Instituto Tecnológico de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos 2

Coordinador Reto: Gildardo Sánchez Ante

Profesor Módulo 1: Alberto de Obeso Oredáin

Profesor Módulo 2: Luis Guillermo Hernández Rojas

Profesor Módulo 3: Rubén Alvarez

Profesor Módulo 4: Adan Octavio Ruiz Martinez

Profesor Módulo 5: Omar Mendoza Montoya

Uso de Imagenes para la Prediccion de Flujo y Descarga de Agua en Ríos

Javier Lizarraga Beyles - A01253105

Natalia Velasco Garcia - A01638047

Jose Luis Rosa Cruz - A01638241

Cesar Ivann Llamas Macias - A01625272

David Alejandro Velázquez Váldez - A01632648

23 de Noviembre, 2022

Contents

List of Figures

List of Tables

1 Introducción

Este trabajo se realizó en colaboración con la Universidad de Nebraska Lincoln, buscando una mejora en el uso de sensores convencionales los cuales miden el nivel y descarga de agua en ríos. Esto es debido a su gran costo, complicada instalación, operación y mantenimiento, lo cual ocasiona inconsistencia o desconfianza en los datos. Se busca la posibilidad de resolverlo con la implementación de modelos predictivos mediante el uso de cámaras para la interpretación de imágenes con inteligencia artificial y ciencia de datos. Se busca utilizar imágenes como una fuente por su fácil acceso económico, además de no necesitar mantenimiento extensivo ni experiencia especializada para su instalación como es el caso de sensores utilizados actualmente. El enfoque es resolver la falta e inconsistencia de datos en los sensores mediante errores. Esto se lograría con una coexistencia de ambas fuentes, necesarias para la medición constante del agua en ríos, lo cual funcionaria para estudios en campos de hidrología en la predicciones de desbordes de los cuerpos de agua estudiados, la construcción de puentes o infraestructura cercana, sistemas de agua y reservas dependientes del suplemento que brindan estos cuerpos de agua, los cuales dependen de un flujo constante y completo de datos para una precisión necesaria para tales trabajos. Todo el trabajo que se presenta se basa en el artículo (Chapman et al., 2022) brindado por los miembros de la universidad de Nebraska hacia nosotros, con nuestro objetivo siendo mejorar el trabajo ya hecho con la búsqueda de alternativas o mejores resultados, validando los recursos usados en esta investigación para encontrar un nuevo valor en estos datos como posible área de oportunidad para un mayor eficiencia en futuras iteraciones. La hipótesis principal es que al dividir los datos por temporadas se podrá entrenar los modelos para sus temporadas seleccionadas con mayor precisión.

2 Descripción de datos

Los datos utilizados son una base de datos obtenida mediante el procesamiento de miles de fotografías tomadas durante intervalos de 1 hora, posicionadas desde el mismo punto fijado en una presa con leves movimientos entre 4^{o} en rotación y 0.25^{o} en traslación. Esto causando un ligero ruido al intentar tomar medidas de las imágenes debido a que fueron tomadas para un documental del área. A su vez se cuenta con medidas de sensores que nos ayudan a medir el nivel de agua y su flujo en intervalos de 15 minutos, contando con datos recopilados desde 2012 a 2019 de ambas fuentes. Estos datos fueron obtenidos del sitio North Platte River State Line Weir localizado entre Wyoming y Nebraska, siendo altamente afectado por el derretimiento de nieve en las cabezas de las montañas, seleccionado por su gran número de imágenes(57,544), la alta resolución que estas poseen(4288x2848 RGB) y su proximidad a la sección de control de rio de la USGS (estación USGS 06674500 NPRSLW, 2020) de donde se obtuvieron los datos de los sensores (State Line Gauge Weir — PBT, 2022). La comparación funciona con calcular la presa que se encuentra en las fotografías (Figure 1), identificando el área por encima de la presa, el área de la espuma blanca que causa el agua al caer y así obtener características que se pueden comparar con los valores de los sensores. A su vez midiendo el área del agua a comparación de la orilla para compararlo con el nivel del agua.

Lo primero que podemos observar es que las muestras clasificadas como agresivas poseen más muestras con valores extremos o irregulares, aunque no pertenecen únicamente a este grupo, también podemos observar que la variable del tiempo no podrá ser utilizada como una variable independiente o una variable que nos ayude a predecir por sí sola un resultado, pues las muestras fueron tomadas en segmentos ya clasificados por uno tras otro, por lo que no aportan mucho valor al modelo, dando la posibilidad de que nuestro modelo sea erróneo al momento de generar conclusiones (e.g. pudiera concluirse que en la mañana todos conducir normal, en medio día empiezan a conducir agresivamente para terminar el

día conduciendo lento) en cambio lo podremos utilizar para comparar cambios bruscos de aceleración y dirección.

