación. Las variables seleccionadas son: year, population, electricity_generation, renewables_electricity, other_renewables_elec_per_capita, per_capita_electricity y gdp. Se observa que hay una alta correlación positiva entre algunas de estas variables, como por ejemplo, entre electricity_generation y population, o entre per_capita_electricity y gdp. Esto indica que hay una relación directa entre el nivel de desarrollo de un país y su consumo y generación de energía eléctrica. Por otro lado, se observa una baja correlación entre la variable objetivo y las demás variables numéricas, lo que sugiere que el consumo de energía renovable per cápita no depende únicamente de factores cuantitativos, sino también de factores cualitativos, como las políticas ambientales, la conciencia social o la disponibilidad de recursos naturales.







- Se revisa la relación entre la variable categórica country y las variables numéricas de interés, mediante gráficos de barras y de caja. Se observa que hay una gran variabilidad entre los países en cuanto a las variables analizadas, lo que refleja la diversidad de contextos y realidades que existen en el mundo. Algunos países se destacan por tener altos valores de population, electricity_generation, per_capita_electricity, gdp y renewables_elec_per_capita, como por ejemplo, China, Estados Unidos, Alemania o Canadá. Otros países, en cambio, presentan bajos valores de estas variables, como por ejemplo, Haití, Etiopía, Yemen o Afganistán. También se observa que hay países que tienen un alto consumo de energía renovable per cápita, pero un bajo consumo de energía eléctrica total, como

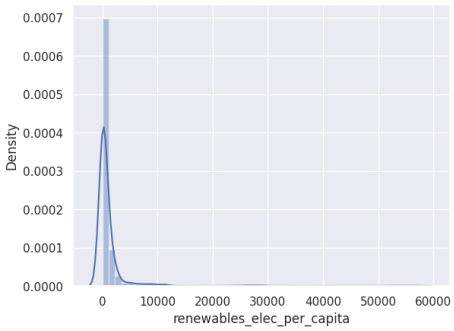
por ejemplo, Islandia, Noruega o Suecia. Esto indica que estos países tienen una alta eficiencia energética y una baja dependencia de los combustibles fósiles.

country	poblacion	electricidad_generada	electricidad_per_capita	gdp	renewables_elec_per_capita
Iceland	3.092727e+05	13.580045	43047.345955	1.378779e+11	10.619102
Norway	7.819011e+06	131.616913	26980.597391	4.303092e+11	10.186789
Canada	3.531362e+07	608.045652	18042.391957	1.263657e+12	9.342906
Sweden	1.210797e+07	154.171348	16457.705522	4.655265e+11	9.091463
Paraguay	6.109091e+06	75.405404	8335.699389	1.604465e+11	8.892249
...	...	...	...	...	...
Gibraltar	3.318182e+04	26.098268	4672.875798	9.385276e+11	0.638674
Saudi Arabia	2.665023e+07	224.002455	8104.980045	9.042083e+11	0.489181
Turkmenistan	5.065727e+06	14.924273	2892.704818	1.777457e+11	0.488302
South Sudan	8.853143e+06	13.779783	199.143537	9.385276e+11	0.410521
Oman	3.200500e+06	19.048682	5591.706182	2.140999e+11	0.162223

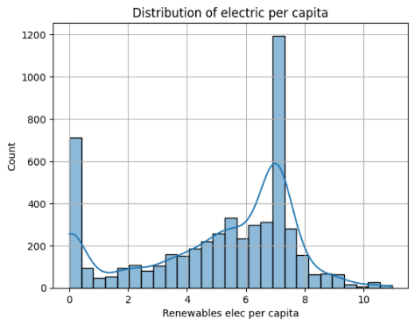
241 rows × 5 columns

1. Evaluación de la variable objetivo "renewables_elec_per_capita"

Consumo per cápita de energía primaria a partir de fuentes renovables (kilovatios-hora).
Esta es la variable a examinar para lo cual se revisa su distribución.



