Por: Cristián Ismael Schaad Concha
Master en Ingeniería Mecánica, UCHILE
https://github.com/Cris-Schaad

Spike Challenge:

Predicción de caudales extremos en Chile

Análisis preliminar del dataset

Antes de comenzar con el análisis, el dataset es ordenado, primero por nombre de estación y luego, para cada una de estas, por fecha de medición. Se encuentran en este dataset 133 estaciones de medición distintas, que en conjunto acumulan un total de 1.411.180 instancias de mediciones de caudal. Existen significativas diferencias entre las cantidades de datos registrados en las estaciones; teniendo sólo 802 mediciones la estación con menor cantidad de registros mientras que 20.706 mediciones tiene la de mayor. En promedio se tienen más de 10.000 datos por estación.

Este dataset abarca 58 años de registros de caudal, desde el 2 de Enero de 1960 hasta el 9 de Marzo del 2018. En la **figura 1** se muestra la distribución de datos por años. Se observa que el número de datos ha aumentado con el tiempo, por lo que los datos más recientes estarán sobrerrepresentados en comparación con los datos más antiguos. Se observa además una significativa disminución para el año 2017, para los que se podría deducir que aún no han sido entregados o no están disponibles.

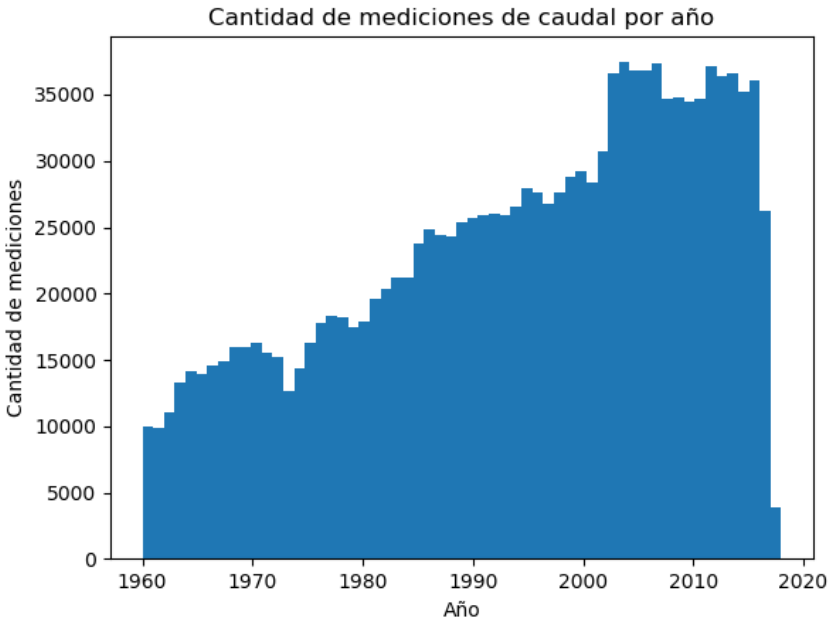


Figura 1

Por otro lado, para el 99.59% del dataset, el tiempo entre datos consecutivos, dentro de una misma estación de medida, es de 24 horas. Para el 0.41% restante (5802 instancias), el tiempo entre registros consecutivos es igual o superior a los 2 días, para los que se puede infiere que la estación de medición no se encontraba en operación en dicho intervalo.

Para los datos de caudal, se cuenta con la totalidad de los registros (1.411.180 mediciones), pero para las precipitaciones promedio hay 27.767 instancias sin datos (2% del total) y para la temperatura máxima promedio hay 151.563 instancias sin datos (10.7% del total). Estas instancias sin registros quedarán fuera de cualquier uso posible en el modelo de predicción. En la **figura 2** se muestra la distribución de estos datos faltantes por cada año.

