Diego Alejandro Peña
Alejandra Guerrero
Luis Enrique García

TABLA DE CONTENIDO

Diego Alejandro Peña
Alejandra Guerrero
Luis Enrique García

1. Planteamientos y objetivos

El objetivo planteado inicialmente fue entregar al usuario final una aplicación que le permitiera generar cambios en los hábitos de consumo de energía eléctrica mediante la predicción del consumo energético que tendría en su vivienda dependiendo de factores como Temporada del año, humedad, electrodomésticos que utilicen, entre otras.

Para esto buscamos un set de datos que nos proporcione información sobre el comportamiento del uso de energía en una vivienda promedio lo más acertada posible. Esto nos permitiría construir un modelo de proyección, que no supere el 20% de error en la predicción.

Una vez encontrado este set de datos en Kaggle, reafirmamos nuestros objetivos principales:

1. Proporcionar información de consumo al usuario que permita describir sus hábitos de uso de energía eléctrica.
2. Obtener la predicción de uso (y costo) que tendrá la energía eléctrica en una vivienda, debido a su comportamiento durante un cierto periodo.

Adicionalmente, el mayor valor que puede generar este modelo de predicción en futuras versiones es permitir a una ciudad planificar nuevas fuentes de energía dependiendo del crecimiento en número de unidades habitaciones. Sin embargo, esto sería en futuros Sprint con mayor volumen de datos.

Los Sprint de este proyecto fueron:

- Análisis exploratorio de datos (Sprint No. 1 – Entrega No. 1)
- Construcción de varios modelos de predicción para hallar el mejor algoritmo de predicción para los datos con los que contamos (Sprint No. 2 – Entrega No. 2).
- Construcción de una página web en donde el usuario pueda consultar predicciones de consumo energético (Sprint No. 3 – Entrega No. 3)

2. Estrategia de validación y selección del modelo

La definición del modelo usado continúo con el mismo enfoque que se trabajó en la entrega No 2 la cual consistió en la construcción de varios modelos de regresión, (para esta entrega se incluyó el modelo de regresión XGBoost) que permitan comparar resultados de predicción, pero cambiando el nivel de detalle de las predicciones y realizando las predicciones por hora y no por minuto. Este último cambio se llevó a cabo al concluir que para un usuario no tiene sentido conocer la estimación de su consumo por minuto, tendría mayor sentido por día o mes. Sin embargo, agrupar los datos por día o mes nos dejaría únicamente 365 registros o 12 registros respectivamente, valores que no serían suficientes para entrenar un modelo.

Diego Alejandro Peña
Alejandra Guerrero
Luis Enrique García

2.1. Estrategia de experimentación para la selección del mejor modelo

Dado que se buscaba predecir el consumo energético “Use [kW]”, se buscó responder el problema considerando regularización. Por tanto, se decidió empezar entrenando modelos simples como regresión lineal y polinómica sin regularización, y con regularización L1 y L2. Por otro lado, se implementaron otro tipo de algoritmos de regresión como árboles de decisión y XGBoost. Cada modelo se lanzó con un conjunto de parámetros de prueba por medio del algoritmo de búsqueda en grilla (GridSearchCV), el cual combina e itera los parámetros para encontrar los que mejores resultados generan; y adicionalmente, se encarga de realizar la validación cruzada (cross validation) con 5 divisiones (folds).

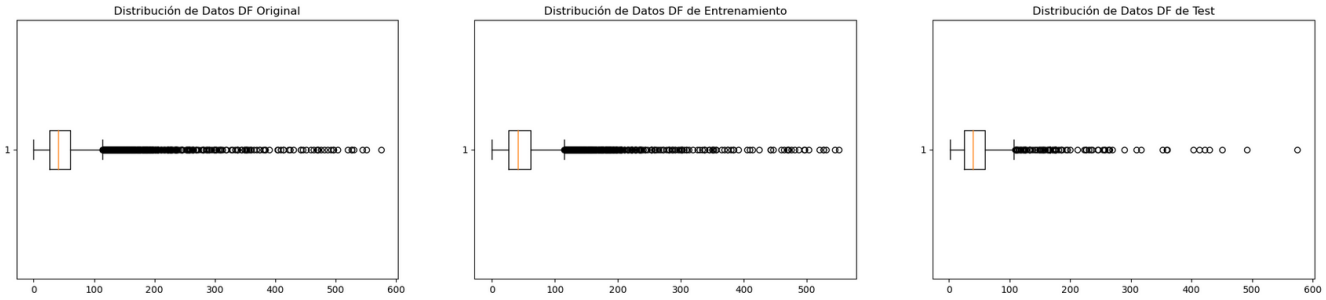
Para la selección de features decidimos considerar todas las variables ambientales y de consumo por electrodoméstico. Como el set de datos tiene una marca de tiempo, inicialmente hicimos pruebas suponiendo que la estación del año, día de la semana, y parte del día eran valores ordinales. Luego, realizamos pruebas asumiendo que las variables no eran ordinales y en consecuencia debían separarse en distintas columnas. Los resultados mejoraron bajo el último supuesto, lo cual creemos que tiene sentido puesto que los algoritmos usados no toman en consideración valores pasados o las características propias de series de tiempo.

