Profesor: Dr. Oldemar Rodríguez Rojas
Análisis de Datos 2
Fecha de Entrega: Domingo 6 de noviembre a las 12 media noche
Instrucciones:

- Las tareas deben ser subida la Aula Virtual antes de las 6:00pm. Luego de esta hora pierde 20 puntos y cada día de retraso adicional perderá 20 puntos más.
- Las tareas son estrictamente individuales.
- Tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso.
- Cada día de entrega tardía tendrá un rebajo de 20 puntos.

# Tarea Número 10

1. **[25 puntos]** La tabla de datos `novatosNBA.csv` contiene diferentes métricas de desempeño de novatos de la NBA en su primera temporada. Para esta tabla, las 21 primeras columnas corresponden a las variables predictoras y la variable Permanencia es la variable a predecir, la cual indica si el jugador permanece en la NBA luego de 5 años. La tabla contiene 1340 filas (individuos) y 21 columnas (variables), con la tabla realice lo siguiente:

   *a*) El objetivo de este ejercicio es analizar la variación del error (usando el enfoque training-testing) para la predicción de la variable `Permanencia`. Para esto repita 5 veces el cálculo de error global de predicción usando el método de los `k vecinos` más cercanos (use `n_neighbors=50`) y con un 75 % de los datos para tabla aprendizaje y un 25 % para la tabla testing. Grafique los resultados.

   *b*) El objetivo de este ejercicio es medir el error para la predicción de la variable `Permanencia`, utilizando validación cruzada con `K grupos` (`K-fold cross-validation`). Para esto usando el método de los `k vecinos` más cercanos (use `n_neighbors=50`) realice una validación cruzada 5 veces con 10 grupos (folds) y grafique el error obtenido en cada iteración, agregue en este gráfico los 5 errores generados en el ejercicio anterior.

   *c*) ¿Qué se puede concluir?

2. **[25 puntos]** Utilizando nuevamente la tabla `novatosNBA.csv` realice lo siguiente:

   *a*) El objetivo de este ejercicio es calibrar el método de `KNeighborsClassifier`. Para esto genere Validaciones Cruzadas con 10 grupos calibrando el modelo de acuerdo con los tres tipos de algoritmos que este permite, es decir, con `ball_tree`, `kd_tree` y `brute`.

   *b*) ¿Se puede determinar con claridad cuál algoritmo es el mejor? ¿Cuál algoritmo usaría con base en la información obtenida?

3. **[25 puntos]** Utilizando nuevamente la tabla `novatosNBA.csv` realice lo siguiente:

*a*) El objetivo de este ejercicio es comparar todos los métodos predictivos vistos en el curso con esta tabla de datos. Para esto genere Validaciones Cruzadas con 10 grupos para los métodos `SVM`, `KNN`, `Árboles`, `Bosques`, `ADA Boosting`, `EXtreme Gradient Boosting`, `Bayes`, `LDA`, y `QDA`. Para KNN use los parámetros obtenidos en las calibraciones realizadas en los ejercicios anteriores (en teoría se deberían calibrar todos los métodos). Luego realice un gráfico de barras para comparar los métodos.

*b*) ¿Se puede determinar con claridad cuál modelo es el mejor? ¿Cuál modelo usaría con base en la información obtenida?

4. [**25 puntos**] Programe una clase denominada `validacion_cruzada` la cual recibe como atributos una lista de modelos predictivos, la cantidad de validaciones cruzadas a aplicar y la cantidad de grupos a formar. Además de los métodos que debe llevar toda clase, programe en esta clase un método que permita aplicar la validación cruzada utilizando los valores de los atributos (Debe aplicar la validación cruzada para cada uno de los modelos de la lista) y un diccionario con la matriz de confusión de cada uno de los modelos. Debe adjuntar una prueba de uso de la clase con al menos 3 modelos.