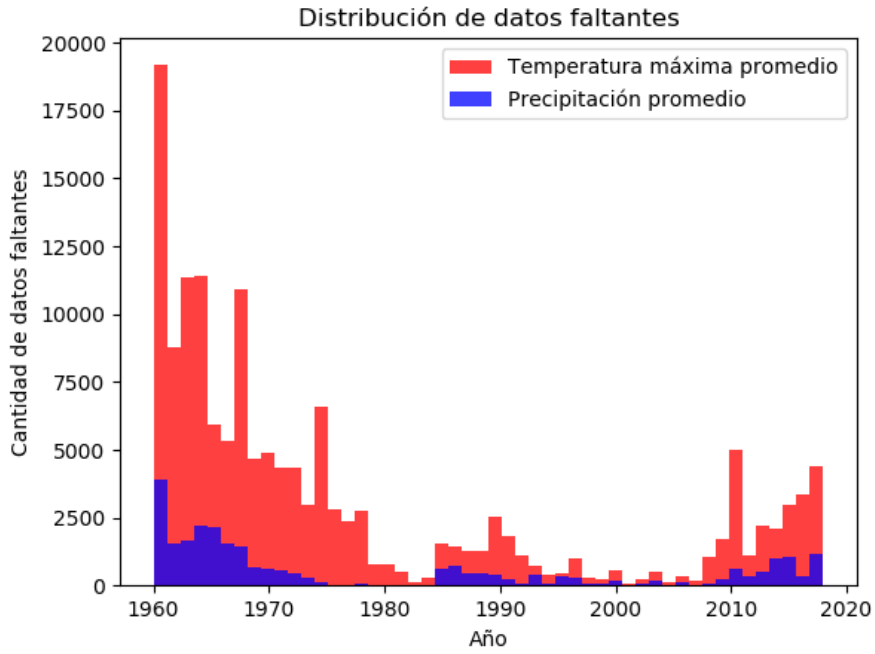


Figura 2

Se ha solicitado además una función (`time_plot_una_estacion`) que grafique 1 de las variables del dataset según el código de la estación, entre un intervalo de fechas arbitrario. En la **figura 3** se muestra como ejemplo el gráfico de la temperatura máxima promedio diaria para la estación "*Rio Ancoa En El Morro*" (código 7355002) entre el 01 de Enero del 2000 y el 1 de Enero del 2005. Se pueden observar claramente la periocidad de esta variable a lo largo del tiempo.

Además, se ha solicitado otra función (`time_plot_estaciones_varias_columnas`) que grafique 1 o más de las variables presentes en el dataset a la vez según el código de la estación entre un intervalo de fechas arbitrario. Dado que las variables están a diferentes escalas, estos datos se normalizan al dividir cada serie de datos por el máximo valor de dicha serie. En la **figura 4** se muestra como ejemplo el gráfico de caudal, temperatura máxima promedio y precipitación promedio diaria para la estación "*Rio Loa En Salida Embalse Conchi*" (código 2104002) entre el 01 de Enero del 2010 y el 1 de Enero del 2015. Nuevamente se pueden observar claramente la periocidad de las variables a lo largo de los años. También se puede observar un salto en la temperatura entre los 2013 y 2014, que corresponde a una situación en la que no se contó con registro alguno de dicha variable.

