SPIKE

**Spike Challenge - Predicción de caudales extremos en Chile**

¡Gracias por participar en el proceso de selección de Spike! Como parte del proceso, este desafío nos ayudará a entender la manera en que te enfrentas a problemas nuevos y, además, podremos evaluar tus conocimientos actuales.

Algunos puntos importantes,

1. Este desafío no te debiera tomar más de 5 horas de tu tiempo. Por lo mismo, no esperamos respuestas muy pulidas ni perfectas.
2. Las preguntas irán aumentando en dificultad, por lo que intenta responder hasta donde puedas. Si por algún motivo hay alguna parte que no lograste completar, no hay problema.
3. Tendrás hasta el Jueves 17 de Octubre a las 23:59 para enviar tus respuestas al desafío.
4. Solo se aceptarán *Jupyter notebook (recomendado)*, *R Markdown* o *R Notebook* como formatos de entrega y solamente *python* o *R*. La idea es sea fácil para nosotros correr lo que ustedes escribieron (que sea reproducible).
5. Lee bien las instrucciones!

Accede a este link para encontrar las instrucciones y el dataset para el desafío: https://github.com/SpikeLab-CL/desafio_spike_cuencas

Saludos!
Spike

**Motivación**

*Nota: este segmento de motivación es para que entiendan un poco más del problema y por qué es relevante. No es necesario que entiendan todo, el resto del desafío los irá guiando en las cosas específicas que deben hacer. En otras palabras:* ***lean esto rápido y luego concéntrense en las preguntas e instrucciones del desafío.***

*Las olas de calor son eventos meteorológicos extremos (temperaturas extremas), que pueden generar impactos negativos en nuestro ecosistema. Estos impactos van desde días de verano muy calurosos, deshidratación, hasta aumento del riesgo de incendios y riesgo de aluviones por crecidas de los ríos (por derretimiento de nieve o lluvias líquidas en zonas altas de la cordillera).*

*Dado que las olas de calor responden a patrones de circulación atmosférica que pueden ser detectadas con días e incluso semanas de anticipación, existe una oportunidad sin precedentes de intentar predecir la ocurrencia de algunos de los impactos negativos asociados a estos eventos extremos. Esto es clave en el contexto de calentamiento global, en donde la frecuencia a escala global de este tipo de eventos extremos ha aumentado en el último siglo* https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2012GL053361.

*Para explorar esta potencial predictibilidad de los impactos asociados a las olas de calor, en particular los eventos de peakflows (crecidas de ríos, los que en casos extremos pueden llegar a desbordar el cauce), en Chile contamos con una amplia red de monitoreo de variables hidro-meteorológicas. Las estaciones meteorológicas que miden temperatura y precipitación son mantenidas por la Dirección General de Aguas (DGA) y la Dirección Meteorológica de Chile. Para potenciar el uso y la generación de conocimiento en base a estos datos públicos, el Centro de Ciencia del Clima y la Resiliencia (CR2) desarrolló un explorador climático (*http://explorador.cr2.cl*), desde el cual se pueden visualizar y descargar estos registros.*

*En el explorador climático se encuentran también los datos de caudal de las estaciones fluviométricas de la DGA, los que representan el agua total que aporta una cuenca hidrográfica a una sección de río. En el CR2 también se generó una base de datos de cuencas, que provee una serie de atributos de cada área aportante a las estaciones fluviométricas (http://camels.cr2.cl), además de series meteorológicas de temperatura y caudal promediadas dentro de la cuenca (es decir, no son datos de estaciones puntuales, sino que agregados espacialmente sobre el polígono de cada cuenca).*

*Algunas de las preguntas claves que en Chile debemos responder son:*

1. ***¿Ha aumentado la frecuencia de olas de calor en Chile?** Para esto se deben analizar las estaciones de temperatura.*
2. ***¿Existe una relación entre olas de calor y eventos extremos de caudal?***
3. *De existir una relación entre olas de calor y eventos extremos de caudal,* ***¿se puede explicar este evento extremo por las características de la cuenca en donde ocurre el peakflow?***

**Instrucciones**

Como dice el título, vamos a predecir los caudales extremos para cuencas en Chile. Para eso armamos un dataset usando datos públicos y reales de estaciones metereológicas.

Cada fila representa una medida de caudal diaria en una estación medidora. La medición de caudal estará asociada a características de la cuenca (fijas) y a mediciones diarias de temperatura y precipitación de estaciones cercanas.

El archivo *caudal_extra.csv* contiene todos los datos que van a necesitar para este desafío. Esta base la produjimos en Spike y sintetiza información de caudal, precipitación y temperatura.

**Nota sobre la producción de la base:** Las estaciones que miden el caudal y las que miden la temperatura y la precipitación no están en el mismo lugar. Para construir esta base, tomamos el polígono de la cuenca aguas arriba que corresponde a la estación de caudal y luego identificamos las estaciones de temperatura y precipitación que se encuentran en ese polígono. En muchos casos hay más de una estación, así que, por simpleza, tomamos el promedio de todas esas estaciones. De esta manera, cada día de medición de caudal va a estar asociada a una única medición de temperatura y de precipitación relevante (aunque puede haber cuencas sin estaciones de temperatura o precipitación).

*La base pesa 256 megas. Si tienen problemas de RAM, tomen un subconjunto de las estaciones y continúen con el desafío.*

Esta base tiene las siguientes variables:

- codigo_estacion: el código de la estación de medición de caudal.
- nombre: nombre del lugar en la cuenca donde está la estación.
- fecha: día de medición.
- caudal: medición de caudal de ese día.
- gauge_id: id de la cuenca.
- precip_promedio: precipitación promedio de ese día en la cuenca.
- temp_max_promedio: temperatura máxima  promedio de ese día en la cuenca.

