Ejercicio ML - Solicitudes de prestaciones

El dataset a utilizar lo tomamos del catálogo de datos abiertos:

<https://catalogodatos.gub.uy/dataset/fondo-nacional-de-recursos-solicitudes-de-prestaciones-fondo-nacional-de-recursos/resource/fbb5238a-934d-4ea4-a5ca-59c6c60ba2fc>

Vamos a predecir si se autoriza o no una solicitud de prestación.

Para realizar el ejercicio tomar como base el código de la demo de sobrevivencia del Titanic.

# **Importar librerías a utilizar**

# **Cargar los datos del archivo datos-solicitudes-2017.xlsx**

data = pd.read\_excel('nombre\_archivo.xlsx')

# **Analizar los datos**

* 1. Ver los datos
  2. Ver cantidad de filas y columnas
  3. Ver estadísticas de los datos
  4. Crear columna categorizando las edades:
     + Desconocido: -1 a 0
     + Niño: 0 a 14
     + Adolescente: 14 a 25
     + Joven: 24 a 35
     + Adulto: 35 a 60
     + Adulto: Mayor 60 a 90
     + Longevo: 90 - 105
  5. Ver el porcentaje de aprobación con respecto a algunas características
     + Hay que crear un nuevo campo para los estados en formato numérico para poder hacer cálculos

data["estado"] = np.where(data["estado\_solicitud"]=="AUTORIZADO", 1, 0)

* 1. Ver si alguna columna tiene datos vacíos
  2. Ver los valores y cantidades de cada columna

# **Eliminar columnas redundantes**

# **Transformar datos**

* 1. Observar si hay valores categóricos
  2. En caso de que haya transformarlos a numéricos

# **Dividir los datos**

* 1. Dividir las variables independientes de la variable dependiente
  2. Dividir nuevamente en datos en entrenamiento (80%) y test (20%)

# **Estandarizar las variables independientes**

# **Crear 4 modelos utilizando distintos algoritmos de clasificación**

* 1. Uno de los modelos elegidos debe ser Random Forest

# **Entrenarlos**

# **Validar los 4 modelos**

* 1. Realizar la matriz de confusión con los datos de test
  2. Utilizar distintas funciones para validar la salida de los datos de entrenamiento
  3. Calcular de forma manual la métrica de exactitud para los datos de test

# **Analizar las características más importantes utilizando el modelo Random Forest**

# **Comparar la salida de la predicción de uno de los modelos con la salida real**

# 