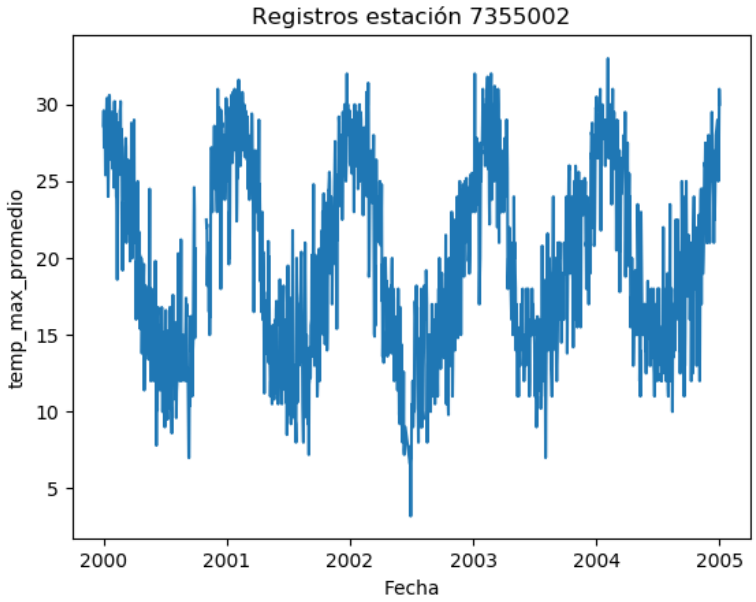


Figura 3

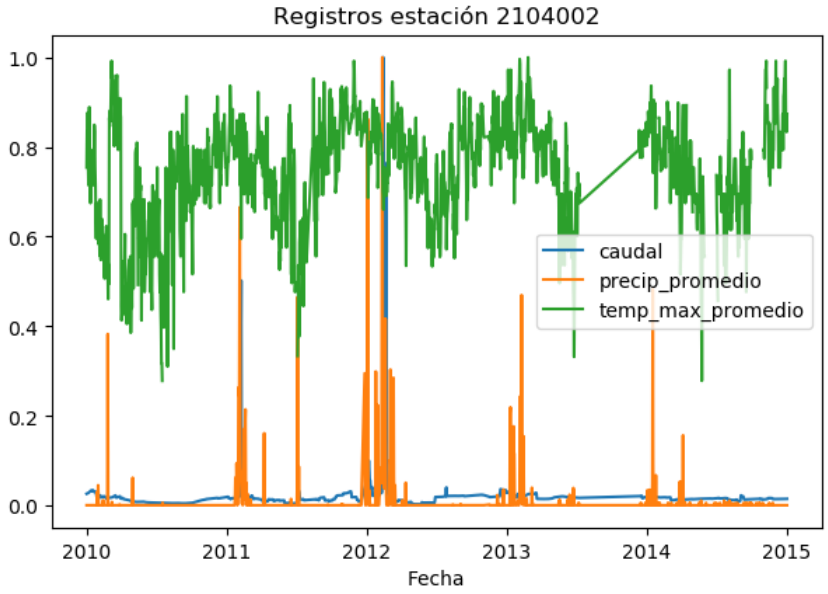


Figura 4

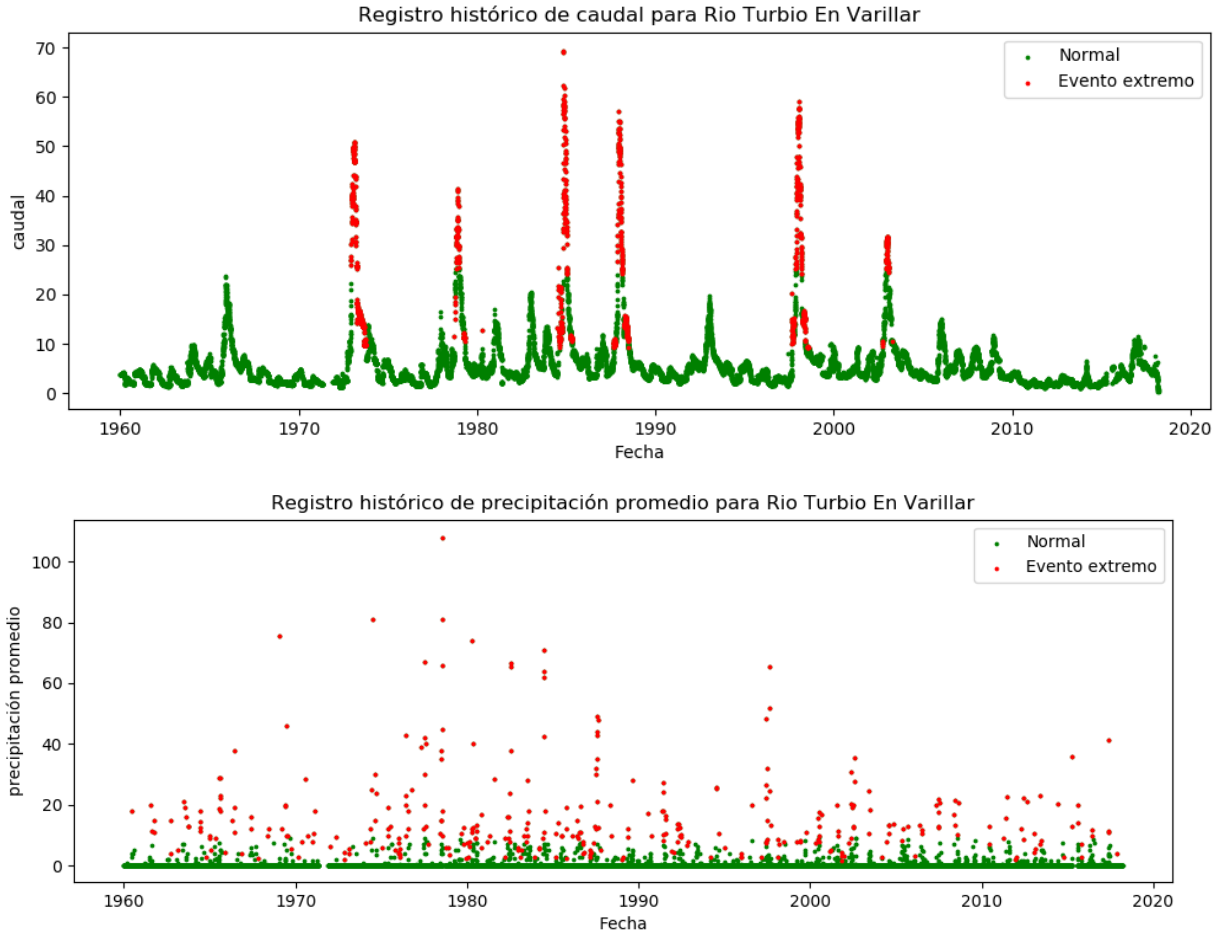
Definición de evento extremo

Como ha sido indicado, los eventos extremos han sido definidos como los datos que se encuentren fuera del 95% de la distribución normal para dicha variable, considerando la distribución de esta variable de manera independiente para cada una de las estaciones de medición y para cada una de las estaciones del año: verano, otoño, invierno y primavera. Esto

último debido a que un evento extremo en cierta estación puede ser considerado como normal para otra.

Para continuar con este análisis, primeramente se generan 4 nuevas variables: "`Verano`", "`Otoño`", "`Invierno`" y "`Primavera`", cada una de estas tomará el valor de 1 si el dato corresponde a una de esas estaciones, y 0 si no. Por simplicidad, se considera que verano corresponde a los meses de Enero, Febrero y Marzo (por completo), otoño a los meses de Abril, Mayo y Junio, y de igual manera para el resto de las estaciones. Luego se crean otras 3 variables: "`caudal_extremo`", "`temp_extremo`" y "`precip_extremo`", que tomarán el valor de 1 si la variable "`caudal`", "`temp_max_promedio`" o "`precip_promedio`" (respectivamente) están fuera del intervalo de confianza del 95%, y 0 si están dentro (es decir, tienen un valor que se considera como normal). Esta medida suena razonable a priori, pero se está imponiendo que el 5% de los registros sean anómalos, cuando la proporción de eventos extremos no es realmente conocida. Una manera de determinar eventos extremos podría ser utilizando un detector de anomalías, basado en la reconstrucción de la data (con un red de *autoencoders*), que determine de forma automática los eventos que no son de normal ocurrencia.

En la **figura 5** se muestran los gráficos de las variables de caudal, temperatura máxima y precipitación (por hora) según su clasificación de evento normal o evento extremo para la estación "*Rio Turbio en Varillar*" durante todo el registro que se tiene de ella. Se puede apreciar a simple vista que los eventos extremos son, en efecto, los datos que más se alejan de la media.



