Taller 2

Parte 1:

Para este ejercicio se utilizará la base de datos *cancer_mama* perteneciente a la Universidad de Wisconsin. Los datos contienen el diagnóstico de cáncer de mamá en 699 mujeres. Las características del diagnóstico corresponden a características obtenidas mediante un análisis de imágenes digitalizadas de masas en los senos. Para mayor información pueden consultar: UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. El objetivo de este ejercicio es predecir el diagnóstico de cáncer de mamá (*Diagnosis*) con base en las características de la masa observada.

1. Hacer un análisis descriptivo de la base de datos. Este análisis debe cumplir con:
    a) Distribución de sus variables. Esto incluye, media, varianza, un histograma, cantidad observaciones vacías por variable, y cualquier otra que vea relevante.
    b) Relación de sus variables con respecto a la que se quiere predecir (*Diagnosis*). Esto incluye las correlaciones de sus variables con su variable objetivo.
    c) Una breve descripción de su base de datos con base en el punto 1.a y 1.b.
2. Pre-procesar su base de datos para un modelo de ML:
    a) Estandarizar, en caso de que sea necesario.
    b) Completar los valores faltantes, en caso de que sea necesario.
3. Implementar correctamente los siguientes modelos de ML (Si el grupo es de menos de 4 personas no hagan el punto 3.b.):
    a) Un modelo de k-NN (k vecinos más cercanos), usando validación cruzada para encontrar el mejor k posible.
    b) Un modelo tipo Árbol, usando validación cruzada para encontrar el mínimo de elementos de su muestra en las hojas (min_samples_leaf) y la máxima profundidad su árbol (max_dept).
    c) Un modelo de Bosque Aleatorio (RandomForest), usando validación cruzada para encontrar el mínimo de elementos de su muestra en las hojas (min_samples_leaf), la máxima profundidad su árbol (max_dept) y el número de estimadores.
    d) Un modelo de regresión logística normal y con regularización Ridge y Lasso, usando validación cruzada para para encontrar el mejor híper-parámetro de regularización (λ).
4. Para los modelos implementados en el punto 3 se debe:
    a) Computar el MSE, accuracy, precision, recall, AUC.
    b) Graficar la curva ROC.
5. Con base en el punto 4 debe argumentar cuál de los modelos implementados en el punto 3 es mejor.

Parte 2:

Para este ejercicio carguen la base de del archivo credit_card.csv. Esta base de datos contiene los siguientes datos:

- **card:** 1 if application for credit card accepted, 0 if not
- **reports:** Number of major derogatory reports
- **age:** Age n years plus twelfths of a year
- **income:** Yearly income (divided by 10,000)
- **share:** Ratio of monthly credit card expenditure to yearly income
- **expenditure:** Average monthly credit card expenditure
- **owner:** yes if owns their home, no if rent
- **selfempl:** yes if self employed, no if not.
- **dependents:** 1 + number of dependents
- **months:** Months living at current address
- **majorcards:** Number of major credit cards held
- **active:** Number of active credit accounts
- **sex:** Male or Female
- **race:** White, Black or Other

1. Entrenar un modelo que permita predecir la probabilidad que la tarjeta de crédito sea aprobada utilizando todas las variables.
2. Tomar el punto de corte óptimo de la curva ROC para categorizar las predicciones.
3. Determinar si su modelo tiene sesgo con respecto a las variables de sexo y raza.
4. Retomar las predicciones del punto 1, hacer una calibración de su modelo con **Platt scaling** y otra con **transformación isotónica.** (Revisen la clase sklearn.calibration.CalibratedClassifierCV)
5. Repetir los puntos 2 y 3 sobre los modelos calibrados.
6. Analizar los resultados obtenidos.