# 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 ***lunes 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 4-6 se entregarán en un jupyter noteboook con los nombres:

T1\_*ApPatNom1*\_ *ApPatNom2*\_ *ApPatNom3*\_4,

T1\_*ApPatNom1*\_ *ApPatNom2*\_ *ApPatNom3*\_5,

T1\_*ApPatNom1*\_ *ApPatNom2*\_ *ApPatNom3*\_6, 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.
2. KNN
3. SVM
4. 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?

1. Describe los siguientes conceptos, menciona para qué se utilizan y da ejemplos de su aplicación:
2. Curvas de Aprendizaje
3. y ajustada (menciona sus diferencias, ventajas y desventajas)
4. Curvas ROC
5. 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*:*
6. *Gradient Descent*
7. *Stochastic Gradient Descent*
8. *Mini-batch Gradient Descent*
9. Con la base de datos del archivo *T2\_diabetes:*
10. Realizar un análisis completo de las tres variables que consideres más importantes (describir/interpretar variables, gráficas, resultados, …)
11. 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
12. Mostrar la gráfica con regiones de decisión del mejor modelo *//plot\_decision\_regions()*
13. Escribir conclusiones
14. Con la base de datos del archivo *T2\_zoo:*
15. Realizar un análisis completo de las tres variables que consideres más importantes (describir/interpretar variables, gráficas, resultados, …)
16. 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
17. Escribir conclusiones
18. Con la base de datos del archivo *T2\_* *spam:*
19. Realizar un análisis completo de las tres variables que consideres más importantes (describir/interpretar variables, gráficas, resultados, …)
20. 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
21. Escribir conclusiones (compara tus resultados con y sin regularización)
22. 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 obten un score menor a 1.82