# Taller de Aprendizaje Automático

## Actividades Taller 1

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Montevideo, 2024

# Tabla de contenido

# Objetivos del Taller

- Desarrollar algunas de las habilidades necesarias para enfrentar un problema de aprendizaje automático de punta a punta
- Adquirir habilidades en el manejo de herramientas de procesamiento de datos: *numpy*, *matplotlib*, *scikit-learn*, *pandas*, *keras*, *tensorflow*, ...
- Instancia de intercambio entre estudiantes y docentes

# Sobre los talleres

- Formato de la clase
  - Se plantea un problema a resolver
  - Los estudiantes trabajan en el problema y los docentes hacen recorridas por los distintos grupos
  - Pasada la mitad de la clase se hace una puesta en común
- Evaluación
  - Se entregarán **dos informes individuales**
    - Informe 1: Cubre los talleres 3, 4 y 5
    - Informe 2: Cubre los talleres 8 y 9
  - El largo de los informes no puede superar las 4 carillas

# Sistema de puntos extra

- Cuestionarios presenciales
  - A la mitad de la clase se habilitará un cuestionario con preguntas relacionadas al Taller
  - Se cerrará a las 12:10hs.
  - Se pueden ganar hasta 5 puntos extras por taller

- Distribución de los puntos

| Cuestionario | Puntos extra para |
|---|---|
| Taller 1 | Entregable 1 |
| Taller 2 | Entregable 1 |
| Taller 3 | Proyecto 1 |
| Taller 4 | Proyecto 1 |
| Taller 5 | Proyecto 1 |
| Taller 6 | Entregable 2 |
| Taller 7 | Entregable 2 |
| Taller 8 | Proyecto 2 |
| Taller 9 | Proyecto 2 |

# Kaggle

- En general se trabajará con datos disponibles en Kaggle
- Cada estudiante tendrá que hacerse un usuario en Kaggle
- Recursos útiles:
  - Documentación Kaggle
  - Kaggle API
  - Tutoriales

# ¿Qué es Pandas?

- Bliblioteca que permite
  - Crear, leer y escribir datos
  - Manipular los datos (seleccionar, indexar, asignar)
  - Generar Estadísticas
  - Agrupar y ordenar datos
  - Preprocesar Datos
  - Renombrar y Combinar
- Tutoriales de Pandas
  - Documentación Oficial
  - Curso Kaggle

# Métodos útiles para actividad de clase

- Levantar datos

```
# Se importa la biblioteca pandas
import pandas as pd

# Se levantan los datos
data_frame = pd.read_csv(filename)
```

- Explorar datos

```
# Visualizar primeras instancias
data_frame.head()

print(data_frame.shape)
(129971, 14)

# Visualizar informacion sobre la base de datos
data_frame.info()
```

# Métodos útiles para actividad de clase

- Acceso a los datos

```
# Seleccionar los valores de una caracteristca
#(una columna)
data_frame['petal_length']
data_frame.petal_length
# Acceder a un elemento
data_frame['petal_length'][0]

###### Seleccion basada en indices  #####
# Obtener la primera fila
data_frame.iloc[0]
# Obtener los primeros m elementos de la columna n
data_frame.iloc[:m, n]

##### Seleccion basada en etiquetas  #####
# Los ultimos valores de 'petal_length'
data_frame.loc[-5:, 'petal_length']
# Valores de 'petal length' y 'petal width'
data_frame.loc[:, ['petal_length', 'petal_width']]
```

# Métodos útiles para actividad de clase

- Mostrar estadísticas

```
# Del data frame
data_frame.describe()

# De una caracteristica en particular
data_frame.petal_length.describe()
data_frame.petal_length.mean() # solo la media
# lista de valores unicos
data_frame.petal_length.unique()
# lista de valores unicos y cuantas veces aparecen
data_frame.petal_length.value_counts()
```

- Agrupamiento de datos
  - *groupby()*

# Preprocesamiento de datos

- Datos faltantes
  - *isnull()*, *notnull()*
  - *fillna()*
  - *dropna()*
- Fechas
  - *to_datetime()*

# Actividad del Titanic

- Generar un modelo que prediga si un pasajero sobrevivirá a partir de atributos personales.
- En la actividad se pondrá en práctica:
  - Manejo de la biblioteca **pandas**
  - Exploración de datos
  - Manejo de **pipelines** de *scikit-learn*
  - Simulación de construcción y puesta en producción de un modelo en la plataforma **kaggle**.

# Actividades parte 2

- Identifique el atributo a predecir
- ¿Es un problema de clases desbalanceadas?
- Identificar y cuantificar datos faltantes.
- Identificar los atributos numéricos y categóricos.
- En caso de contar con datos categóricos identifique las categorías.
- Obtener el porcentaje de pasajeros del conjunto de entrenamiento que sobrevivió.
- Obtener el porcentaje de pasajeros dentro de cada categoría que sobrevivió. Asegúrese de poder responder preguntas del tipo: ¿Qué porcentaje de mujeres sobrevivieron? ¿Cuál fue el porcentaje de paraseros de primera clase (PClass 1) que sobrevivió?

# Bosson de Higgs