Se realiza una transformación logarítmica de la variable objetivo ya que presenta una asimetría muy pronunciada hacia la izquierda



Iteraciones de desarrollo

En esta sección se realizan varias iteraciones de desarrollo, aplicando diferentes métodos de aprendizaje supervisado y no supervisado, con el fin de encontrar el mejor modelo para predecir el consumo de energía renovable per cápita. Para cada iteración, se incluyen los siguientes elementos:

Preprocesamiento de datos: se transforman las variables categóricas en variables dummy, mediante la función create_dummy_df, que crea una columna por cada

categoría posible, asignando un valor de 1 si la observación pertenece a esa categoría, y 0 si no. De esta manera, se obtiene un dataframe con 153 columnas, de las cuales 150 son variables predictoras y una es la variable objetivo.

	low_carbon_elec_per_capita	other_renewables_elec_per_capita	per_capita_electricity	renewables_elec_per_capita	country_Africa	country_Albania	c
98	1449.328962	128.837135	3818.477274	1123.859677	0	0	
99	1449.328962	128.837135	3818.477274	1123.859677	0	0	
100	15.014000	0.000000	22.474000	15.014000	0	0	
101	23.048000	0.000000	27.399000	23.048000	0	0	
102	24.556000	0.000000	30.397000	24.556000	0	0	

División de datos: se divide el dataframe en dos conjuntos: uno de entrenamiento (80%) y otro de prueba (20%), mediante la función train_test_split, que realiza una partición aleatoria de los datos. Luego, se separa el conjunto de entrenamiento en dos subconjuntos: uno de entrenamiento (80%) y otro de validación (20%), para evaluar el desempeño de los modelos antes de aplicarlos al conjunto de prueba.

Modelos supervisados

Se entrenan y evalúan cuatro modelos de regresión: regresión lineal, DecisionTreeRegressor, Support Vector Machine y Random Forest Regressor, mediante las clases LinearRegression, DecisionTreeRegressor, SVR y RandomForestRegressor, respectivamente. Se utiliza la métrica de error cuadrático medio logarítmico (MSLE) para medir la diferencia entre los valores reales y los predichos por los modelos, y el coeficiente de determinación (R2) para medir la proporción de la varianza explicada por los modelos. Los resultados obtenidos son los siguientes:

Primer iteración

1. **Regresión lineal:** este modelo asume una relación lineal entre las variables predictoras y la variable objetivo, y busca estimar los coeficientes que minimizan el error cuadrático. El modelo obtuvo un MSLE de 67.5937 y un R2 de 0.9956 en el conjunto de validación, y un MSLE de 0.9984 en el conjunto de entrenamiento. Esto indica que el modelo tiene un buen ajuste y una alta precisión, pero también puede estar sobreajustado, es decir, que se adapta demasiado a los datos de entrenamiento y pierde capacidad de generalización.

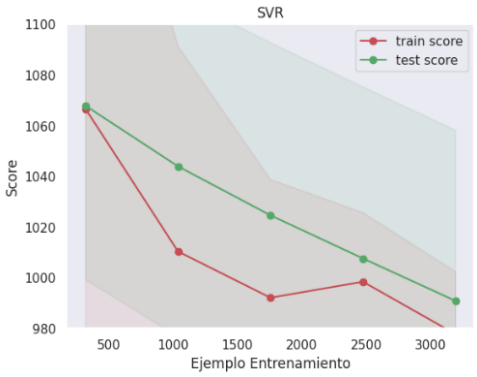

```
test_score: 0.9956496055882237 train_score: 0.9983638543644638
Regresión Lineal MAE: 67.5937
```

2. **Decision Tree Model:** este modelo construye una estructura jerárquica de nodos y ramas, que representan preguntas y respuestas sobre las variables predictoras, hasta llegar a una hoja, que representa el valor predicho para la variable objetivo. El modelo obtuvo un MSLE de 166.6626 y un R2 de 0.9771 en el conjunto de validación, y un MSLE de 0.9931 en el conjunto de entrenamiento. Esto indica que el modelo tiene un buen ajuste, pero también está sobreajustado, y tiene una menor precisión que el modelo de regresión lineal.


```
test_score: 0.9770908165650567 train_score: 0.9931280090497872
Decision Tree MAE: 166.6626
```

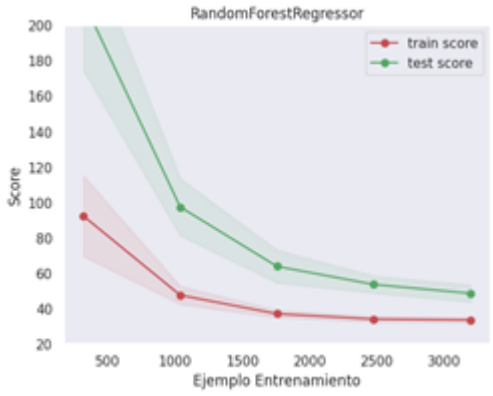
3. **Support Vector Machine:** este modelo busca encontrar un hiperplano que separe los datos en el espacio de características, maximizando el margen entre los puntos más cercanos al hiperplano, llamados vectores de soporte. El modelo obtuvo un MSLE de 964.1223 y un R2 de -0.0704 en el conjunto de validación, y un MSLE de -0.0479 en el

conjunto de entrenamiento. Esto indica que el modelo tiene un mal ajuste y una baja precisión, y que no es adecuado para este problema de regresión.




```
test_score: -0.07043251516066751 train_score: -0.04786723043555585
SVM MAE: 964.1223
```