2.2. Selección de medida de error

Cada modelo fue evaluado teniendo en cuenta métricas de error acordes a un problema de regresión. En este caso se consideró: el MAE porque es menos susceptible a errores, mientras que el RMSE castiga en mayor medida los errores grandes.

2.3. División de los datos

El conjunto inicial de datos se agrupó por hora, pasando de 503.910 registros a 8.398 registros. Posteriormente, el set se dividió para dejar un 80% de los datos para el entrenamiento del modelo y un 20% para las pruebas del modelo y evaluación de su desempeño. Los dos sets de datos se compararon con el dataset original para corroborar que la distribución de los datos se mantuviera igual. En este pasó se identificó que, a pesar de que los datos fueron agrupados, siguen presentando una gran cantidad de outliers para los cuales los modelos no pueden identificar la forma de estimar un valor que se acerque al valor real atípico. Al indagar por la presencia de esos outliers, creemos que es un comportamiento que tiene sentido para la variable energía en rangos de hora, puesto que el uso de ciertos electrodomésticos ocurre de forma esporádica.



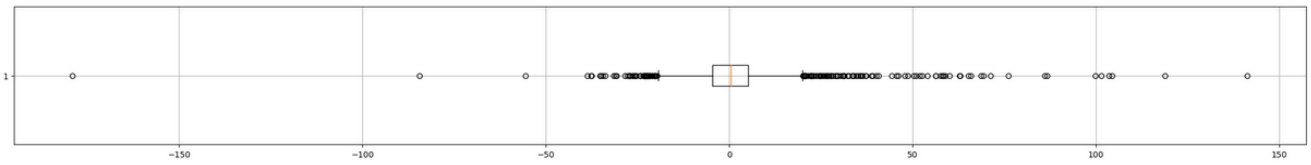
3. Construcción y evaluación del mejor modelo

Se construyeron 6 modelos para el análisis de los datos con el objetivo de realizar una predicción de consumo de energía lo más acertada posible. Los resultados de cada modelo se muestran a continuación:

Modelo	Indicador	Entrenamiento	Prueba	Diferencia
Regresion lineal simple	MAE	18.729700	18.395920	0.333780
Regresion Polinomial sin regularización	MAE	7.377071	8.384821	-1.007750
Regresion Polinomial Lasso L1	MAE	7.371626	8.371111	-0.999485
Regresion Polinomial Ridge L2	MAE	7.222146	7.488357	-0.266211
Arbol de Decisión	MAE	11.399242	13.014461	-1.615219
XGBoost	MAE	6.924382	8.543422	-1.619040
Regresion lineal simple	RMSE	30.633959	29.735524	0.898435
Regresion Polinomial sin regularización	RMSE	12.320524	14.805394	-2.484870
Regresion Polinomial Lasso L1	RMSE	12.882060	14.069748	-1.187688
Regresion Polinomial Ridge L2	RMSE	12.323311	14.781928	-2.458617
Arbol de Decisión	RMSE	17.493118	21.109245	-3.616127
XGBoost	RMSE	10.885534	15.139550	-4.254016

En conclusión, los modelos entrenados no presentaron problemas de overfitting, gracias a la selección de features que se realizó en distintas iteraciones de ejecución de los modelos. Como tal, el margen de error de los modelos no varía considerablemente entre modelo, excepto árboles de decisiones que tuvo el peor resultado. Vale la pena destacar que el modelo de regresión polinomial no mejoró al aplicar regularización L1 y L2; esto puede deberse a que

múltiples campos se encontraban normalizados en el set de datos original. Finalmente, como el MAE y RMSE menor fue el del modelo que usa el algoritmo XGBoost, se decidió analizar sus resultados.