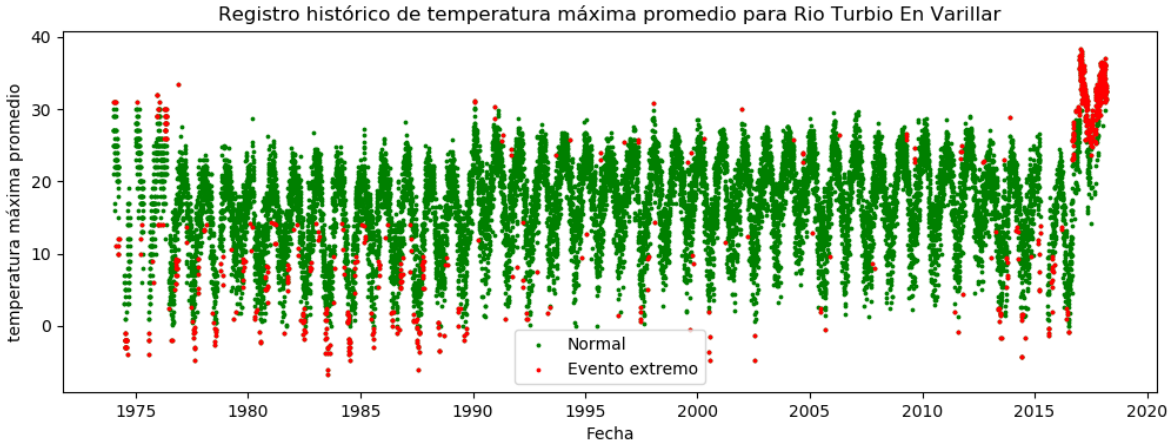
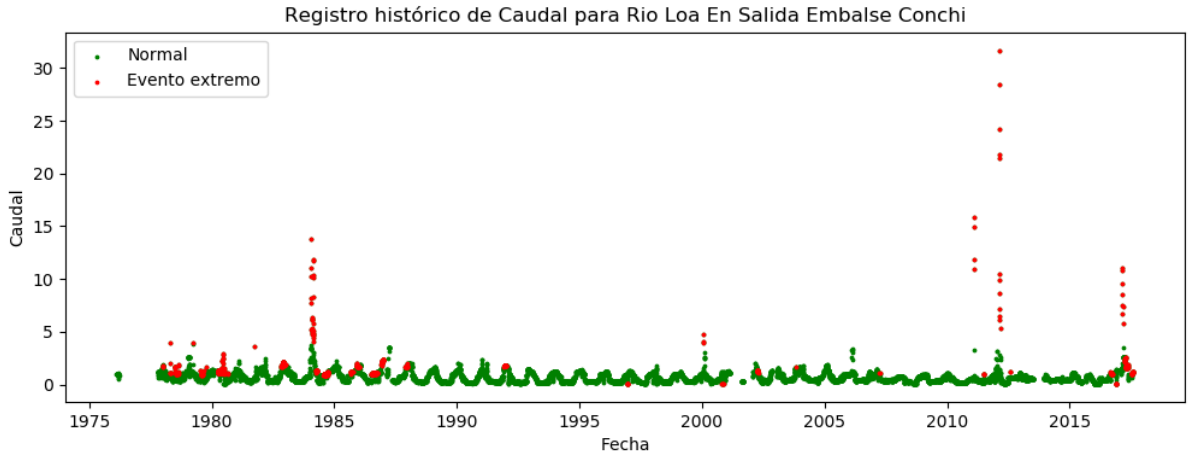
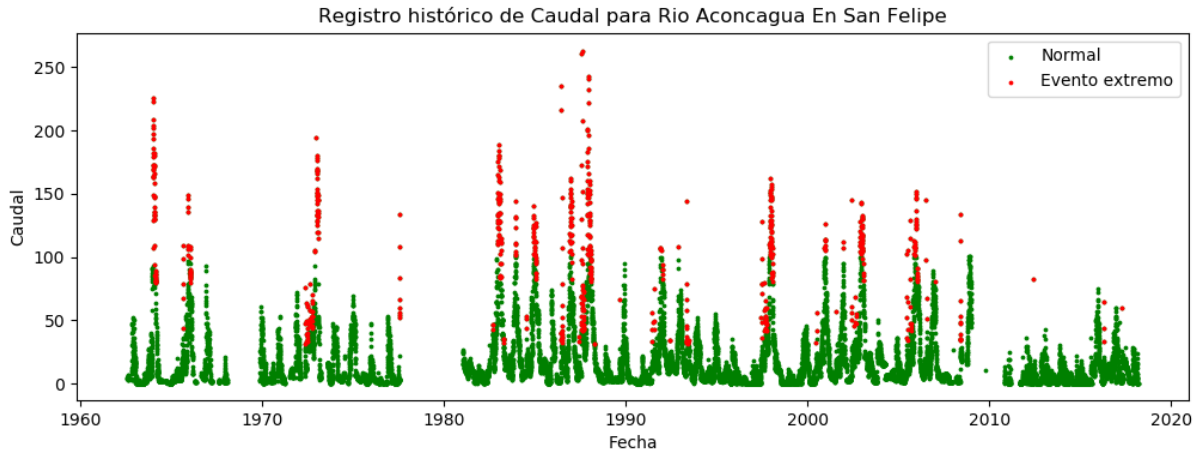


Figura 5

En los gráficos de la **figura 6** se muestran los registros de caudal histórico en 3 ríos ubicados en distintas zonas geográficas de Chile; el Río Loa en el Norte de Chile, el Río Aconcagua en el centro y el Río Aysen en la Patagonia. Como se puede apreciar en dichos gráficos, el comportamiento entre estos ríos es bastante disimilar entre sí, dado que la frecuencia de ocurrencia de eventos extremos aumenta significativamente a medida que se avanza hacia el sur.