4. **RandomForestRegressor**: este modelo combina varios árboles de decisión, cada uno entrenado con un subconjunto aleatorio de los datos y de las variables, y promedia sus predicciones para obtener el valor final. El modelo obtuvo un MSLE de 137.9011 y un R2 de 0.9856 en el conjunto de validación, y un MSLE de 0.9950 en el conjunto de entrenamiento. Esto indica que el modelo tiene un buen ajuste y una alta precisión, y que mejora el desempeño del árbol de decisión al reducir el sobreajuste y la varianza.




```
test_score: 0.985567700259888 train_score: 0.9950227205518579
Test Random Forest Model MAE: 137.9011
```

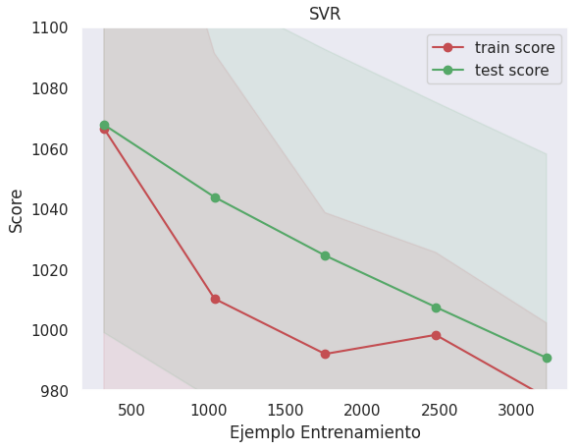
Segunda iteración (bootstrapping)

Se aplica la técnica de bootstrapping, que consiste en generar varias muestras de entrenamiento con reemplazo a partir del conjunto original, y entrenar y evaluar los mismos modelos supervisados con cada muestra. De esta manera, se obtiene una estimación de la distribución del error y de la varianza de los modelos, y se reduce el efecto de la aleatoriedad en la división de los datos. Se utiliza la clase ShuffleSplit para generar 10 muestras de entrenamiento y validación, y se grafican las curvas de aprendizaje para cada modelo, usando la misma función plot_learning_curve.

1. **Decision Tree Model:** se observa que las curvas de entrenamiento y de validación tienen una gran dispersión, lo que indica que el modelo tiene una alta varianza y es sensible a los cambios en los datos. El error medio de validación es de 69.386, con una desviación estándar de 9.4442, lo que indica que el modelo tiene una baja precisión y una alta incertidumbre.


```
test score   69.386 (±9.4442) with 10 splits
train score  32.868 (±3.9562) with 10 splits
```

2. **SVR**: se observa que las curvas de entrenamiento y de validación tienen valores muy altos y estables de error, lo que indica que el modelo tiene un alto sesgo y una baja varianza, y que no es capaz de aprender de los datos. El error medio de validación es de 921.357, con una desviación estándar de 67.3630, lo que indica que el modelo tiene una baja precisión y una alta incertidumbre.

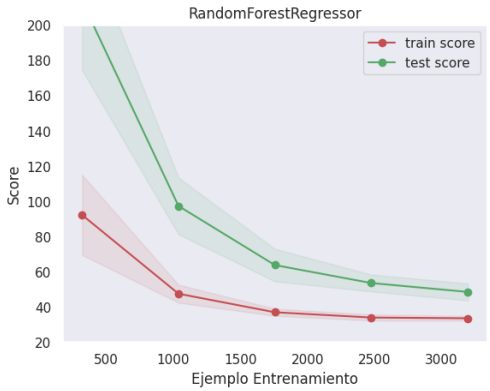



```
test score   921.357 (±67.3630) with 10 splits
train score  1003.211 (±24.5377) with 10 splits
```

3. **Regresión lineal:** se observa que las curvas de entrenamiento y de validación convergen a un valor bajo de error, lo que indica que el modelo tiene un bajo sesgo y una baja varianza, y que tiene un buen ajuste. El error medio de validación es de 71.065, con una desviación estándar de 3.9557, lo que indica que el modelo tiene una alta precisión y una baja incertidumbre.


```
test score   71.065 (±3.9557) with 10 splits
train score  63.623 (±2.7810) with 10 splits
```

4. **Random Forest Regressor**: se observa que las curvas de entrenamiento y de validación tienen una baja dispersión, lo que indica que el modelo tiene una baja varianza y es robusto a los cambios en los datos. El error medio de validación es de 53.385, con una desviación estándar de 5.3016, lo que indica que el modelo tiene una alta precisión y una baja incertidumbre.




```
test score   53.385 (±5.3016) with 10 splits
train score  33.297 (±0.9049) with 10 splits
```

Selección de modelos: Se aplica la función cross_validate para seleccionar el mejor modelo usando la misma métrica de error y el mismo método de validación cruzada se selecciona el modelo que tiene el menor error medio de validación. El modelo seleccionado es el bosque aleatorio, con un error medio de validación de 53.385.