En la gráfica anterior se evidencia el error que se obtuvo para el modelo XGBoost. Se observa la presencia de outliers, es decir de diferencias que están lejos de los valores normales y que se relaciona con el comportamiento de la variable objetivo. Como se comentó previamente, para los modelos es difícil identificar patrones que le permitan acercarse al valor real de consumo de energía cuando existen tantos valores atípicos.

Considerando que la media de consumo es 51.5, del set de errores marcamos diferencias de 10 kW como los peores casos de predicción, lo cual arrojó como resultado un conjunto de datos del 25% del conjunto de pruebas. Es decir, un cuarto de las predicciones tiene diferencias mayores a 10 kW al compararlas con el valor real. Por tanto, concluimos que el modelo no es fiable ya que el valor estimado puede ser mucho mayor o menor al valor real de consumo, con lo cual un usuario correría el riesgo de no tomar las mejores decisiones para ajustar su consumo a su presupuesto.

4. Despliegue del modelo

El modelo seleccionado se exportó en un archivo con extensión “joblib”, lo que permitió implementarlo en un servidor para hacer uso de este desde una aplicación web con interfaz amigable para el usuario.

4.1. Aplicación Web

Se diseñó e implementó una aplicación web con el fin de proveer la información deseada a un usuario final que desea conocer el consumo energético que tendrá en el futuro. La aplicación permite visualizar los datos usados en el proyecto para ver su comportamiento a través del año y de los días. Asimismo, se pueden consultar las variables climáticas de interés que se presentan cada día y que son usadas para realizar la predicción.

Por otra parte, tras realizar pruebas iniciales a la aplicación desarrollada, se encontró que el servidor usado para la predicción de los valores no tiene los recursos suficientes para hacer las predicciones en tiempo real y enviar los resultados en cuestión de microsegundos a la interfaz web. Por lo tanto, se decidió realizar una predicción de consumo energético de cada uno de los días presentes en el año, generando un archivo CSV que contiene toda la

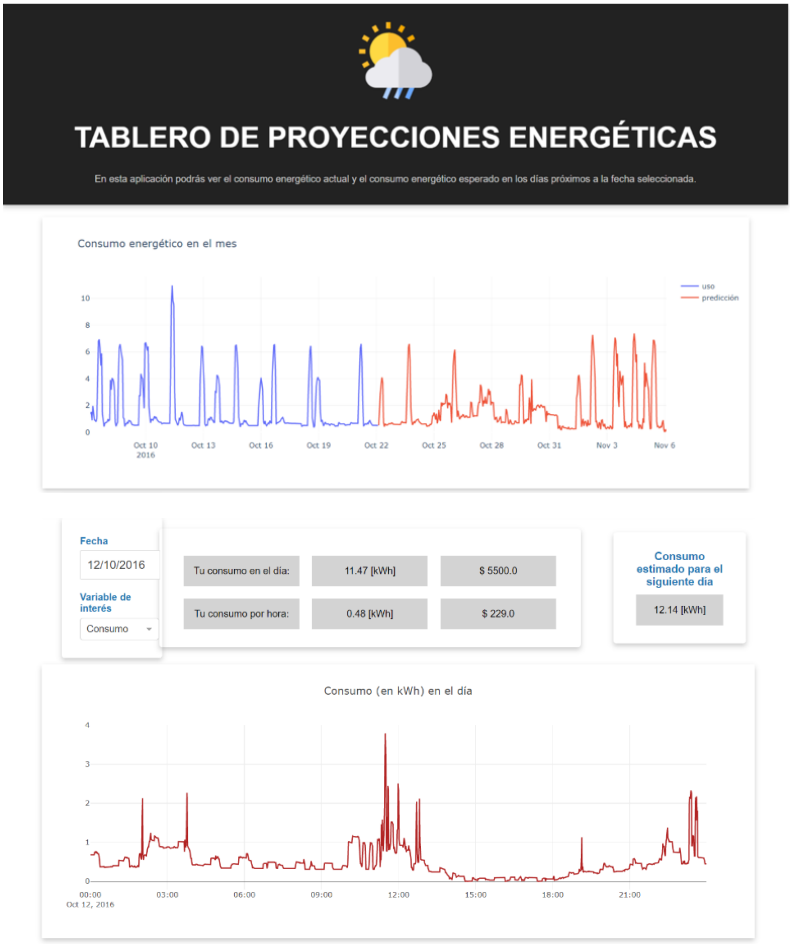
información de las predicciones realizadas por el modelo seleccionado, de forma que se puedan precargar en la página web y así mejorar la experiencia de usuario.

