Universidad de la República

Facultad de Ingeniería

Maestría en Ciencia de Datos y Aprendizaje Automático

Introducción a la Ciencia de Datos 2023

Tarea Final. Presentación de un Anteproyecto.

Grupo 1:
Ignacio Corrales

Julio de 2023.

Resumen.

El proyecto planteado refiere a un abordaje de aprendizaje automático para el problema de pronósticos de caudales hidrológicos futuros, en un punto de un río. Usando datos de niveles observados en distintos puntos de un curso hídrico, precipitaciones ocurridas en la cuenca y pronósticos de precipitaciones futuras sobre la misma, analizar la capacidad de ajuste de distintos modelos para la predicción del caudal futuro.

Motivación.

La Represa Hidroeléctrica de Salto Grande, instalada en el curso del río Uruguay, además de ser de vital importancia para la República Argentina, es para Uruguay el principal actor en el concierto del mercado eléctrico.
Desde su completa puesta en funcionamiento, en 1983, ha generado históricamente entre el 55% y 70% de la energía anual consumida por los hogares e industrias del Uruguay. Desde la irrupción de la energía eólica en el sistema interconectado, esos guarismos han bajado para ubicarse en un 40%, lo cual sigue siendo significativo, y a su vez, han aumentado su importancia en cuanto a la regulación de frecuencia del sistema, debido a que los aerogeneradores carecen de dicha capacidad.
El punto de partida del proceso de producción de una represa hidroeléctrica, es el Pronóstico Hidrológico, el cual consiste en pronosticar los caudales de agua que llegarán al embalse en un futuro cercano. Un pronóstico de buena calidad permite, por un lado, optimizar el recurso hídrico y la gestión del embalse, brindando un mayor potencial de generación eléctrica; y por otro, operar la represa dentro de los parámetros de seguridad tanto para la obra civil de la represa en sí, como para las poblaciones ribereñas instaladas a ambas márgenes del río, permitiendo incluso, mitigar inundaciones y reduciendo la cantidad de hogares y familias afectadas en eventos de crecidas.
Para realizar la tarea de pronóstico, Salto Grande se sirve de los datos reportados por su red telemétrica de más de setenta estaciones, instaladas en distintos puntos de la cuenca del río Uruguay. Existen estaciones meteorológicas, que reportan únicamente datos de lluvias ocurridas y estaciones hidrometeorológicas, las cuales se instalan en la orilla del río o alguno de sus afluentes principales y además de reportar lluvias, reporta niveles del curso de agua.
Los modelos matemáticos que se usan para el cálculo del pronóstico, tienen base en los fenómenos físicos involucrados en el proceso hidrodinámico e hidráulico. En concreto, en Salto Grande, el pronóstico se hace en 2 pasos:

1- Modelo de transformación de precipitación a caudal (utilizando hidrograma unitario).
2- Modelo de propagación o enrutamiento.

En el primer paso se usa un modelo que ajusta parámetros que modelan la física que rige el escurrimiento de la lluvia ocurrida hacia los cursos de agua y las características del terreno, como ser la cobertura del suelo en cada punto de la cuenca, el estado de humedad previa del suelo, etc.
Para el segundo paso, se usa el modelo de enrutamiento Muskingum-Cunge, que se sirve de los caudales calculados en el paso previo y modela su propagación por el cauce del río hasta su llegada al embalse.

Objetivo.

El objetivo del trabajo planteado, es comprobar la factibilidad de incorporar técnicas de aprendizaje automático, sustituyendo o complementando los modelos físico-estadísticos utilizados a la fecha, con modelos como los vistos en el curso, para la generación de pronósticos hidrológicos.

Datos de entrada.

A continuación, se enumeran los datos de entrada con los que se cuenta, sus características y frecuencia.

Precipitación observada.