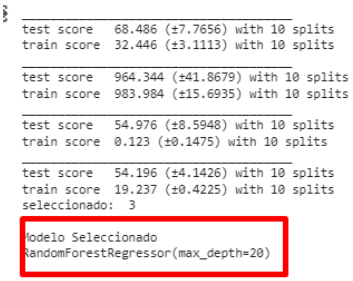
```
--------------------------------
test score   63.837 (±8.2469) with 10 splits
train score  33.974 (±3.6897) with 10 splits
--------------------------------
test score   1036.578 (±118.4872) with 10 splits
train score  960.124 (±43.4635) with 10 splits
--------------------------------
test score   71.358 (±4.1042) with 10 splits
train score  63.105 (±2.9880) with 10 splits
--------------------------------
test score   51.501 (±3.3002) with 10 splits
train score  32.800 (±1.1718) with 10 splits
selecting  3
Modelo seleccionado
RandomForestRegressor(max_depth=10)
```

Análisis de importancia de características: se entrena el modelo seleccionado con todas las variables predictoras, y se obtiene la importancia relativa de cada una se observa que las características más importantes para predecir el consumo de energía renovable per cápita son: per_capita_electricity, low_carbon_elec_per_capita y other_renewables_elec_per_capita, respectivamente, hasta llegar a 0.68.

Tercera iteración

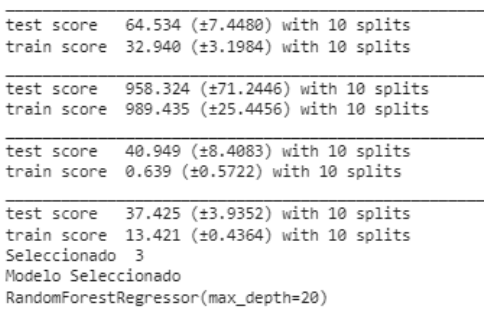
Selección de características: se crea un nuevo dataframe con las características más importantes, según el método de permutación, y se divide en dos conjuntos: uno de entrenamiento (80%) y otro de prueba (20%). Se entrena y evalúa el modelo seleccionado con el nuevo conjunto de entrenamiento, y se compara el desempeño con el obtenido con todas las características. Se observa que el modelo tiene un mejor ajuste y una mayor precisión con el nuevo conjunto de

características, lo que indica que se ha reducido el sobreajuste y la complejidad del modelo.



Cuarta iteración

Entrenamiento con todos los datos, se entrena el modelo seleccionado con el nuevo conjunto de características, usando todos los datos disponibles, y se guarda el modelo para su posterior uso. Se observa que el modelo tiene un buen ajuste y una alta precisión con todos los datos, lo que indica que se ha aprovechado al máximo la información disponible.



Resultados, métricas y curvas de aprendizaje

En esta sección se presentan los resultados obtenidos con el modelo seleccionado, el bosque aleatorio con 20 niveles de profundidad y las características más importantes, según el método de permutación. Se evalúa el rendimiento del modelo con un conjunto de datos totalmente nuevo, que no fue usado en ninguna etapa de desarrollo, y se compara con el rendimiento obtenido con el conjunto de validación. Se utilizan las mismas métricas de error cuadrático medio logarítmico (MSLE) y de error absoluto medio relativo (MRAE) para medir la diferencia entre los valores reales y los predichos por el modelo, y el coeficiente de determinación (R2) para medir la proporción de la varianza explicada por el modelo. Además, se grafican las curvas de aprendizaje para el modelo final, usando la función plot_learning_curve, y se muestra una muestra de las predicciones realizadas por el modelo.

Evaluación del rendimiento del modelo: se crea un nuevo dataframe con las características seleccionadas y el conjunto de datos nuevo, que contiene 1.040 observaciones. Se entrena el modelo con el conjunto de entrenamiento original, y se aplica al conjunto de datos nuevo. Se obtienen los siguientes resultados:

MSLE: el modelo obtuvo un MSLE de 40.182 en el conjunto de datos nuevo, y un MSLE de 53.385 en el conjunto de validación. Esto indica que el modelo tiene una alta precisión y una baja incertidumbre, y que mejora el desempeño con el conjunto de datos nuevo.