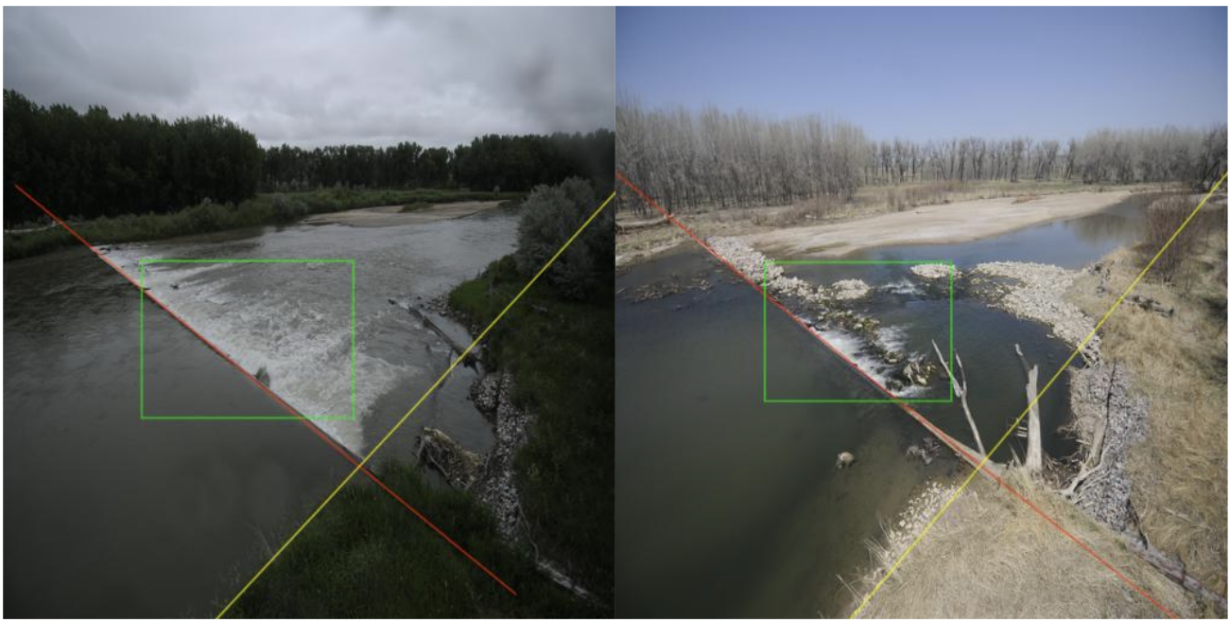

Figure 1: Comparación de dos imágenes del conjunto que se utilizaran en distintas etapas, la línea roja representando la presa, el recuadro verde marcando el área de descarga y la línea amarilla marcando la orilla.

Estos datos ya fueron anteriormente limpiados por nuestros colaboradores en la universidad de Nebraska, removiendo datos faltantes y acoplando los datos de los sensores con sus respectivas fotografías, a su vez se eliminaron las imágenes que no presentaban una vista clara de la presa, impidiendo así la interpretación computacional del nivel y flujo del agua, esto sucedió por las siguientes razones: nieve, hielo o escombros en el sitio, humedad o hielo en el lente de la cámara, nivel de agua elevado, escombros en partes de la presa, imágenes muy oscuras para su correcto procesamiento.

3 Metodologías

Los modelos generados fueron para estimar el nivel y flujo del agua, esperando predecir estas variables en casos futuros sin necesidad de las variables específicas del ambiente, solo enfocándonos en la captura de imágenes para una predicción certera. Con este aclarado podemos decir que los datos con los que tratamos fueron obtenidos por nuestros colaboradores de la universidad de Nebraska, donde se encargaron de procesar las imágenes para obtener los datos con los que se trabajarán.

El primer paso fue recrear los resultados obtenidos por el equipo de Nebraska. Estos se replicaron utilizando los datos que obtuvieron mediante el procesamiento de imágenes y aplicando los modelos seleccionados, la selección de estos modelos será explicada en la siguiente sección. El motivo de este paso fue para comenzar en el mismo punto en el que se concluyó el trabajo pasado, verificando que los datos sean correctos y la situación se pudo replicar exitosamente. Al replicar los resultados se optó por implementar modelos similares de regresión, debido a que el objetivo de estos modelos se enfoca en la predicción de variables, la selección de los modelos será especificada y justificada en la siguiente sección, en este caso el nivel del agua y su flujo. Nuestro enfoque fue encontrar modelos que nos brindaran resultados distintos a los ya obtenidos, buscando variación en los resultados utilizando la precisión como métrica, estos son vistos con mayor detalle en la sección de modelos.