Se tiene en una base de datos SQL Server una tabla donde se registran, para cada dupla (ID Estación, Fecha-Hora) la precipitación ocurrida en los 15 minutos previos a la Fecha-Hora indicada, representado como un entero que indican los milímetros medidos por la estación.
En esa tabla se registran valores de 72 estaciones distintas, de las cuales se sabe su ubicación geográfica (latitud, longitud).

Rango de datos: mayo de 2013 a la fecha.
Frecuencia de datos: datos cada 15 minutos.

Niveles observados.

De manera análoga a la precipitación, se tiene en el mismo SQL Server, el valor cada 15 minutos del nivel medido reportado por las 28 estaciones hidrometeorológicas a distintas alturas del río, aguas arriba.

Precipitaciones previstas.

Se propone usar un producto generado por Salto Grande, que, basado en el modelo ECMWF de pronósticos de precipitaciones, genera un pronóstico de precipitaciones promedio por subcuenca en pasos de tres horas.

De las 70 subcuencas del río Uruguay, existen 38 que son de interés, por estás aguas arriba de la represa. Son las que a continuación se ilustran en azul.

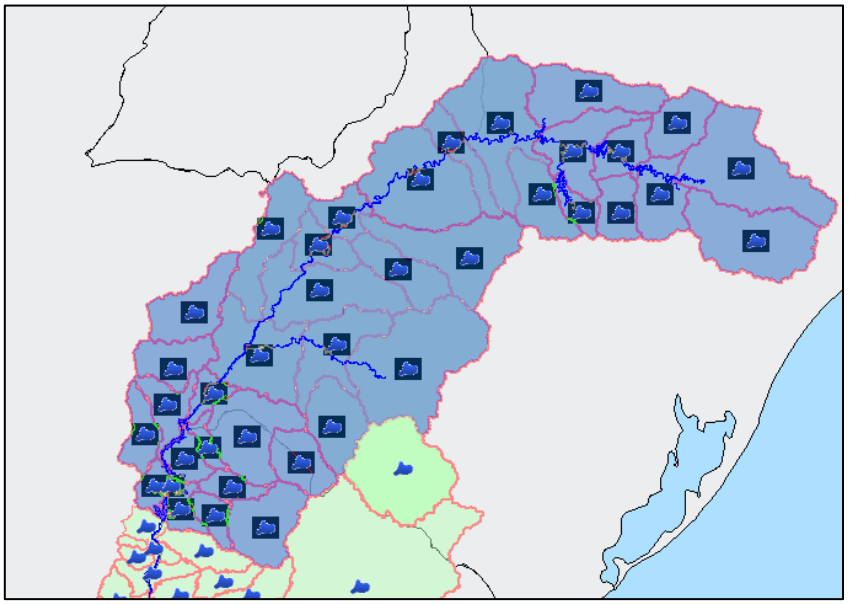

Ilustración 1: Subcuencas de interés

Los datos con los que se cuentan, almacenados en una Base de datos SQL Server, tienen una fecha de elaboración del pronóstico y para cada una de las subcuencas, las precipitaciones que ocurrirán a futuro, cada tres horas por los próximos 15 días.
Por conocimiento de la cuenca del río Uruguay, se propone usar únicamente los primeros 7 días del pronóstico.

Rango de datos: enero de 2014 a la fecha.
Frecuencia de datos: datos cada 12 horas.

Datos de salida.

Pronóstico de caudales.

La salida de los modelos actuales, es una serie de caudales que se pronostica que ingresen al embalse, en pasos de tres horas, por los próximos 7 días (un vector de 56 entradas).
Para este trabajo, se propone una salida con paso diario (caudal promedio diario, o sea 7 entradas).
Para validación, se tiene la serie histórica de caudales observados.

Plan de Proyecto.