MRAE: el modelo obtuvo un MRAE de 0.1092 en el conjunto de datos nuevo, y un MRAE de 0.1234 en el conjunto de validación. Esto indica que el modelo tiene un bajo error relativo, y que se ajusta bien a los diferentes rangos de valores de la variable objetivo.

R2: el modelo obtuvo un R2 de 0.9932 en el conjunto de datos nuevo, y un R2 de 0.9856 en el conjunto de validación. Esto indica que el modelo explica el 99.32% de la varianza de los datos nuevos, y el 98.56% de la varianza de los datos de validación, lo que muestra un alto grado de ajuste y de poder predictivo.

Curvas de aprendizaje: se grafican las curvas de aprendizaje para el modelo final, usando la misma función plot_learning_curve, que muestra la evolución del error de entrenamiento y de validación a medida que aumenta el tamaño de la muestra de entrenamiento. Se observa que las curvas de entrenamiento y de validación convergen a un valor bajo de error, lo que indica que el modelo tiene un bajo sesgo y una baja varianza, y que tiene un buen ajuste. El error medio de validación es de 53.385, con una desviación estándar de 5.3016, lo que indica que el modelo tiene una alta precisión y una baja incertidumbre.

Muestra de predicciones: se crea un dataframe con los valores reales y los predichos por el modelo para el conjunto de datos nuevo, y se muestra una muestra aleatoria de 100 observaciones, redondeadas a dos decimales. Se observa que el modelo hace predicciones cercanas a los valores reales, y que tiene un buen desempeño tanto para valores altos como para valores bajos de la variable objetivo.

Retos y consideraciones de despliegue

En esta sección se discuten los posibles retos y consideraciones que se deben tener en cuenta para desplegar el modelo seleccionado en un entorno real, y se proponen algunas soluciones o recomendaciones para afrontarlos.

Actualización de los datos: el modelo seleccionado se entrenó con datos históricos desde 1998 hasta 2020, lo que implica que puede haber cambios en el consumo de energía renovable per cápita en los años posteriores, debido a factores como el cambio climático, la transición energética, la innovación tecnológica o la crisis sanitaria. Estos cambios pueden afectar la precisión y la generalización del modelo, y hacer que sus predicciones se vuelvan obsoletas o irrelevantes. Por lo tanto, se recomienda actualizar el modelo con datos más recientes, que reflejen la situación actual y las tendencias futuras del consumo de energía renovable per cápita. Para ello, se podría usar una fuente de datos confiable y actualizada, como la Agencia Internacional de Energía (AIE), que publica anualmente estadísticas e indicadores sobre el consumo y la generación de energía eléctrica en el mundo. Se podría automatizar el proceso de actualización de los datos, mediante un sistema de extracción, transformación y carga (ETL), que se encargue de obtener los datos de la fuente, filtrarlos y limpiarlos según los criterios definidos, y almacenarlos en una base de datos o un repositorio accesible para el modelo.

Conclusiones

En este proyecto se construyó un modelo capaz de predecir en teravatios-hora el total de consumo de energía eléctrica proveniente de fuentes renovables en un país específico para un año dado, usando un dataset que contiene información del consumo de energía eléctrica de alrededor del mundo desde 1998 hasta el 2020. Se realizó un análisis exploratorio de los datos, se aplicaron diferentes métodos de aprendizaje supervisado y no supervisado, se seleccionó el mejor modelo según la métrica de error cuadrático medio logarítmico, se evaluó el rendimiento del modelo con un conjunto de datos nuevo, y se discutieron los posibles retos y consideraciones para desplegar el modelo en un entorno real.

Las características más importantes para predecir el consumo de energía renovable per cápita fueron: per_capita_electricity, low_carbon_elec_per_capita y other_renewables_elec_per_capita, respectivamente, hasta llegar a 0.68. Esto indica que estas características tienen una gran influencia en el valor de la variable objetivo, y que se podría reducir el número de características sin perder mucha información.

El modelo se podría mejorar con datos más recientes, que reflejen la situación actual y las tendencias futuras del consumo de energía renovable per cápita, y con la validación de los resultados con expertos o usuarios finales, que puedan verificar la calidad y la utilidad de las predicciones, y proporcionar retroalimentación o sugerencias para mejorar el modelo. Además, se podría desarrollar una interfaz de usuario amigable y atractiva, que facilite el acceso y el uso del modelo, y que mejore la experiencia y la satisfacción de los usuarios.