La información mostrada en las gráficas de la aplicación tiene como fin informar al usuario del comportamiento que se observó en el conjunto de datos para la fecha seleccionada, presentando información del año en la gráfica superior (correspondiente al año 2016) e información del día y la variable climática seleccionados. Este enforque descriptivo se realizó teniendo en cuenta los comentarios recibidos en entregas anteriores y del enfoque del trabajo realizado, ya que realizar predicciones de una fecha futura, de un año diferente y de ubicaciones geográficas diferentes está fuera del enfoque de este trabajo.

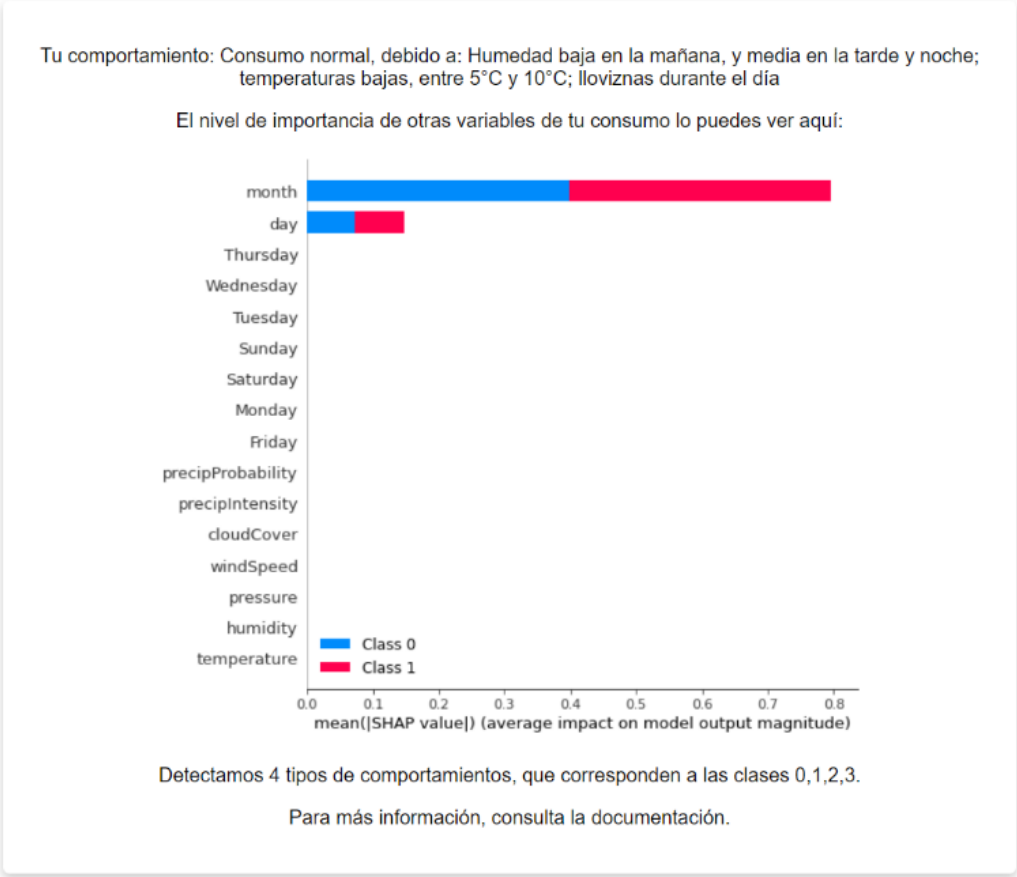
La aplicación web se desplegó en un servidor con el fin de que cualquier usuario tenga acceso a esta: https://energy-forcasting.herokuapp.com

A continuación, realizaremos una breve descripción de dicha aplicación:

- Interfaz gráfica (que se puede consultar mediante cualquier navegador sea móvil o en portátiles) en la que el usuario puede seleccionar una fecha de predicción y una variable de su interés.

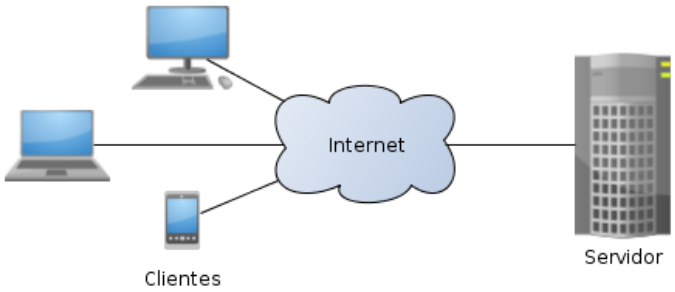


Una vez se seleccione se generará la respectiva predicción de consumo del siguiente día. En una gráfica adicional se evidencian las variables más importantes que influyen en dicha predicción y una descripción del comportamiento energético del usuario de acuerdo a la variable seleccionada.