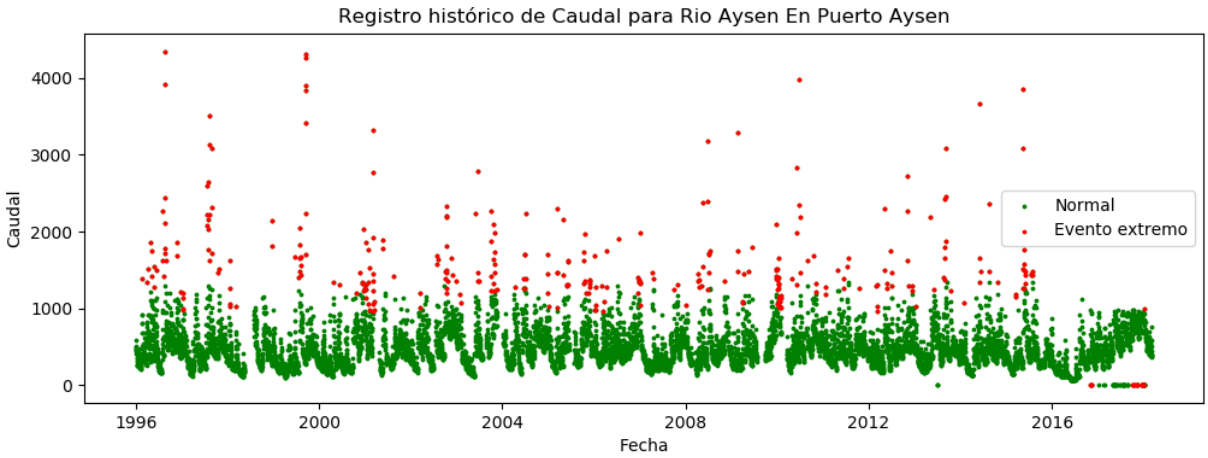
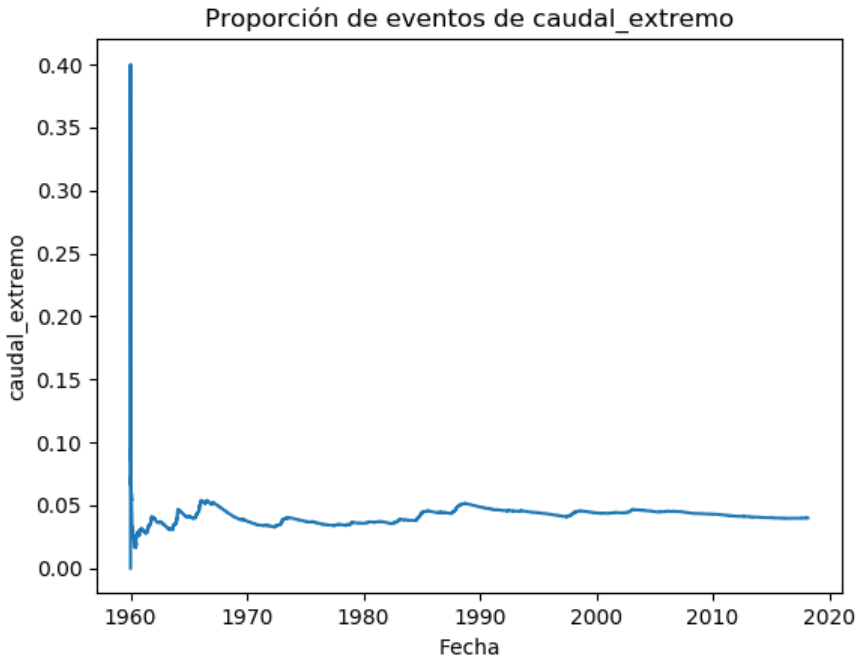


Figura 6

Además, en los gráficos de la **figura 7** se muestra la proporción histórica de eventos extremos, de todo el dataset, para el caudal, precipitaciones y temperatura máxima promedio. Según estos, **NO** se puede concluir que la proporción de eventos extremos haya incrementado con los años. Es más, estos muestran tener una proporción que se ha mantenido constante; de 5% para los caudales, 3% para las precipitaciones y 5% para las temperaturas. Esta proporción era esperable dado que se usó como definición que los datos extremos están fuera del 95% de la distribución normal, por lo que el restante 5% debiese ser extremo.



