EJERCICIO PROPUESTO PARA LA EVALUACIÓN DEL MÓDULO MINERÍA DE DATOS II.

La evaluación del módulo Minería de Datos II se lleva cabo con tres bases de datos. Una de ellas para aplicar los métodos bayesianos y del Gradient Boosting y otros dos conjuntos de imágenes, a elegir uno de ellos, para aplicar una red convolucional para realizar la clasificación.

Primer ejercicio

Las instituciones financieras incurren en pérdidas significativas debido al incumplimiento de los préstamos para vehículos. Esto ha provocado un endurecimiento en la suscripción y un aumento de las tasas de rechazo de dichos préstamos.

La necesidad de disponer de un mejor modelo de calificación del riesgo crediticio justifica la realización de un estudio destinado a estimar los determinantes del incumplimiento de los préstamos para vehículos.

Por tanto, el trabajo consistirá en predecir con precisión la probabilidad de que el prestatario incumpla el pago del préstamo para el vehículo en el primer EMI (Cuotas mensuales equivalentes) en la fecha de vencimiento. Al hacerlo, se asegurará de que los clientes capaces de pagar no sean rechazados y que se puedan identificar determinantes importantes que se pueden utilizar para minimizar más las tasas de incumplimiento.

En los conjuntos de datos que se proporcionan se recoge la siguiente información sobre el préstamo y el prestatario:

- Información del prestatario (datos demográficos como edad, prueba de identidad, etc.).
- Información del préstamo (detalles del desembolso, relación préstamo-valor, etc.).
- Datos e historial de la Oficina (puntaje de la Oficina, número de cuentas activas, estado de otros préstamos, historial crediticio, etc.).

Se dispone de tres conjuntos de datos, correspondientes al entrenamiento (**train.csv**), al test (**test.csv**) y a la descripción de variables (**data_dictionary.csv**).

Algunas cuestiones sobre el tratamiento de la base de datos:

- La variable explicativa es loan_default.
- De cara a la modelización, se prescinde de las siguientes variables: UniqueID, branch_id, supplier_id, Current_pincode_ID, State_ID, Employee_code_ID, MobileNo_Avl_Flag.
- En relación con la variable PERFORM_CNS.SCORE.DESCRIPTION, se consideran los valores de la A a la M, el resto son desconocidos.
- La variable edad se calcula de la siguiente manera: Age = DisbursalDate - Date.of.Birth.
- Las variables numéricas son las siguientes:
  disbursed_amount, asset_cost, PRI.NO.OF.ACCTS, PRI.ACTIVE.ACCTS, PRI.OVERDUE.ACCTS, PRI.CURRENT.BALANCE, PRI.SANCTIONED.AMOUNT, PRI.DISBURSED.AMOUNT, SEC.NO.OF.ACCTS, SEC.ACTIVE.ACCTS, SEC.OVERDUE.ACCTS, SEC.CURRENT.BALANCE, SEC.SANCTIONED.AMOUNT, SEC.DISBURSED.AMOUNT, PRIMARY.INSTAL.AMT, SEC.INSTAL.AMT, NEW.ACCTS.IN.LAST.SIX.MONTHS, DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS, NO.OF_INQUIRIES, Age, NEW.ACCTS.IN.LAST.SIX.MONTHS, DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS
- Las variables categóricas son las siguientes:

manufacturer_id, Aadhar_flag, PAN_flag, VoterID_flag, Driving_flag, Passport_flag, PERFORM_CNS.SCORE, NEW.ACCTS.IN.LAST.SIX.MONTHS, DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS, AVERAGE.ACCT.AGE, NO.OF_INQUIRIES, PERFORM_CNS.SCORE.DESCRIPTION, AVERAGE.ACCT.AGE, CREDIT.HISTORY.LENGTH, Employment Type

Con esta base de datos se solicita que se razone y explique, fundamentalmente a través de los modelos explicados en el módulo ocho, cuál o cuáles de ellos recomendaría. Como ya se dispone del conocimiento de una amplia variedad de algoritmos y métodos de Data Science estudiados en este máster, se anima a que se comparen el mayor número de algoritmos de clasificación para determinar cuáles de ellos presentan el mejor rendimiento.

Segundo Ejercicio

Para evaluar la parte de Deep Learning se debe de elegir un conjunto de imágenes de los dos que se proponen.