Una alternativa que se buscó fue la especialización de los modelos a temporadas, dividiendo los datos en verano, otoño, invierno y primavera según las fechas que tienen los datos con los que contamos. Gracias a esto se creó un modelo para cada temporada por su cambio tan drástico en el río entre cada una de las estaciones, al mismo tiempo interpretaremos las imágenes fijando el área de donde se enfoca la información obtenida de la base de datos mediante el software de procesamiento de imágenes creado por nuestros compañeros de Nebraska. Este modelo se preparó no generando nuevos datos sino analizando la imagen

para relacionarlo con los valores de los sensores, de esta forma utilizaremos otra fuente de datos nueva, no la información interpretada sino la imagen en sí. Por último se optó por probar una serie de tiempo para la predicción de datos faltantes, por la naturaleza de los datos la cuales cuentan con tiempos específicos.

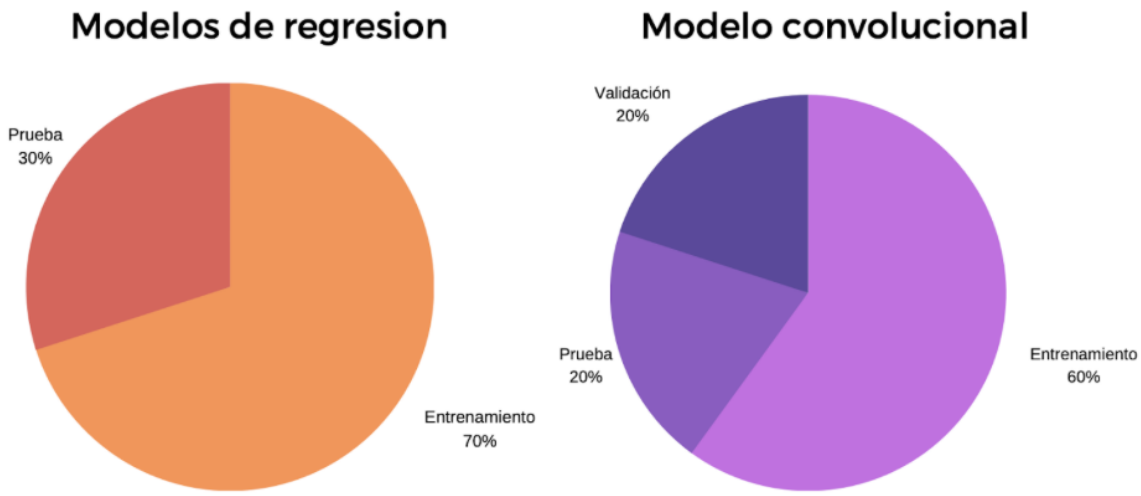

Figure 2: División de variables en los dos casos de modelos utilizados

Debido a las circunstancias distintas de nuestros puntos a seguir, los casos de prueba se organizaron de la siguiente manera (Figure 2), para los modelos de regresión en donde se reutilizaron los datos generados por el equipo de Nebraska, se dividió el 70

4 Modelos

Para recrear los resultados pasados debemos tomar en cuenta que se utilizó el modelo con mejores resultados, se especifica en su artículo (Chapman et al., 2022) que las características principales de sus modelos seleccionados son sostenibilidad para implementarlo con datos de nivel y flujo escalares, disponibilidad en herramientas conocidas de ciencia de datos como Weka (Frank et al., 2016), SciKit Learn (Buitinck et al., 2013), and R (R Core Team, 2016) y por ultimo modelos utilizados en otros casos similares, el modelo que seleccionamos fue:

- Random Forest Regression (RFR)

Ya contando con el modelo pasado se identificaron modelos similares a los anteriores para comparar el desempeño de ambos.

- Modelo de regresión linear
- Modelo de regresión Ridge
- Modelo de regresión Lasso
- Modelo de regresión K-Nearest Neighbor (KNN)
- Modelo de regresión con arbol de decision

Estos modelos se implementaron por su simplicidad, tanto como gran desempeño en trabajos de colinealidad y por la posibilidad de obtener mejores resultados para compararlos.

Cada modelos requería de dos versiones, una que estimara el nivel del agua (stage) y otro que estimara su flujo (discharge), por lo que al dividirlo por estaciones obtuvimos 48 versiones (12 por temporada), lo cual aplicamos a los modelos con mejor precisión.