1. Tratamiento de los datos de entrada y preprocesamiento.

El primer paso es mitigar los problemas de calidad existentes en el juego de datos.
Tanto en los datos de precipitación, como en los de niveles, se observa una serie de problemas:

- **Datos faltantes:** El crecimiento de la red telemétrica, dado por la sucesiva incorporación de estaciones a lo largo del tiempo, hace que muchas de las estaciones que hoy reportan, no reportaran al inicio del período. A su vez, problemas en las estaciones o con sus instrumentos, fallos en la alimentación o derivados de los canales de comunicación, hacen que existan baches de datos para muchas de las estaciones.
- **Datos fuera de rango (outliers):** Existe datos reportadas por las estaciones cuyos valores no se corresponden con datos reales. Precipitaciones o niveles menores a 0 o mayores a cierto valor máximo esperable para un lapso de 15 minutos.
- **Datos espacialmente incongruentes:** Existen datos que no son correctos, pero que resulta algo más difícil de detectar, ya que sus valores están dentro de los umbrales considerados como normales. Sin embargo, el análisis del contexto espacial, pueden delatar su incongruencia. Por ejemplo, en un evento de precipitación intensa, donde todas las estaciones cercanas a la analizada están reportando lluvias intensas, pero esta estación reporta valores 0.
- **Datos temporalmente incongruentes:** Similarmente al caso anterior, si una estación viene reportando niveles similares cada 15 minutos, un valor muy disímil en el medio, puede tratarse de un dato no confiable.

El primer paso propuesto, es la detección y eliminación de estos datos *outliers*.
Luego, se requiere generar datos para los puntos y fechas en donde no hay o en donde hemos eliminado valores *outliers* previamente. Para ello se sugieren distintas aproximaciones, dependiendo del caso.
Para precipitaciones, en el caso donde falten pocos datos puntuales para una estación en un período acotado de tiempo, se emularán los datos faltantes promediando, por ejemplo, los 3 valores anteriores y los 3 posteriores al instante de tiempo faltante registrados por esa estación.

Para el caso en que faltan datos por períodos prolongados de tiempo, se procederá a generar los datos como una interpolación de los datos registrados por las 4 estaciones más cercanas.

Para los niveles, los datos faltantes (o eliminados por outliers), se sustituirán interpolando los últimos datos registrados antes de la brecha y los inmediatamente posteriores.

El siguiente paso, será acumular (sumar) las precipitaciones en pasos de 6 horas y promediar los niveles en igual período.

Realizado este paso, nos quedaremos con datos cada 6 horas o, lo que es lo mismo, 4 muestreos diario por variable.

2. Juego de datos.

Cada uno de los datos, será entonces la representación de un estado hidrológio de la cuenca en un instante de tiempo y, a partir de ellos, se pretende inferir un vector de dimensión 7 que representa los caudales diarios promedio que se esperan en los próximos 7 días.
Para representar ese estado hidrológico, se usarán las precipitaciones ocurridas en los últimos 7 días previos a ese instante, en cada uno de los 72 puntos de medición, con paso de 6 horas, más los niveles registrados en los últimos 7 días en cada uno de los 28 puntos de medición, también cada 6 horas y las precipitaciones pronosticadas en cada una de las 38 subcuencas, por los próximos 7 días, también acumuladas cada 6 horas.
En otras palabras, las *features* o características que se tendrán serán:

- Por cada estación meteorológica (72), los últimos 28 valores de precipitación registrados (7 días por 4 mediciones diarias) ordenados cronológicamente.
- Por cada estación hidrométrica (28), los últimos 28 niveles registrados (7 días por 4 mediciones diarias), ordenados cronológicamente.
- Por cada subcuenca (38), 28 valores con las precipitaciones esperadas en los próximos 7 días (4 valores por día), ordenados cronológicamente.

Esto determina la existencia de unas 3.864 *features* para caracterizar cada uno de los datos.

Como la frecuencia de datos original es de 15 minutos, el set de datos que se podrá generar desfasando las acumulaciones en 15 minutos, es grande (9 años de datos x 365 x 24 x 4 = 315.360 datos).