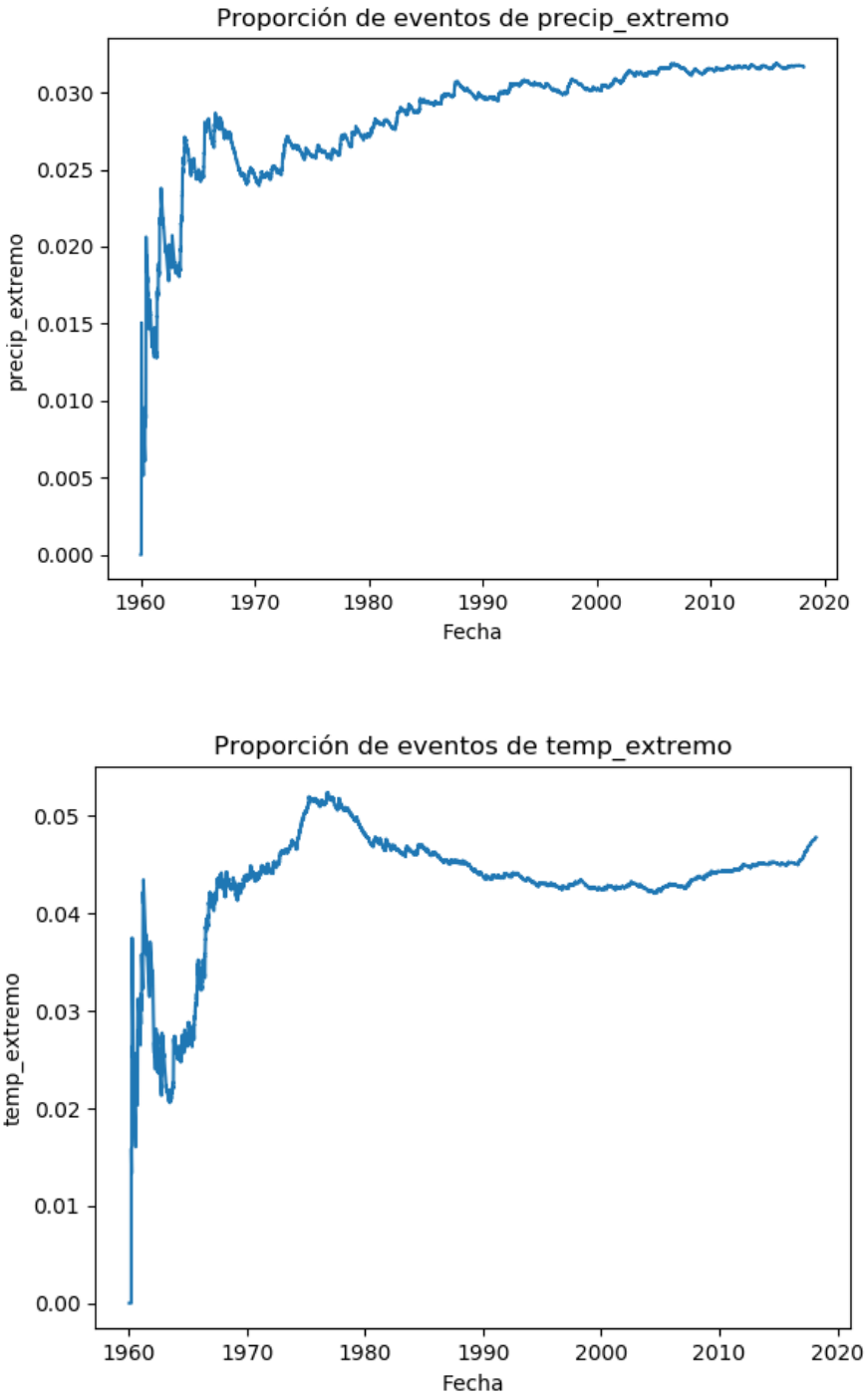

Proporción de eventos de precip_extremo
precip_extremo
0.030
0.025
0.020
0.015
0.010
0.005
0.000
1960
1970
1980
1990
2000
2010
2020
Fecha
Proporción de eventos de temp_extremo
temp_extremo
0.05
0.04
0.03
0.02
0.01
0.00
1960
1970
1980
1990
2000
2010
2020
Fecha

Figura 7

Predicción de eventos extremos con Red Neuronal Recurrente

Para la predicción de eventos extremos (sólo para caudales) se utilizará un algoritmo de **redes neuronales recurrentes (RNN),** específicamente una red basada en la arquitectura **LSTM** (*Long-Short Term Memory*). Este tipo de algoritmos se caracteriza por aprender de manera autónoma los parámetros necesarios para representar los datos suministrados.

La red en total será bastante simple; tendrá 1 capa de LSTM de 64 neuronas, de la que se tomará el último output temporal y será dado a una red neuronal típica de 64 neuronas con función de activación sigmoide. De esta última red se obtiene la predicción, es decir, si el evento futuro a predecir estará dentro de los valores normales o será extremo (fuera del 95% de la distribución normal de datos). Debido este tipo de predicción, este problema se considera como un problema de clasificación con 2 clases. Por otro lado, dado que las clases están evidentemente desbalanceadas, las clases con menor cantidad de datos (las de eventos extremos) son ponderadas por la inversa de su proporción en el dataset, para aumentar su significancia y lograr un mejor entrenamiento.