Al utilizar las imágenes se realizaron dos pruebas, recortar imágenes para mostrar el área del flujo por la presa y el nivel del agua y aplicar un filtro de escalas grises como fuentes (Figure 3) se manejaron otros modelos especializados, ayudando a relacionar las imágenes con sus respectivos valores en los sensores. La esperanza es que este mismo modelo pueda identificar con solo la imagen los valores ya sea del nivel del agua tanto como el del flujo. El modelo específico para imágenes fue:

-Modelo de red convulsionar

Todos los modelos y sus resultados están accesibles en el repositorio del proyecto, anexado en este escrito.

Figure 3: Comparación entre imagen original e imagen utilizada con un recorte y filtro de escala de grises.

5 Resultados

Los primeros resultados obtenidos fueron la recreación de los resultados del equipo de Nebraska con el modelo de RFR como lo indicamos en la sección anterior. Los resultados obtuvieron la precisión del trabajo anterior, por lo que se pudo asegurar que los puntos de comienzo fueron exactos para continuar con una búsqueda de alternativas o mejoras. Los siguientes resultados fueron de los modelos de regresión lineal nuevos. En las series de tiempo no se lograron resultados aceptables, pues en cada caso solo se obtiene un promedio de los demás datos para remplazar los datos faltantes (Figure 4).

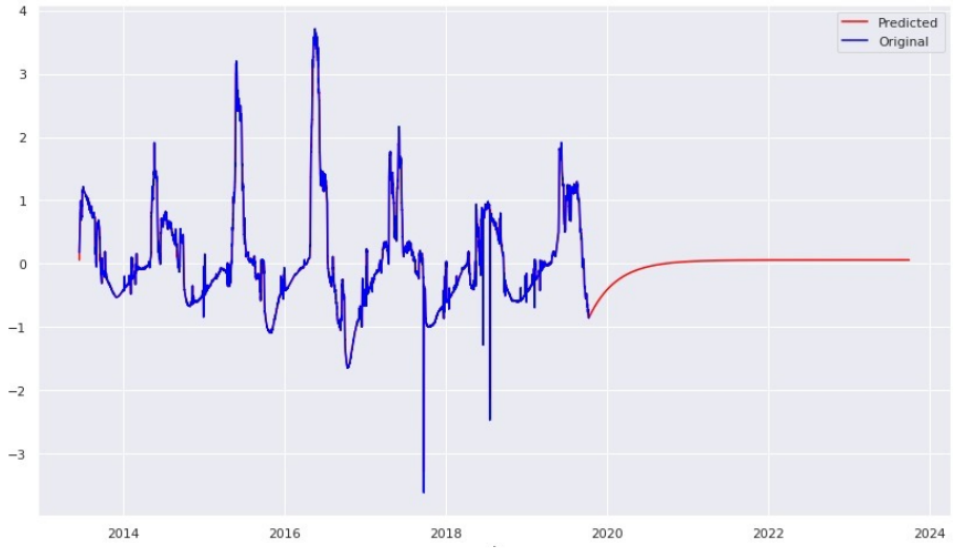

Figure 4: Resultados de predicción del nivel del agua con series de tiempo.

La variable utilizada para medir su rendimiento fue el MSE (Mean Square Error) y MAE (Mean Absolute Error). Estos representan la distancia del valor verdadero y el promedio del valor obtenido al cuadrado midiendo el error que puede obtener un modelo, mientras que el segundo es el promedio de este valor absoluto. Con esto se observa la precisión del modelo en los resultados tras evaluar sus pérdidas.

Por último se obtuvieron los resultados del modelo convolucional evaluado con las mismas variables, gracias a esto se logró una comparación aceptable (Table 1).

	Lineal	Ridge	Lasso	KNN	Decisión	RFR	CNN
MSE	0.784	0.245	0.443	0.195	0.398	0.398	0.143
% MAE	0.537	0.363	0.599	0.281	0.388	0.388	0.307

Table 1: Comparación de resultados entre todos los modelos seleccionados.

Los mejores resultados se obtuvieron en todos los modelos al entrenar los modelos con los datos de una estación específica evaluandolos con esa misma temporada, la única excepción fue el modelo CNN en donde se entrenó con todos los datos y se evaluó por temporada.

Para elegir los mejores resultados se debe tomar en cuenta que aunque KNN muestra mejores resultados de forma general se encontró que estaba sobre entrenado. Esto se observa con que al seleccionar el valor de K el error es muy bajo por lo que no es preciso, mientras que al aumentarlo se dispara el error, dándonos a entender que el error aumentará hasta llegar a un número de K aceptable para el número de datos (Figure 5).

