



Copyright © 2021 NeuralWorks. Confidential and proprietary

# Challenge Data Scientist

## Instrucciones

- Debes entregar tu solución en un repositorio GitHub
- En el repositorio deben estar todos los archivos utilizados para la resolución de tu desafío.
- La solución debe estar implementada en un notebook .ipynb utilizando python 3, indicando claramente la pregunta que estás resolviendo. No serán revisados otros lenguajes como R o similar.
- Recuerda que no estamos en tu cabeza! Escribe los supuestos que estás asumiendo.
- Para este desafío te recomendamos que describas claramente cómo mejorar cada parte de tu ejercicio en caso de que tenga opción de mejora.
- Debes enviar el link al repositorio vía mail a [rayen@neuralworks.cl](mailto:rayen@neuralworks.cl) contestando el correo en el que se te envió este enunciado.

## Problema

El problema consiste en predecir la probabilidad de atraso de los vuelos que aterrizan o despegan del aeropuerto de Santiago de Chile (SCL). Para eso les entregamos un dataset usando datos públicos y reales donde cada fila corresponde a un vuelo que aterrizó o despegó de SCL. Para cada vuelo se cuenta con la siguiente información:

**Fecha-I** : Fecha y hora programada del vuelo.

**Vlo-I** : Número de vuelo programado.

**Ori-I** : Código de ciudad de origen programado.

**Des-I** : Código de ciudad de destino programado.

**Emp-I** : Código aerolínea de vuelo programado.

**Fecha-O** : Fecha y hora de operación del vuelo.

**Vlo-O** : Número de vuelo de operación del vuelo.

**Ori-O** : Código de ciudad de origen de operación

**Des-O** : Código de ciudad de destino de operación.

**Emp-O** : Código aerolínea de vuelo operado.

**DIA** : Día del mes de operación del vuelo.

**MES** : Número de mes de operación del vuelo.

**AÑO** : Año de operación del vuelo.

**DIANOM** : Día de la semana de operación del vuelo.

**TIPOVUELO** : Tipo de vuelo, I =Internacional, N =Nacional. **OPERA** : Nombre de aerolínea que opera.

**SIGLAORI** : Nombre ciudad origen.

**SIGLADES** : Nombre ciudad destino.

## Desafío

1. ¿Cómo se distribuyen los datos? ¿Qué te llama la atención o cuál es tu conclusión sobre esto?
2. Genera las columnas adicionales y luego expórtelas en un archivo `synthetic_features.csv` :
  - o `temporada_alta` : 1 si Fecha-I está entre 15-Dic y 3-Mar, o 15-Jul y 31-Jul, o 11-Sep y 30-Sep, 0 si no.
  - o `dif_min` : diferencia en minutos entre Fecha-O y Fecha-I .
  - o `atraso_15` : 1 si `dif_min` > 15, 0 si no.
  - o `periodo_dia` : mañana (entre 5:00 y 11:59), tarde (entre 12:00 y 18:59) y noche (entre 19:00 y 4:59), en base a Fecha-I .
3. ¿Cómo se compone la tasa de atraso por destino, aerolínea, mes del año, día de la semana, temporada, tipo de vuelo?  
¿Qué variables esperarías que más influyeran en predecir atrasos?
4. Entrena uno o varios modelos (usando el/los algoritmo(s) que prefieras) para estimar la probabilidad de atraso de un vuelo. Siéntete libre de generar variables adicionales y/o complementar con variables externas.
5. Evalúa tu modelo. ¿Qué performance tiene? ¿Qué métricas usaste para evaluar esa performance y por qué? ¿Por qué elegiste ese algoritmo en particular? ¿Qué variables son las que más influyen en la predicción? ¿Cómo podrías mejorar la performance?

## Aspectos a considerar

Orden y claridad al momento de plantear un análisis, idea, código, etc.

Creatividad para resolver el desafío.

Código versionado en Git.

No vamos a revisar excel, macros, códigos en R.

No vamos a revisar desafíos que no lleguen en la fecha indicada.

Ante cualquier duda deja explícitos tus supuestos.

No vivimos en tu cabeza, trata de expresarte lo mejor posible para explicar tus decisiones y respuestas.