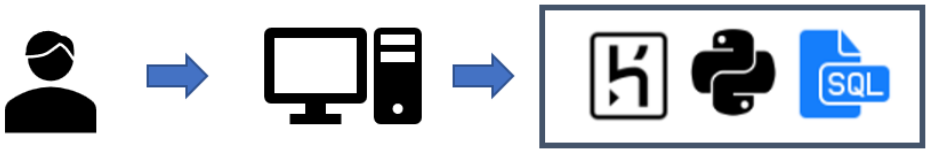
4.2. Arquitectura implementada

Se implementó una arquitectura cliente servidor que permite a los usuarios acceder al recurso desde cualquier dispositivo electrónico que tenga acceso a internet y un navegador web.



La aplicación está realizada usando Dash, el cual es un framework para construir aplicaciones de datos usando Python. Dash está escrito usando Plotly.js y React.js y permite construir y desplegar aplicaciones de datos con interfaces a la medida.

La aplicación construida en este framework actúa como un monolito que permite desplegar la aplicación fácilmente, ya que contiene el código de la página a mostrar y la lógica que debe seguir.



5. Conclusiones

Teniendo en cuenta el objetivo del proyecto, se logró describir el consumo de energía y realizar predicciones del consumo de energía en kilovatios de un hogar. Sin embargo, el resultado que se obtuvo no es del todo aceptable puesto que un 25% de las predicciones se caracterizan por márgenes de error que superan los 10 kW por hora, lo cual se relaciona con el comportamiento atípico que caracteriza a la variable objetivo. Por otro lado, se construyó una interfaz web para que un usuario pueda consultar el consumo real de la vivienda y de los factores ambientales por día, de forma que pueda relacionar su consumo con dichas variables y ver las predicciones propuestas, aunque están sujetas al modelo y RMSE obtenido.

Consideramos que sí cumplimos uno de los objetivos planteados, que es permitir a un usuario analizar su consumo. Pero, no fue posible cumplir con el objetivo de estimar confiablemente su consumo energético a futuro. Esto se debe a que el problema es de tipo serie de tiempo o forecasting, e intentamos resolverlo como un problema de ML estático y no dinámico. Del mismo modo, la variable que pretendíamos estimar (consumo energético) se caracteriza por tener una distribución con muchos outliers, ya que el uso de ciertos electrodomésticos es esporádico.

Otro reto que se presentó está relacionado con que las condiciones de los datos no eran suficientes puesto que el conjunto de datos estaba limitado a registros de un solo año (2016) y para una sola persona. Aunque teníamos más de medio millón de registros (correspondientes al consumo por cada minuto), si intentábamos abordar el problema con frecuencias de datos más cercanas a los intereses reales de los consumidores como mes o día (y de esa manera mejorar la distribución de los datos disminuyendo la presencia de outliers), nos enfrentaríamos al dilema de únicamente contar con 365 o 12 registros, dependiendo de la frecuencia elegida, lo cual no sería suficiente para generar un modelo. Por tanto, optamos por agrupar los datos por hora, para contar con al menos 8.398 registros, sabiendo que el problema de los valores atípicos se mantendría vigente.

Para obtener mejores resultados, la fuente de datos debería contar con más datos históricos de forma que sea posible comparar el consumo energético entre distintos períodos de tiempo,

incluir datos de más personas con sus características demográficas para realizar comparaciones más interesantes y hacer visible otros hábitos de consumo energético, y así tener la posibilidad de generar otros modelos con un público objetivo más amplio como regiones y no apartamentos.

6. Autoevaluación y evaluación grupal:

Aporte por fase del proyecto

	Análisis	Implementación	Evaluación	Documentación
Alejandra Guerrero	30%	33%	30%	40%
Luis Enrique García	40%	33%	30%	30%
Diego Peña	30%	33%	40%	30%

Evaluación (sobre 5)

	Autoevaluación	Evaluación Diego	Evaluación Alejandra	Evaluación Luis
Alejandra Guerrero	5	5	5	5
Luis Enrique García	5	5	5	5
Diego Alejandro Peña	5	5	5	5