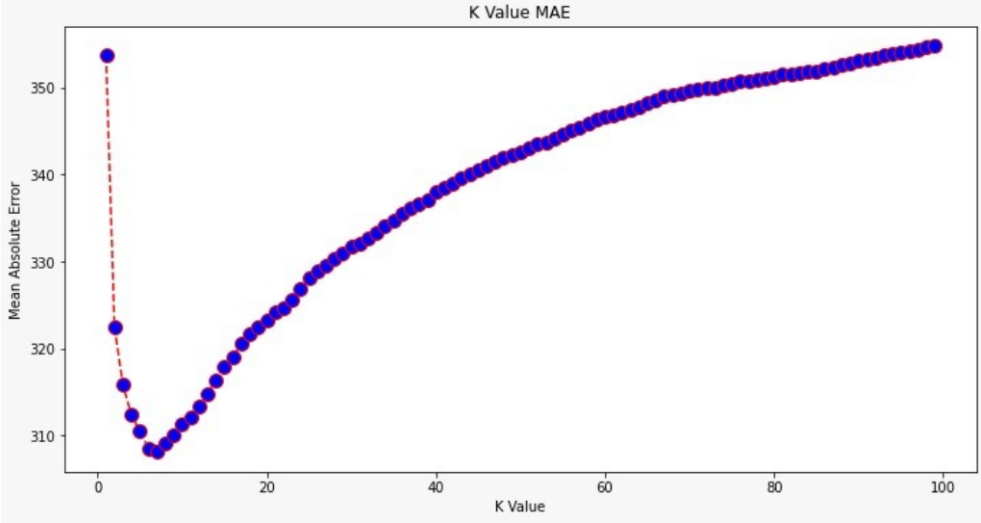

Figure 5: Demostración del valor de K a comparación del error.

Ya con los resultados obtenidos se logró hacer una comparación en donde se encontró que el modelo de Ridge demostró mejores resultados (Table 2), aunque el modelo de RFR

mostró mayor precisión separado por estaciones (Table 3). El modelo CNN no mostraba resultados constantes variando la precisión, aunque en el mejor caso fue el mejor modelo (Table 4), considerando que la fuente de datos fue distinta.

Ridge	Primavera	Verano	Otoño	Invierno
MSE	0.423	0.332	0.475	0.797
% MAE	0.523	0.491	0.572	0.799

Table 2: Resultados del modelo Ridge separado por temporada.

RFR	Primavera	Verano	Otoño	Invierno
MSE	0.176	0.176	0.176	0.175
% MAE	0.290	0.290	0.292	0.291

Table 3: Resultados del modelo RFR separado por temporada.

CNN	Primavera	Verano	Otoño	Invierno
MSE	0.175	0.094	0.148	0.168
% MAE	0.342	0.256	0.272	0.356

Table 4: Resultados del modelo CNN separado por temporada.

6 Conclusion

Durante el trabajo se consideraron varias alternativas, dando la oportunidad de estudiar los datos a un fondo sorprendente, e impulsando a experimentar con soluciones creativas e innovadoras. La solución que mejores resultados fue tomar los datos y dividirlos por temporadas (primavera, verano, otoño e invierno), especializando un modelo para cada temporada, a su vez se implementó una solución que solo necesitaría imágenes con un leve procesamiento sin necesidad de un análisis complejo de las imágenes. Aunque el anterior muestra inconsistencia de resultados precisos se considera que fue la solución más creativa debido a que una vez entrenado podrá funcionar sin necesidad de alteraciones al sitio, implementando la solución de la hipótesis que creamos desde un inicio. Para concluir confirmamos la hipótesis inicial en la cual se creía que un modelo entrenado para temporadas específicas demostraría mejores resultados. Con esto aclarado los resultados con mayor precisión y consistencia de buenos resultados fue el modelo de RFR aplicado desde un comienzo por el equipo de Nebraska con la alteración de datos y segmentación de conjunto de entrenamiento y prueba.

7 Referencias

- Chapman, K. W., Gilmore, T. E., Chapman, C. D., Mehrubeoglu, M. Mittelstet, A. R. (2022). Camera-based Water Stage and Discharge Prediction with Machine Learning. University of Nebraska Lincoln.

- State Line Gauge Weir — PBT. (2022). https://plattebasintimelapse.com/explore/galleries/state-line-gauge-weir/

- Frank, E., Hall, M. A., and Witten, I. H.: The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, fourth edn., $https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016-appendix.pdf$, 2016.

- Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122, 2013.

- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2016.

- $GitHub - DavidVelazquez584/AI_WaterStageDischarge$. (s. f.). GitHub. $https://github.com/DavidVelazquez584/AI_WaterStageDischarge$