

# Seminario de Aplicaciones Actuariales

## Seminario de Estadística I

### Aplicaciones de Ciencia de Datos con Python

Profesor: Dr. Arrigo Coen Coria

Ayudante: Act. Miriam Colín

#### Tarea 2: Algoritmos de Clasificación

Instrucciones:

- La entrega será el **viernes 16 de agosto**. Puede ser de manera individual o en equipos de a lo más 3 alumnos.
- Las preguntas 1-3 se entregarán en un pdf con el nombre:  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_1\_3
- Cada una de las preguntas 4-7 se entregarán en un jupyter notebook con los nombres:  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_4,  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_5,  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_6,  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_7, respectivamente.
- Responde las siguientes preguntas y realiza lo que se pide.

1. Describe los siguientes algoritmos, da ejemplos de su aplicación y escribe sus ventajas y desventajas al implementarlo.
  - a) KNN
  - b) SVM
  - c) Regresión Logística

Cada una de tus respuestas debe responder lo siguiente:

- ¿Cuál es la idea fundamental del algoritmo?
- ¿Qué definiciones tiene esta metodología?
- ¿Qué tipo de datos puede utilizar el algoritmo y si requieren los datos alguna transformación?
- ¿Qué diferencias tiene con respecto a otros algoritmos?

2. Describe los siguientes conceptos, menciona para qué se utilizan y da ejemplos de su aplicación:
  - a) Curvas de Aprendizaje
  - b)  $R^2$  y  $R^2$  ajustada (menciona sus diferencias, ventajas y desventajas)
  - c) Curvas ROC

3. Describe cuáles son las características de cada uno de los miembros de la familia de *Gradient Descent*, menciona sus diferencias y da ejemplos de su aplicación:
  - a) *Gradient Descent*
  - b) *Stochastic Gradient Descent*
  - c) *Mini-batch Gradient Descent*
4. Con la base de datos del archivo *T2\_diabetes*:
  - a) Realizar un análisis completo de las tres variables (describir/interpretar variables, gráficas, resultados, ...)
  - b) Ajusta un modelo para predecir si la persona tiene o no diabetes (variable *diabetes*) utilizando todas las columnas de la base para ajustar un modelo de KNN con 3 valores distintos de *k* y comparar los resultados
  - c) Mostrar la gráfica con regiones de decisión del mejor modelo  
`//plot_decision_regions()`
  - d) Escribir conclusiones
5. Con la base de datos del archivo *T2\_zoo*:
  - a) Realizar un análisis completo de las tres variables que consideres más importantes (describir/interpretar variables, gráficas, resultados, ...)
  - b) Ajusta un modelo para predecir el tipo de animal (variable *type*) utilizando todas las columnas de la base utilizando un modelo de SVM con 2 *kernels* distintos (elegir dos: *linear*, *poly*, *rbf*, *sigmoid*) y comparar los resultados
  - c) Escribir conclusiones
6. Con la base de datos del archivo *T2\_spam*:
  - a) Realizar un análisis completo de las tres variables que consideres más importantes (describir/interpretar variables, gráficas, resultados, ...)
  - b) Ajusta un modelo para predecir si el correo es spam o no (variable *spam*) utilizando todas las columnas de la base ajustando un modelo de Regresión Logística
  - c) Escribir conclusiones (compara tus resultados con y sin regularización)
7. Ajusta un modelo de KNN, SVM, Regresión Logística, o una mezcla ponderada de estos tres KNN/SVM/RL para clasificar la salud de los embarazos de la base de datos <https://www.kaggle.com/c/tabular-playground-series-jun-2021/submissions>, y obtén un score menor a 1.82