La **primera base de datos** de imágenes es el conocido conjunto de datos denominado CIFAR 10 que procede del Canadian Institute for Advanced Research y que consta de 60.000 imágenes (50.000 de entrenamiento y 10.000 de test) en color de 32x32 píxeles repartidas en 10 clases y que se puede leer directamente desde la librería Keras.

El **segundo conjunto** de imágenes es una construcción artificial de personas con y sin mascarilla. La base de datos ha sido obtenida de la siguiente dirección de internet: https://github.com/prajnasb/observations . Hay que descargar toda la información entera (https://github.com/prajnasb/observations/archive/master.zip URL de la descarga directa). Posteriormente tenemos que acceder a la carpeta **observations-master\experiements\dest_folder** donde tendremos las carpetas con las imágenes de entrenamiento, **train**, validación, **val** y test, **test**. Cogemos estas tres carpetas y las llevamos a nuestro directorio de trabajo para trabajar con ellas.

Cada una de estas carpetas contendrá una carpeta llamada with_mask y otra denominada without_mask, que contendrán las imágenes que tienen mascarilla y las que no. El conjunto de datos se reparte de la siguiente manera:

- train with_mask: 658
- train without_mask: 658
- val with_mask: 71
- val without_mask: 71
- test with_mask: 97
- test without_mask: 97

Una vez elegido uno de los dos conjuntos de datos se solicita que se realice una **clasificación de las imágenes** que **incluya** cualquier **estrategia para aumentar en lo posible la precisión**: Data Aumentation, Regularización, BatchNormalization, Features Extraction, Fine Tune, modelos preentrenados…. Se valorará que el código esté comentado viendo que se entiende los mecanismos que se están aplicando para conseguir una mejor solución.

Este ejercicio se puede realizar tanto en R como en Python. Algunas **recomendaciones** para la resolución de este segundo ejercicio son: habilitar la GPU del equipo informático (en caso de tener), utilizar Google colaboratory o reducir el número de iteraciones del modelo (epochs).

Calificación de los Ejercicios

El primer ejercicio tendrá una valoración de 6,5 puntos y el ejercicio de Deep learning de 3,5 puntos.

Tareas a realizar en el primer ejercicio.

1. **Preprocesado de la información: tablas, gráficos, balanceo de la muestra, etcétera (1,5 puntos)**
2. **Aplicación de Modelos (3 puntos)**
3. **Resumen de la información obtenida y comparación de modelos por métodos gráficos y contrastes de hipótesis (2 puntos)**

Aclaraciones sobre la importancia de los diferentes análisis y forma de evaluación

- De forma general apreciamos bastante que se realicen comentarios sobre los resultados que se obtengan.
- En la primera e importante tarea a realizar como es el preprocesado, además de la estadística descriptiva de los datos, se deberán realizar los siguientes procedimientosi:
    - Selección de variables. Elija con alguna técnica un conjunto pequeño de variables explicativas.
    - Balanceo de muestras: Utilice las variables seleccionadas en el paso anterior para construir una muestra equilibrada de la clase mayoritaria.
- En el apartado 2 se valora la inclusión de otros modelos explicados en otros módulos (que se conozcan que suelen tener buenos resultados), así como la simulación de los modelos con diferentes configuraciones. También se apreciará la presentación de los diferentes gráficos de los análisis y, sobre todo, de la curva ROC.
- El epígrafe 3 debe contener al menos un cuadro resumen con los resultados de todos los modelos utilizados, tanto con los datos de entrenamiento como de test, donde se especifiquen varias medidas de evaluación (porcentaje de aciertos por clases, área bajo la curva ROC…)
- También es importante y valorable la presentación de los resultados: orden, estética, creatividad, aportaciones en programación, etcétera.
- Se tiene que presentar un documento que indique claramente los análisis efectuados así como el código utilizado. NO ES NECESARIO PRESENTARLO EN RMARKDOWN.
- **RECORDARLES QUE ES MUY IMPORTANTE LA INTERPRETACIÓN DE LOS RESULTADOS, NO SOLO EL CÓDIGO DE LOS DIFERENTES MODELOS QUE APLIQUEN.**

En la resolución de la tareas encomendadas se utilizará exclusivamente cualquier librería de R, de Python o el programa WEKA.

[i] El preprocesado que se sugiere también se puede programar en Python o en Weka