Los datos que se suministrarán a la red deben ser restructurados a un formato acorde para la red; deben ser entregados en series de tiempo. Por ello, estos segmentan en series de tiempo contengan 24 horas de información, es decir, series de tiempo con 24 instancias temporales de 1 hora. Las variables a utilizar serán 4: “caudal”, “precip_promedio”, “temp_max_promedio” y “dia del año”, que toma un valor del 1 al 365 (o, 366 si es bisiesto), ya que se considera que esta variable podría dar mejor precisión con respecto a la estacionalidad del registro. Con estos series de tiempo, formalmente **el objetivo de la red LSTM será la de predecir la categoría del evento a 24 horas en el futuro, considerando las últimas 24 horas de registros.** La única restricción para esta red, es que los datos tengan un valor numérico (y no NaN). Si no se cuenta con el registro durante las últimas 24 horas de una de las variables mencionadas, no se podrá usar en la red.

Para entrenar la red, el dataset es sub-dividido en un *training set* y un *test set.* El primero corresponde a todos los datos del 80% de las 133 estaciones de medición, y el *test set* al 20% restante. Se debe destacar que estos conjuntos son totalmente disyuntivos, no hay datos de las estaciones de medición que estén en ambos sets a la vez. Posteriormente, los datos del 10% de las estaciones en el *training set* es retirada para ser usada como set de validación durante el entrenamiento. Todas estas divisiones en sub-conjuntos son realizadas de forma aleatoria.

Por otro lado, para facilitar el aprendizaje, los datos de cada variable del *training set* son normalizadas entre 0 y 1, y las variables del *validation set* y del *test set* son ajustadas a esta normalización. La red es finalmente entrenada usando el optimizador Adam, con tasa de aprendizaje de 0.001, en lotes (*batches*) de 65536 series de tiempo durante 100 épocas de entrenamiento. La función de costo a minimizar en esta red será la de entropía cruzada binaria, que sirve para problemas de clasificación binarios (de sólo 2 clases).

En la **figura 8** se muestra la curva de aprendizaje de la red, es decir, la evolución de la función de costo. Se observa que a las 100 épocas la función de costo de la validación ya ha convergido a un valor estable.

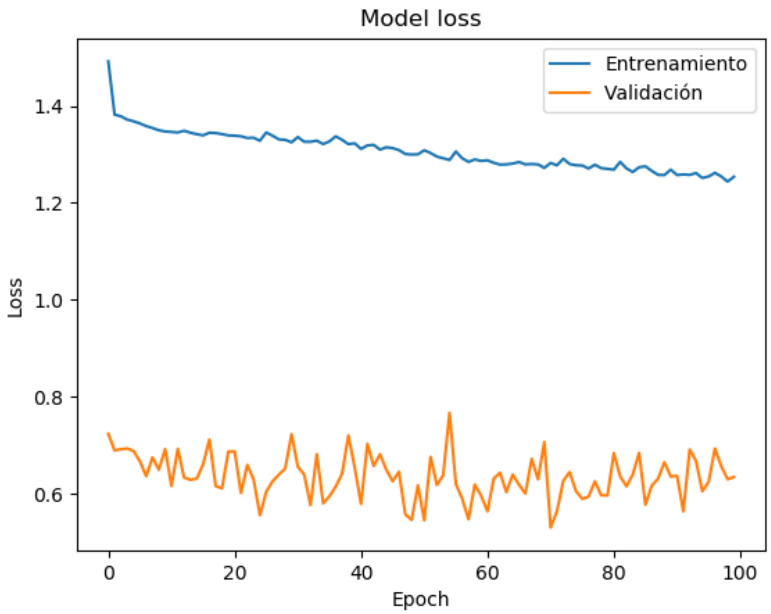


Figura 8

Finalmente los resultados para la red LSTM se muestran en la figura 9. Como es posible notar, la red predice con un 79% de exactitud los eventos normales del *test set*, pero para los eventos extremos tiene un 54% de exactitud. Esto muestra que el modelo no tuvo el éxito esperado para la predicción de eventos extremos. Es posible que con una mayor cantidad de información mejor sean las predicciones. Una variable que podría ser útil es la latitud del rio, ya que como se mostró anteriormente, las anomalías son muy diferentes entre sí, dependiendo fuertemente de las ubicaciones geográficas.