Hemos creado además un canal en Gitter para que todos puedan colaborar entre sí. Ingresa en https://gitter.im/desafioSpike (se requiere cuenta github o gitlab).

**Desafío**

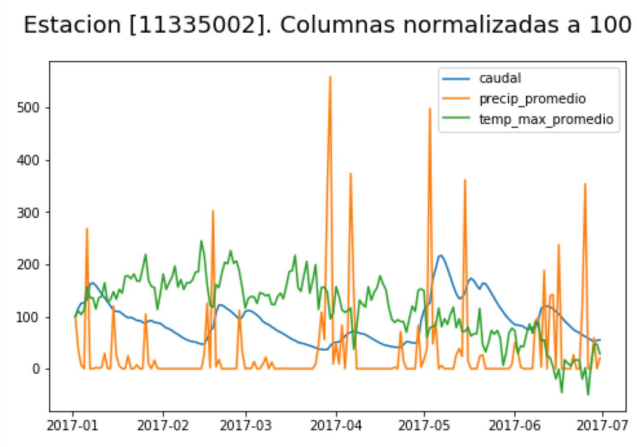
1. Baja el archivo *caudal_extra.csv*. Puedes bajarlo de BigQuery (vean el código al final de este documento), o bien, desde Github (https://github.com/SpikeLab-CL/desafio_spike_cuencas).
2. Analiza el dataset *caudal_extra.csv*. ¿Qué puedes decir de los datos, distribuciones, missing, u otros? ¿Hay algo que te llame la atención? ¿Por qué hay tantos valores missing? Pensar en la manera en que se elaboró el dataset, descrito más arriba. (Entregable: texto/imágenes)
3. Plots de precipitación, temperatura y caudal
    a. Escribir una función que tome como input una estación y haga plot de los datos para una columna. Debiese tener estos argumentos:

```
def time_plot_una_estacion(codigo_estacion, columna, fecha_min, fecha_max):
    ...
```

    b. Ahora escribir una función que haga plots de varias columnas, para poder visualizar caudal, precipitación y temperatura al mismo tiempo. Como las series están en diferentes escalas, sugerimos normalizarlas antes de hacer el plot (por ejemplo, dividiendo por la primera observación de cada seria)

```
def time_plot_estaciones_varias_columnas(codigo_estacion, columnas, fecha_min, fecha_max):
    ...
    ...
```

    Por ejemplo:



4. Crea tres variables llamadas
    - caudal_extremo
    - temp_extremo
    - precip_extremo

Dichas variables toman valor 1 un día si el caudal/temperatura/precipitación (según sea el caso) observado ese día es extremo. Esto significa que es mayor de lo "esperado". Para capturar esta idea, el valor de caudal, por ejemplo, toma valor 1 si está sobre el percentil 95 de ese caudal para esa estación del año (Verano, Primavera, Otoño, Invierno). Toma valor 0 cuando está bajo ese percentil. En otras palabras, para cada estación de medición y para cada estación del año, debes considerar la distribución histórica de caudal/temperatura/precipitación para elegir ese percentil 95.

Esta medida toma en cuenta la estacionalidad, pues, por ejemplo, una temperatura de 25 grados en invierno puede ser extrema, pero en verano es normal. También toma en cuenta que cada cuenca (o estación) es diferente. Lo que es extremo para una cuenca no lo es para la otra.

¿Les parece razonable esta medida para capturar algo “extremo”? ¿Usarían otra? ¿Cuál? (*Solamente descríbanla, no la codifiquen! Vamos a usar la definición de Spike para esta desafío*)

5. Analicen la variable *caudal_extremo*. Los comportamientos en diferentes cuencas son muy diferentes?
6. Hagan un plot del porcentaje de eventos extremos a través del tiempo (caudal_extremo, temp_extremo, precip_extremo). Se han vuelto más o menos comunes?
7. Predicción de caudal extremo. Entrena uno o varios modelos (usando el/los algoritmo(s) que prefieras) para estimar la probabilidad de un caudal extremo (la variable binaria caudal_extremo). Siéntete libre de generar variables adicionales y/o complementar con variables externas.
   ¿Qué datos podemos usar y cuáles no? Por supuesto, no podemos usar datos del futuro, pero ¿es lícito usar información del mismo día? ¿del día anterior? Todo depende de cómo propongas que el modelo se puede usar. Haz una propuesta de cómo usar tu modelo en la práctica (por ejemplo: una vez entrenado, voy a tomar los datos de XXX hasta el lunes y predecir para el día siguiente). Dada la propuesta, declara restricciones de información para caudal, temperatura y precipitación

8. Análisis de resultado del modelo
    a. Qué performance tiene el modelo? Qué métricas usaste para medir esa performance? Cuáles son las variables más importantes? Qué opinión te merecen los resultados?
    b. Si quisiéramos capturar alrededor de un 70% de los eventos de caudales extremos. Cuál es la precisión de tu modelo con ese porcentaje de captura? Cuéntanos si te parece útil.
9. Sube tu jupyter notebook, R Markdown, u otro a un repositorio de Github. (Por favor, usa una licencia como MIT para que otros puedan usar lo que avanzaste para combatir el cambio climático). Luego envíanos el link a jobs@spikelab.xyz usando el asunto: “Spike Challenge”.

## Para bajar el archivo desde BigQuery

Instalar pandas-gbq package.

Proyecto: spikelab
Tabla: public.caudal_extra_min

```
df = pd.read_gbq("SELECT * FROM public.caudal_extra_min",
project_id="spikelab")
```