Para la separación de los datos de entrenamiento y de pruebas (en una proporción 70-30), se deberá tener en cuenta los picos de creciente ocurridos con poca frecuencia, de manera que, mediante una estratificación supervisada, haya una representación similar en ambos conjuntos de lo que es un estado de río “normal” y lo que refiere a estados de crecientes o atípicos.
De hecho, se puede incluso sopesar la posibilidad de entrenar dos modelos diferentes para ambos regímenes de río, dada la diferencia sustancial en cantidad de datos de uno y otro tipo en el conjunto de datos y las diferencias marcadas del comportamiento hidrológico en una y otra situación.
Otra opción a explorar es desfasar las acumulaciones de los datos de régimen normal cada 3 horas y desfasar los datos de creciente en 15 minutos, de manera de generar una mayor cantidad de datos de este estilo para mitigar el sesgo en el modelo entrenado.

3. Selección de características.

Mediante un análisis de la varianza explicada que proporciona cada una de las *features* arriba descritas, se podrá seleccionar un subconjunto de las mismas que disminuya la dimensión de nuestros vectores de datos y así aumentar la eficiencia de los modelos.
Otra aproximación posible para lograr el objetivo de disminuir la complejidad del universo de estudio, es la eliminación recursiva de características hasta alcanzar un criterio de parada determinado.
También sería válido mantener todas las características y dejar que el modelo de aprendizaje automático, como ser una red neuronal, elimine en su primera capa, aquellas características que aportan menos información.

4. Selección de modelo.

Se plantea comparar varios modelos y con varias configuraciones de hiperparámetros para

evaluar cual configuración tiene un mejor ajuste.
Los modelos que se plantean estudiar son los siguientes:

- **Redes Neuronales Recurrentes (RNN):** Se desea evaluar si una arquitectura como la Long Short-Term Memory, puede aprender las dependencias temporarias que existen en los datos planteados.
- **Redes Neuronales Convolucionales (CNN):** Se desea comprobar la capacidad de una arquitectura de este estilo, para aprender los patrones espaciales que también caracterizan a los datos de entrada.
- **Una combinación de las anteriores:** Combinar RNN y CNN para capturar tanto las dependencias temporales como las características espaciales en los datos.

A su vez, cada uno de ellos se evaluará con distinta configuración de los siguientes hiperparámetros:

- Cantidad de capas.
- Cantidad de neuronas por capa.
- Función de activación.
- Learning Rate.
- Batch Size.
- Number of epochs.
- Técnica de regularización.

Para cada modelo, se realizará la validación cruzada (con una cantidad de pliegues a determinar) y se registrará el rendimiento obtenido en cada pliegue con el error cuadrático medio como métrica (suma de los cuadrados de las diferencias en cada una de las 7 entrada de los vectores *Yr* e *Yp*, siendo *Yr* el dato real e *Yp* el pronóstico generado por el modelo). Luego se calculará la media y la desviación estándar de estos valores para obtener una estimación del rendimiento promedio y la variabilidad del modelo en diferentes divisiones de datos.
Con una visualización acorde, por ejemplo, una gráfica de violín, se procederá a seleccionar el modelo a utilizar.

5. Entrenamiento y validación.

Con el modelo y sus parámetros elegidos, se entrenará el modelo con todo el conjunto de entrenamiento y se analizará su rendimiento en el conjunto reservado para test, registrándose el error cuadrático medio.

6. Comparación con el modelo actual

El último paso propuesto para este proyecto, es el de tomar los pronósticos producidos por el modelo utilizado hoy en día y aplicarles las normalizaciones utilizadas de acumulaciones, promedios y demás, de forma de llevarlos a pronósticos de caudal medio diario para los siguientes 7 días, tal como las salidas del modelo de aprendizaje automático que se ha entrenado en los pasos previos del proyecto.
Una vez realizado esto, se podrá calcular el error cuadrático medio de este modelo sobre el conjunto de test y comparar ambas metodologías.
Esta comparación, de arrojar guarismos optimistas, puede ser el punto de partida de un nuevo proyecto que profundice sobre estas técnicas, dedicando mayor tiempo y recursos para éste.